UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
Câmpus de Marília

ELISMAR VICENTE DOS REIS

# **Proposta de metodologia para abordagem terminológica da análise de domínio baseada em mineração de texto:** uma aplicação na Ciência da Informação.

Marília
2023

ELISMAR VICENTE DOS REIS

# **Proposta de metodologia para abordagem terminológica da análise de domínio baseada em mineração de texto:** uma aplicação na Ciência da Informação.

Tese apresentada ao Programa de Pós-Graduação em Ciência da Informação como parte das exigências para obtenção do título de Doutor em Ciência da Informação pela Faculdade de Filosofia e Ciências, Universidade Estadual Paulista (UNESP) Campus de Marília.

**Área de concentração:** Informação, Tecnologia e Conhecimento.

**Orientadora:** Dr.ª Ely Francina Tannuri de Oliveira.

**Coorientador**: Dr. Ricardo César Gonçalves Sant'Ana.

Marília
2023

R375p

Reis, Elismar Vicente dos

Proposta de metodologia para abordagem terminológica da análise de domínio baseada em mineração de texto : uma aplicação na Ciência da Informação / Elismar Vicente dos Reis. -- Marília, 2023

151 p.

Tese (doutorado) - Universidade Estadual Paulista (Unesp), Faculdade de Filosofia e Ciências, Marília

Orientadora: Ely Francina Tannuri de Oliveira

Coorientador: Ricardo César Gonçalves Sant'Ana

1. metodologia. 2. terminologia. 3. análise de domínio. 4. mineração de texto. 5. literatura científica. I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Biblioteca da Faculdade de Filosofia e Ciências, Marília. Dados fornecidos pelo autor(a).

## IMPACTO POTENCIAL DESTA PESQUISA

Esta tese pode contribuir aos Objetivos do Desenvolvimento Sustentável (ODS) no Item 9, que versa sobre “construir infraestruturas resilientes, promover a industrialização inclusiva e sustentável e fomentar a inovação”. O conteúdo do subitem 9.5 refere-se a “fortalecer a pesquisa científica, melhorar as capacidades tecnológicas de setores industriais em todos os países, particularmente os países em desenvolvimento, inclusive, até 2030, incentivando a inovação e aumentando substancialmente o número de trabalhadores de pesquisa e desenvolvimento por milhão de pessoas e os gastos público e privado em pesquisa e desenvolvimento”. Embora o projeto decorrente não abarque a totalidade desse subitem, apresenta real potencialidade de ajudar em alguns pontos, pois trata-se de metodologia destinada a inovar pesquisas científicas no campo da Ciência da Informação, mas que pode servir também a outras áreas, possibilitando o aumento do número de pesquisas com o emprego da estrutura metodológica desenvolvida. Outrossim, com relação ao item 17, que tem como objetivo “reforçar os meios de implementação e revitalizar a parceria global para o desenvolvimento sustentável”, serve-se ao tema tecnologia no subitem 17.8, o qual trata da questão de “operacionalizar plenamente o Banco de Tecnologia e o mecanismo de capacitação em ciência, tecnologia e inovação para os países menos desenvolvidos até 2017, e aumentar o uso de tecnologias de capacitação, em particular das tecnologias de informação e comunicação”. Ainda no campo da tecnologia, mas com o viés voltado para pesquisas da literatura científica, a tese pode ainda contribuir com o item 12, que tem como mote “garantir padrões de consumo e de produção sustentáveis”, pois, a proposta metodológica viabiliza a investigação automatizada para descoberta de padrões e elementos ocultos nos textos, revelando informações desconhecidas e diferentes perspectivas. Ao investigar publicações específicas de determinada área, a possibilidade de novos conhecimentos pode cooperar com o item 12.a, que busca “apoiar países em desenvolvimento a fortalecer suas capacidades científicas e tecnológicas para mudar para padrões mais sustentáveis de produção e consumo”.

## POTENTIAL IMPACT OF THIS RESEARCH

This doctoral dissertation contributes to the to the Sustainable Development Goals (SDGs) in Goal 9, which deals with “build resilient infrastructures, promote inclusive and sustainable industrialization and foster innovation”. The content of target 9.5 refers to “enhance scientific research, upgrade the technological capabilities of industrial sectors in all countries, in particular developing countries, including, by 2030, encouraging innovation and substantially increasing the number of research and development workers and per 1 million people and public and private research and development spending”. Although the resulting project does not cover the entirety of this target, it can potentially help in some points, as the methodology is aimed at innovating scientific research in the field of Information Science. In addition, it can also serve other areas, enabling the increase in the number of investigations with the use of the developed methodological structure. Furthermore, regarding Goal 17, which aims to “strengthen the means of implementation and revitalize the global partnership for sustainable development”, the technology theme is served in target 17.8, which deals with the issue of “fully operationalize the technology bank and science, technology and innovation capacity-building mechanist for least developed countries by 2017 and enhance the use of enabling technology, in particular information and communication technology”. Still in the field of technology, but with a bias towards research in the scientific literature, the dissertation could also have a valuable contribution to goal 12, whose motto is “ensure sustainable consumption and production patterns”, as the methodological proposal enables the automated investigation to discover patterns and hidden elements in texts, revealing unknown information and different perspectives. When investigating specific publications in a given area, the possibility of building new knowledge can cooperate with item 12.a, which seeks to “support developing countries to strengthen their scientific and technological capacity to move towards more sustainable patterns of production and consumption”.

ELISMAR VICENTE DOS REIS

**Proposta de metodologia para abordagem terminológica da análise de domínio baseada em mineração de texto:** uma aplicação na Ciência da Informação.

Tese apresentada ao Programa de Pós-Graduação em Ciência da Informação da Universidade Estadual Paulista “Júlio de Mesquita Filho” (UNESP), como requisito parcial para o título de Doutor em Ciência da Informação.

Área de concentração: Informação, Tecnologia e Conhecimento.
Linha de Pesquisa: Produção e Organização da Informação.

Banca Examinadora

_____
Prof.ª Dr.ª Ely Francina Tannuri de Oliveira
UNESP – Campus de Marília
Orientadora

_____
Prof.ª Dr.ª Maria Cláudia Cabrini Grácio
UNESP – Campus de Marília

_____
Prof. Dr. Fernando de Assis Rodrigues
UFPA - Belém

_____
Prof. Dr. Guilherme Ataíde Dias
UFPB – João Pessoa

_____
Prof.ª Dr.ª Elaine Parra Affonso
FATEC - Campus de Presidente Prudente

Marília, 30 de março de 2023.

*Para Aparecido Vicente dos Reis e Maria Benedita dos Reis, meus pais, que foram morar junto a Deus. Não consigo expressar em palavras o que meu coração está sentindo. A humildade evidente e as durezas de nossas vidas me fizeram crescer, mas, principalmente, seus ensinamentos de simplicidade e amor me guiaram nas caminhadas. Gostaria apenas que estivessem aqui para poder abraçá-los e compartilhar este momento. Muito do que sou devo a vocês. Que Deus os tenha.*

## AGRADECIMENTOS

À minha querida e estimada orientadora, Ely Tannuri, que além de me conduzir nesta caminhada acadêmica, me proporcionou deliciosos cafés, sorrisos contagiantes e valiosos conselhos para a vida pessoal e profissional. Todas as reflexões em nossas breves, mas intensas conversas, em muito me enriqueceram. Dos medos e contratempos do trajeto, creio que rumamos à direção correta. Talvez sem chegar à perfeição, mas com a certeza de ter dado nosso melhor. Sempre me estendeu a mão, sempre compreensiva e presente, me fez entender que suas cobranças sempre foram no intuito de me ajudar. Você é uma pessoa iluminada professora, tens e terá sempre um lugar muito especial no meu coração.

Ao professor Ricardo Sant´Ana, que no andamento da jornada passou a fazer parte deste projeto. Se inteirou do assunto e inteirou a lacuna de sutilezas que nos faltavam. Confesso que os questionamentos iniciais me deixaram apreensivo, mas percebi que seus encaminhamentos instigavam e contribuíam para aperfeiçoamento da proposta. Reconheço que sua participação nos trouxe olhares bastante apurados e foram de grande valia para o aprimoramento do estudo. Obrigado por aceitar a empreitada e compartilhar conosco um sopro de vossa sabedoria.

À minha eterna namorada Aline Luqui, a qual me presenteou com o bem mais precioso até hoje e que não mede esforços para transpormos juntos os desafios da vida. Ao meu amado filho Murilo Vicente Afonso Reis, que no decurso dessa tese nasceu. A vida brindou-me com a coincidência de tê-lo acolhido em meus braços no mesmo dia em que meu pai comemorava aniversário. Não há palavras para descrever o quanto sou grato a vocês dois e não há como expressar a felicidade que sinto em tê-los ao meu lado. Simplesmente amo muito o que somos.

Aos meus irmãos, sobrinhos e sobrinhas, que são alento à falta que sinto dos que já se foram. Embora distantes pelos rumos que a vida deu a cada um, saibam que estão e estarão sempre em meus pensamentos, orações e coração. Vocês também são alicerce para minha vida e motivos para eu continuar sorrindo.

## RESUMO

Esta tese teve como finalidade elaborar uma metodologia para operacionalização da abordagem terminológica da análise de domínio. Estudos terminológicos investigam padrões das linguagens dos discursos, e suas análises baseiam-se em unidades de significação nos textos, sem esquecer do contexto sociocultural. Nessa mesma linha, a análise de domínio busca descobrir estruturas de conhecimento, padrões de linguagem e comportamento de cooperação nos domínios. A mineração de texto serve para automatizar a extração de regularidades, padrões ou tendências nos documentos em linguagem natural. Portanto, a análise de domínio propõe os objetos a serem investigados e a mineração de texto fornece os meios para as descobertas. Por isso, para desenvolver a metodologia, fez-se uso das técnicas de dedução de frequência de termos e análise por categorias temáticas, advindas do campo da linguística e automatizadas pela mineração de texto. Desenvolveu-se um fluxograma canônico, que nesta tese concebeu-se no *software* Knime. O software é composto por módulos para pré-processamento, transformação e mineração de textos para descoberta de conhecimento. A metodologia foi aplicada em 287 resumos de estudos apresentados pelo GT7 nos ENANCIBs de 2012 a 2018. A automatização proporcionou melhorias nas questões de limitação humana quanto a leitura, exploração e registro de grandes volumes de dados. Por meio da técnica de dedução de frequência de termos, foram encontradas especificidades desconhecidas na linguagem dos resumos, relacionadas à quantidade de termos que compõem as sentenças e termos mais recorrentes. Por meio do algoritmo de Alocação Latente de Dirichlet (*Latent Dirichlet Allocation* – LDA), identificou-se cinco tópicos, cada um constituído por dez palavras, que representam os principais temas do *corpus*. O algoritmo possibilitou ainda identificar *clusters* de resumos com interlocução temática. Desse modo, encontraram-se adjacências nas comunicações do grupo de autores, que mesmo sem se conhecerem, desenvolveram textos convergentes, formando uma comunidade com discursos correlatos. O algoritmo *snowball* foi empregue para realização do *stemming*, que agrupa palavras de mesmo radical, pois considera-se que tais termos possuem significado semelhante e proximidade gramatical. A radicalização reduziu o conjunto inicial de termos de 5.820 para 3.657, simplificando e limitando a quantidade, o que pode auxiliar processos de indexação, buscas, recuperação da informação e custo computacional. Os resultados foram promissores, pois conseguiu-se automatizar análises de texto e de conteúdo. Conclui-se que a metodologia pode contribuir com a comunidade científica para realização de pesquisas em linguagem natural, de busca e recuperação da informação, e para descoberta de padrões e articulações temáticas dos textos. Pode corroborar também para expansão de estudos da literatura científica fora das bases bibliográficas mais conhecidas. As temáticas da área de linguística, assim como análise de domínio e mineração de texto, possuem evidente consonância conceitual, demonstrando a pertinência da pesquisa. A metodologia proporcionou a operacionalização da abordagem terminológica de forma automatizada e em consonância a análise de domínio, pois, os algoritmos utilizados consideram indicadores em relação ao conjunto total dos textos, revelando perspectivas informacionais coletivas e não individuais.



**Palavras-chave:** metodologia; terminologia; análise de domínio; mineração de texto; literatura científica.

## ABSTRACT

This study aimed to develop a methodology for operationalizing the terminological approach to domain analysis. Terminological studies investigate language patterns of discourses, and their analyzes are based on units of meaning in texts, without forgetting the sociocultural context. Along the same lines, domain analysis seeks to discover knowledge structures, language patterns and cooperation behavior in domains. Text mining serves to automate the extraction of regularities, patterns or trends in natural language documents. Therefore, domain analysis proposes the objects to be investigated and text mining provides the means for discoveries. Therefore, in order to develop the methodology, we used the techniques of deducing the frequency of terms and analysis by thematic categories, coming from the field of linguistics and automated by text mining. A canonical flowchart was developed, which in this thesis was conceived in the Knime software. The software consists of modules for pre-processing, transforming and mining texts for knowledge discovery. The methodology was applied to 287 abstracts of studies presented by GT7 at the ENANCIBs from 2012 to 2018. Automation provided improvements in human limitation issues regarding the reading, exploration and recording of large volumes of data. Through the technique of deducing the frequency of terms, unknown specificities were found in the language of the abstracts, related to the number of terms that make up the most recurrent sentences and terms. Using the Latent Dirichlet Allocation (LDA) algorithm, five topics were identified, each consisting of ten words, representing the main themes of the corpus. The algorithm also made it possible to identify clusters of abstracts with thematic interlocution. In this way, adjacencies were found in the communications of the group of authors, who, even without knowing each other, developed converging texts, forming a community with related discourses. The snowball algorithm was used to carry out stemming, which groups words with the same root, since it is considered that such terms have similar meaning and grammatical proximity. Radicalization reduced the initial set of terms from 5.820 to 3.657, simplifying and limiting the quantity, which can help indexing processes, searches, information retrieval and computational cost. The results were promising, as it was possible to automate text and content analysis. It is concluded that the methodology can contribute to the scientific community to carry out research in natural language, to search and retrieve information, and to discover patterns and thematic articulations of the texts. It can also corroborate the expansion of studies in the scientific literature outside the most well-known bibliographic bases. The themes in the area of linguistics, as well as domain analysis and text mining, have an evident conceptual consonance, demonstrating the pertinence of the research. The methodology provided the operationalization of the terminological approach in an automated way and in line with the domain analysis, since the algorithms used consider indicators in relation to the total set of texts, revealing collective and not individual informational perspectives.



**Keywords:** methodology; terminology; domain analysis; text mining; scientific literature.

## LISTA DE ILUSTRAÇÕES

**SUMÁRIO**

## PREFÁCIO

As páginas que antecedem a introdução deste estudo compõem um elemento pré-textual, que tem por finalidade mostrar ao leitor uma síntese de minha trajetória e, ao mesmo tempo, apresentar as motivações que me direcionaram à temática para elaboração da tese. É um breve apanhado de minha caminhada, iniciada após aprovação no meu primeiro vestibular, no qual encerro justamente expondo o que aprendi e o que sinto ao concluir mais um passo muito importante em minha vida.

No ano de 1996, iniciei minha graduação na Faculdade de Tecnologia de São Paulo (FATEC), extensão de Ourinhos, tendo concluído o curso superior de Tecnologia em Processamento de Dados no primeiro semestre de 2001. Esse momento foi um misto de sentimentos, em que a felicidade da formatura contrasta com o vazio dos dias seguintes, pois há pouco eu era estudante e agora seria desempregado. Passei um tempo trabalhando em um quartinho emprestado nos fundos da casa um colega. Fazia manutenção, vendia peças e computadores e dividia os lucros com ele.

Meu primeiro contato com sala de aula foi em agosto de 2001, quando consegui emprego e comecei a trabalhar em uma escola de informática. Nessa escola, eu ministrava aulas de montagem e manutenção de computadores, instalação de sistemas operacionais e *software* aplicativos, informática básica e linguagem de programação. Nos anos de 2002 e 2003, fiz especialização em Redes de Computadores, na Universidade Norte do Paraná (UNOPAR), na cidade de Londrina.

Ainda durante o ano de 2002, mesmo trabalhando na escola de informática, fui contratado como auxiliar de redes de computadores, na Faculdade de Ibaiti (PR) (FEATI), para trabalhar no período noturno. Eu era responsável por dar apoio aos usuários e realizar configuração, instalação de sistemas operacionais, *software* aplicativos, bem como todo tipo de auxílio ao funcionamento e disponibilidade dos equipamentos e da rede, a fim de manter o bom andamento das aulas nos laboratórios de informática e dos setores administrativos.

No final de 2002, recebi uma proposta do diretor da FEATI para me dedicar somente à faculdade, de modo que passei a trabalhar apenas na instituição de ensino superior. Em 2003, por conta do desligamento do administrador de redes, novamente o diretor fez uma proposta, dessa vez, para que eu assumisse o cargo vago. Era uma oportunidade bastante interessante, pois, como estava no decurso da especialização,

eu poderia pôr em prática o que aprendia na teoria. Passei, então, a ter contato mais direto com instalação e configuração de servidores, de equipamentos de redes, além de ter autonomia para propor a implantação e atualização de serviços como DNS, firewalls, servidores de e-mail, de páginas de internet, proxy etc.

Atuei como administrador de redes na FEATI até o ano de 2012. Em 2005, a coordenação do curso de Sistemas de Informação perguntou-me sobre o interesse em lecionar, explicando que no semestre seguinte, que se iniciaria em uma semana, o curso precisaria de professor para as disciplinas de *Redes de Computadores I* e *II*, consecutivamente. Assim, iniciou-se minha caminhada como professor, a qual se segue até a presente data. Nesse intervalo de 2005 a 2012, além de administrar a rede, passei a lecionar diversas outras disciplinas do curso, sempre relacionadas a sistemas operacionais e redes. Também me tornei professor de informática básica de outros cursos oferecidos na instituição, como as graduações em Administração e Pedagogia, e nos cursos técnicos em Segurança do Trabalho e Enfermagem.

Foi um período de muito crescimento pessoal e profissional. Enquanto professor, o contato com as pessoas me fez enxergar além do *hardware* e do *software.* Compreendi que dividir conhecimento era, sobretudo, oportunizar alguém a mudar de vida. Isso passou a fazer mais sentido quando alguns alunos aprendiam a enviar um simples e-mail. O que para muitos era algo rotineiro, para os menos favorecidos era algo extraordinário. Percebi uma lacuna de conhecimento muito grande entre as classes sociais e passei a me interessar bastante pelo universo da educação.

No meio dessa caminhada, entre 2008 e 2009, fiz o curso de especialização MBA em Gestão de Negócios e Pessoas, oferecido pela própria FEATI. Diante das facilidades de não precisar se deslocar à outra cidade e dos incentivos financeiros oferecidos aos funcionários, tratei a situação como uma grande oportunidade. Embora profissionalmente não tenha seguido nessa área, a experiência foi muito enriquecedora, pois, tive contato com vários apontamentos relacionando tecnologia e sociedade, o que também despertou meu interesse sobre esses temas.

Ainda nessa jornada, em janeiro de 2009, eu assumi o cargo de professor de informática do Estado do Paraná. Passei a lecionar matérias da área nos cursos técnicos oferecidos pelo Colégio Estadual Aldo Dallago (CEAD), da cidade de Ibaiti, ficando no quadro próprio do magistério do estado até o ano de 2012. No período de 2009 a 2012, por conta da quantidade de horas trabalhadas, tive que pedir diminuição das quarenta horas da atribuição de administrador de redes, passando a exercer vinte

horas nessa função, vinte horas como professor da faculdade e vinte horas como professor do Estado. Eram três turnos de labuta diária.

No mês de fevereiro de 2012, tomei posse como professor em dedicação exclusiva do ensino básico, técnico e tecnológico do Instituto Federal do Paraná (IFPR), campus Jacarezinho, o qual se faz meu endereço profissional até a presente data. A condição de dedicação integral ao IFPR fez com que eu pedisse exoneração do cargo de professor do Estado e o desligamento das funções que exercia na faculdade FEATI. Entre os meses de abril de 2013 e março de 2014, como aprimoramento para docência, concluí minha licenciatura com habilitação em Informática, no Programa Especial de Formação Pedagógica, oferecido pela Universidade Tecnológica Federal do Paraná (UTFPR), campus de Cornélio Procópio.

Ainda no ano de 2013, entrei como aluno especial no Programa de Pós-graduação em Ciência da Informação na Universidade Estadual de Londrina (PR) (UEL), vindo a ingressar como aluno regular do mestrado em março de 2014. No programa, trabalhei com Análise de Redes Sociais (ARS), investigando prioritariamente quais plataformas a geração Z utilizava para formação de redes de compartilhamento de informação. Com a ARS, pude verificar a interligação entre informação, tecnologia e sociedade, ao mapear o percurso informacional entre os agentes/atores em um ambiente específico. A conclusão dessa etapa se deu em dezembro de 2015, ao defender a dissertação intitulada Plataformas infocomunicacionais e o compartilhamento da informação: estudo da Geração Z.

No meio do mestrado, confesso que pensei em desistir, pois 2014 foi um ano de tristezas e alegrias. Após uma queda, meu pai bateu a cabeça e passou cerca de um mês na UTI, vindo a falecer no final de setembro. Para complementar, meu casamento estava planejado para o início de novembro e não havia mais como retroceder. Mesmo com a dor, meu matrimônio foi um acontecimento de muita felicidade, que me deu forças a continuar. Contudo, no meio de 2015, descobrimos que minha mãe estava doente; e em estágio bastante avançado. Foram quase dozes meses de correrias e viagens intermináveis na tentativa de reverter o quadro, até que o mais triste aconteceu. Eu e meus irmãos perdíamos mais um pedaço de nós.

Após esse período de vida complicado, por conta dessas adversidades, fiquei bastante desanimado em dar continuidade aos estudos e acabei me ausentando do meio, dedicando-me apenas à sala da aula. Somente no ano de 2018, voltei a me interessar novamente pela esfera acadêmica e consegui ingressar como aluno regular

no Programa de Pós-graduação em Ciência da Informação da Universidade Estadual Paulista “Júlio de Mesquita Filho” (UNESP), campus Marília-SP. Essa tese é fruto do percurso aqui descrito, o qual tenta abarcar as nuances de minha trajetória profissional e os percalços da vida. Além da formação inicial na área tecnológica, dos direcionamentos seguintes como professor, das valiosas trocas de experiências dessa profissão e dos encaminhamentos acadêmicos para o campo da CI, algumas disciplinas no programa de doutorado me conduziram ao tema desta tese.

A disciplina Questões Bibliométricas em Produção e Organização da Informação fizeram entender a importância das investigações das produções científicas e apresentaram as dificuldades na execução de alguns estudos. As disciplinas Usuário e produção da informação sob a perspectiva da Análise de Domínio e Metateoria e Análise de Domínio em Organização do Conhecimento despertaram a curiosidade ante o aspecto metodológico dos estudos analíticos de domínio. O que era um indício para o tema proposto ficou translúcido ao participar da disciplina Inteligência Artificial e Ciência de Dados aplicadas à Ciência da Informação.

E assim nasceu este projeto, que une as áreas às quais tenho me dedicado e que penso coroar essa minha caminhada pessoal e acadêmica. A presente tese trata de uma proposição de metodologia que faz uso de aparatos tecnológicos, para automatizar processos e viabilizar a execução de tarefas complexas ante as limitações humanas em ler, armazenar e correlacionar padrões em textos. Procurei empregar a tecnologia como facilitadora de atividades investigativas aos pesquisadores (seres humanos), principalmente para visualização de perspectivas coletivas (análise de domínio) e não apenas individuais para estudos na área da CI.

Confesso que meu sentimento é de muita felicidade. Creio ter conseguido aliar os conhecimentos da graduação, das especializações, do mestrado e doutorado. Posso debater nas áreas em que atuo sobre o poder da informação e das interações humanas. Devo ponderar a tecnologia como facilitadora e não como solução para tudo. Para que a equidade seja alcançada, todos devem ter os mesmos direitos. Temos como missão demonstrar aos menos favorecidos que, informação e educação dão ao indivíduo o poder de transformar o ambiente ao seu redor, que existem possibilidades de mudança de vida, e isso é o principal.

# 1 INTRODUÇÃO

O dinamismo dos meios de comunicação amparados pela tecnologia alavancou novas perspectivas no trato da informação, não só por uma questão de mudança de suporte, mas principalmente pelas possibilidades de alcance que os meios digitais são capazes de prover. Para Capurro e Hjørland (2007, p. 149), a definição de informação[1] como conhecimento comunicado ocupa um papel nuclear na sociedade, pois, "o desenvolvimento e a disseminação do uso de redes de computadores desde a segunda grande guerra mundial e a emergência da Ciência da Informação como uma disciplina nos anos 50 são evidências disso."

Os autores entendem que o conhecimento e a sua comunicação são fenômenos básicos de toda sociedade humana, mas o que caracteriza a chamada sociedade da informação foi o surgimento da tecnologia da informação e seus impactos a nível global. "É lugar comum considerar-se a informação como condição básica para o desenvolvimento econômico juntamente com o capital, o trabalho e a matéria-prima, mas o que torna a informação especialmente significativa na atualidade é sua natureza digital." (CAPURRO; HJØRLAND, 2007, p. 149).

Sant´Ana (2020, p. 12-13) traz uma reflexão bastante contemporânea e pertinente ao enfatizar o determinismo imperativo dos enredos computacionais na atualidade como um fenômeno universal e irreversível, argumentando sobre o imbricamento entre ser humano e tecnologia, que "nos leva a construir percepções do real cada vez mais dependentes da intermediação maquínica. Nossa relação com a informação não foge à regra e mais, tende a ser a marca determinante dos paradigmas socioculturais predominantes neste início de milênio."

As mudanças tecnológicas vivenciadas também contribuíram para que cada vez mais dados não estruturados[2] tenham que ser recuperados, pois, na cultura moderna, o texto é o veículo mais comum para a troca formal de informações. (WITTEN, 2004). Tais textos, disponíveis em páginas de internet, redes sociais ou

---

[1] Este estudo adota a definição de informação como conhecimento comunicado, conforme Capurro e Hjørland (2007, p. 149).

[2] Dados não estruturados referem-se à informação disponível em plataformas digitais, sem qualquer estrutura organizacional quanto a forma de armazenamento ou recuperação, como por exemplo, uma página de internet ou um arquivo em formato pdf. Em contrapartida, os dados estruturados podem ser armazenados em um banco de dados relacional, por exemplo. Tais definições de dados são utilizadas no âmbito da Tecnologia em Sistemas de Informação, da área de Informática, notadamente na disciplina de Banco de Dados.

base de dados, podem conter informações inicialmente desconhecidas, mas potencialmente valiosas do ponto de vista comercial e científico. Essa vasta quantidade de materiais *online* propicia o acesso à literatura científica, independentemente da localização de seus pretensos leitores.

Destaque-se que, apesar da rotulagem atribuída como dados não estruturados, para o campo da Linguística, por exemplo, por mais simples que seja um documento textual, ele possui algum tipo de estrutura. Segundo Feldman e Sanger (2007), mesmo um texto que seja considerado limitado, pode ser rico semântica e sintaticamente, ainda que tais paralelismos não estejam explícitos em seu conteúdo. Os elementos tipográficos (pontuação, maiúsculo/minúsculo, números, caracteres especiais, sublinhados, espaçamentos etc.) indicam subcomponentes relevantes como parágrafos, títulos, autores, entre outros, e a própria sequência de palavras pode ser estruturalmente significativa para o documento.

Ao se transpor isso para a esfera acadêmica e reportar aos estudos científicos, a maioria das publicações são classificadas como documentos em formato livre ou fracamente estruturados, pois oferecem relativamente pouco em termos de indicadores tipográficos ou marcações para denotar sua estrutura. Ressalta-se que não só a tecnologia ampliou o acesso aos estudos provenientes dos ambientes acadêmicos, outros fatores podem ser considerados significativos para o estabelecimento da literatura científica no país, como o surgimento da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), a qual favoreceu de forma relevante o crescimento dos programas de pós-graduação.

A CAPES tem papel considerável para a ciência no país, pois, o órgão contribuiu para perpetuar os programas de especialização e alavancar os eventos e publicações científicas. Muito do que já se produziu e ainda se produz de literatura nessa esfera teve a cooperação da entidade. Ressalte-se que o Brasil, em 40 anos, teve um crescimento de mais de 800% no que diz respeito à pós-graduação, passando de 699 cursos em 1976 para 6131 no ano de 2016. Sobretudo, no final dos anos 1990 (1462 cursos), o ritmo de crescimento fica mais evidente, saltando para 1791, no ano de 1995, 2621, em 2000, 3224, em 2005, e chegando a 4757, no ano de 2010. (NOBRE; FREITAS, 2017).

As ações da CAPES são pautadas em programas voltados para: investimento na formação acadêmica (no Brasil e em outros países) com bolsas de estudos e de Fomento aos Cursos; avaliação da pós-graduação com abertura de cursos *stricto*

*sensu*, reconhecimento e renovação de funcionamento; promoção de cooperação científica bilateral com acordos entre universidades nacionais e estrangeiras, financiando intercambio discente/docente, proximidade curricular e aproveitamento de créditos; disponibilização de acesso e divulgação da produção científica por meio do Portal de Periódicos, que possui importantes Revistas e Bases de Dados Científicas (nacionais e internacionais) indexadas. (BRASIL, 2022, n.p.).

Para este estudo, tem-se especial atenção ao programa de acesso e divulgação da produção acadêmica, pois, com a expansão dos programas de mestrado e doutorado e aumento das publicações de estudos desenvolvidos nas universidades, muitos pesquisadores passaram a investigar a própria produção científica. Além da disponibilização do Portal de Periódicos da CAPES para acesso a revistas e bases de dados renomadas, o empenho dos movimentos de bases bibliográficas de acesso aberto[3] também merece ser destacado. Esse conjunto de fenômenos favoreceu a disseminação das informações geradas nas universidades, contribuindo consideravelmente para expansão de acesso às literaturas científicas internacional e brasileira.

Frente o aumento do número de usuários de computadores e internet, das inovações tecnológicas e avanços no campo científico, um cenário de inter-relações parece emergir, favorecendo o compartilhamento dos estudos acadêmicos. (DAVIES, 1989). Além do processo de interação nas plataformas infocomunicacionais (REIS; TOMAÉL, 2016), os resultados dos eventos científicos passaram a ter como suporte os meios eletrônicos, aumentando significativamente a quantidade e disponibilidade de dados no formato *online.* (SANT´ANA, 2020).

Tal panorama conflui às observações de Macias-Chapula (1998), que pondera sobre a questão de publicar as pesquisas ser um compromisso dos cientistas; e que o conhecimento produzido deve ser acessível à comunidade. Para o autor, as publicações propiciariam um cenário de troca de conhecimento entre grupos de pesquisadores (intragrupos; intergrupos). Como apontado, desde essa época havia inclinação para estudos sobre a produção científica (MACIAS-CHAPULA, 1998), os quais foram se ampliando, e passaram a caracterizar a importância desse tipo de pesquisa. (OLIVEIRA; GRÁCIO, 2009; OLIVEIRA, 2013; SANT´ANA, 2020).

---

[3] Acesso Aberto (AA): publicações disponíveis online a todos, sem nenhum custo ou restrições limitadas a respeito de reutilização.

Estudiosos de diferentes países têm se dedicado a investigar as produções acadêmicas, principalmente literaturas indexadas em bases bibliográficas mundialmente reconhecidas como *Web of Science (WoS), Scopus* e *LISA.* (LU; WOLFRAM, 2012; AMORIM; CAFÉ, 2014, 2016; ROSAS; GRÁCIO, 2015; GUIMARÃES *et al.*, 2017; MANHIQUE; CASARIN, 2018; HSU; LI, 2019; TRABADELA-ROBLES *et al.*, 2020; MOKHTARPOUR, KHASSEH, 2021). No entanto, encontram certa dificuldade de acesso a literatura de países apontados como periféricos, caso do Brasil, já que a tônica nas bases mencionadas é a indexação de revistas internacionais. (OLIVEIRA; GRÁCIO, 2009).

Diferentes pesquisas são realizadas com dados extraídos dessas bases, e como bem evidenciado por Oliveira e Grácio (2009), estudos da literatura científica no país ficam prejudicados. Por isso, compreendendo a produção científica como transmissora de conhecimento e reputando as publicações o modo que conhecimento chega à sociedade e, ainda, creditando à ciência o papel de agente difusora de melhorias para as comunidades, entende-se que instrumentos que auxiliem a investigação da literatura proveniente de eventos científicos não indexados não deveriam ser descartados. As comunicações desses eventos, mesmo não integrados as bases bibliográficas mais conhecidas, são fontes importantes de conhecimento.

O processo de disponibilização das pesquisas científicas é interessante, pois torna possível o alcance de novos conhecimentos, contudo, encontrar informações que levem ao conhecimento não é tarefa simples. (DAVIES, 1989). Se, por um lado, há facilidade de acesso, por outro, dificulta-se o refinamento e visualização das informações. Visto que a capacidade humana de exploração e registro de grandes volumes de dados ainda é limitada, torna-se inviável para a maioria das pessoas lerem os textos por si só. (WITTEN, 2004). Esse número elevado de informações sem nenhuma forma eficiente de tratá-las, denominou-se sobrecarga de informação. (DAVIES, 1989; GOLDSCHMIDT; PASSOS, 2005).

Para amenizar a limitação humana em relação ao tratamento da informação, ferramentas e técnicas foram criadas, implantadas e continuam sendo aperfeiçoadas, como estudos estatísticos aplicados no processamento e análises da informação. Informações estruturadas em banco de dados – classificadas, indexadas, organizadas e com ferramentas sofisticadas e rápidas para busca e recuperação da informação – têm sido objeto de estudos, a fim de se extraírem conhecimentos para apoio à tomada

de decisão. Igualmente, esses processos são empregados a dados não estruturados, como no caso das comunicações científicas.

Devido ao tamanho de coleções documentais como *LISA*, *WoS* e *Scopus*, ou qualquer outra, mesmo com recortes do universo da pesquisa, tentativas manuais de correlacionar os documentos seriam processos complexos. Por isso, entende-se que métodos automatizados presentes na Mineração de Texto (MT) podem contribuir tanto na rapidez quanto na eficiência das investigações. Em alguns casos, "não são apenas um complemento útil, mas um requisito básico para que os pesquisadores possam, de maneira prática, reconhecer padrões sutis em muitos documentos em linguagem natural." (FELDMAN; SANGER, 2007, p. 2). Em complemento, Aranha e Passos (2006, p. 1) definem que a "mineração de textos consiste em extrair regularidades, padrões[4] ou tendências de grandes volumes de textos em linguagem natural, normalmente, para objetivos específicos." Por meio de técnicas de mineração de texto, torna-se possível realizar a busca de conhecimento útil, em dados[5] não estruturados ou semiestruturados.

A Ciência da Informação (CI), que tem como escopo investigar a origem, organização, armazenamento, transformação e transmissão da informação em meios naturais ou artificiais (BORKO, 1968), poderia valer-se de técnicas de mineração de texto que se adequem a procedimentos de pesquisa da área. Por exemplo, a aplicação de processos computacionais parece cabível em investigações que busquem os principais termos de um conjunto de textos em determinado domínio.

Na CI, existem subáreas que direcionam interesses para temáticas específicas, como no caso dos Estudos Métricos da Informação (EMI), que analisam produções científicas regionais ou institucionais por meio de indicadores bibliométricos. (OLIVERIA, 2013). Uma subárea que vem obtendo visibilidade considerável na CI é a Análise de Domínio (AD)[6], que se destaca como aporte metodológico do paradigma sociocognitivo (HJØRLAND; ALBRECHTSEN, 1995), e inclusive tem a bibliometria

---

[4] Neste estudo adota-se a definição de padrão conforme Fayyad, Piatetsky-Shapiro e Smyth (1996), como uma expressão que descreve um subconjunto de dados ou um modelo aplicável ao subconjunto. Extrair um padrão também designa ajustar um modelo aos dados; encontrar estrutura a partir de dados; ou, em geral, fazer qualquer descrição de alto nível de um conjunto de dados.

[5] Dados são um conjunto de fatos, por exemplo, casos em um banco de dados. (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996). Nesta pesquisa dados referem-se aos resumos das comunicações científicas, ou ao conjunto de termos constantes em cada resumo.

[6] A Análise de Domínio (AD) será mais detalhada no decorrer desta pesquisa.

como uma das abordagens (HJØRLAND, 2002; OLIVEIRA, 2013) para operacionalização das análises.

Outra abordagem dos estudos analíticos de domínio (AD) é a terminológica[7] (HJØRLAND, 2002), vista como bastante cabível na área. Tem-se que levantamentos terminológicos dos discursos[8] refletem o pensamento dos autores. O campo da Linguística se entrelaça com diversas áreas, incluindo a CI, já que a linguagem é o modo como ocorrem as comunicações. Segundo Hjørland e Albrechtsen (1995, p. 420), "o sistema de linguagem não deve ser basicamente entendido como algo autodependente e isolado, mas como uma ferramenta para a interação humana no material, social, e mundo psicológico."

Adentrando, então, no campo da Linguística, mesmo que de maneira pouco aprofundada, a Socioterminologia de Gaudin (1993, 2014) e Faulstich (2006) e a Teoria Comunicativa da Terminologia (TCT) de Cabré (2005) apontam que, ao tomar os textos como elementos principais nas pesquisas, os discursos ali presentes estão carregados de impressões sociológicas (dimensão social) dos seus autores e que as expressões[9] utilizadas nesses discursos manifestam o funcionamento[10] e não a função[11] dos vocábulos dentro dos textos.

Este estudo atém-se a dois métodos da área de linguagens. O método de dedução de frequência é mais voltado ao texto propriamente e contribui para uma espécie de instrumentalização matemática para se fazer o tratamento da informação textual. Já o método de análise por categorias temáticas está mais ligado à análise de conteúdo; e pode ser utilizada para verificar o sentido do texto ou fazer comparações entre eles. (PÊCHEUX, 1997). Em complemento, Marian (2015) menciona a subárea denominada Linguística de Corpus (LC), que utiliza tecnologia computacional para trabalhar a linguagem ante aspectos probabilísticos, remetendo que a presença dos termos nos discursos não é algo casual. Portanto, entende-se que procedimentos para realização de levantamentos terminológicos podem viabilizar as análises de texto e conteúdo.

---

[7] Terminológico(a) e terminologia serão utilizados como sinônimos do levantamento terminológico aqui proposto, que será feito com as técnicas de dedução de frequência de termos e análise por categorias temáticas.

[8] Discurso neste estudo, entende-se a produção escrita, a linguagem utilizada pelos autores nas comunicações científicas.

[9] Entenda-se, nesta pesquisa, expressão, vocábulo e palavra como sinônimos de termo.

[10] Algo com efeito de funcionar, ação, atividade.

[11] Utilizado como algo com utilidade ou uso.

Os métodos de dedução de frequência de termos e o de análise por categorias temáticas podem ser empregados com algoritmos existentes na mineração de texto; e auxiliar estudos em subáreas da CI, como por exemplo, realizar a busca dos termos mais utilizados em um conjunto de discursos, por meio de ferramentas computacionais. (MARIAN, 2015). Diante das possibilidades de automatizar processos, para apontar entrelaçamento informacional e proporcionar uma visão mais relacional de um *corpus* textual, é que se entende como oportuna a utilização de técnicas da Linguística e Mineração de Texto para realização de estudos na área CI. Ressalte-se que, mesmo perpassando por outros campos, esta pesquisa não pretende se aprofundar a conceitos de outras áreas, pois, em muitos momentos serão apenas fragmentos teóricos para compreensão do projeto.

Para assimilar o imbricamento entre o campo da mineração de texto, das linguagens e da ciência da informação, retrocede-se brevemente ao princípio da Análise de Domínio (AD), pois, é por meio dela que a CI se insere ao contexto desta tese. A literatura menciona três paradigmas epistemológicos que fizeram ou fazem parte dos estudos na área. O paradigma físico situa-se em uma epistemologia material, na qual a informação tem caráter puramente técnico e quantitativo, transmissível por computadores e sem preocupação semântica. (SARACEVIC, 1995; CAPURRO, 2003). O paradigma cognitivo detém seu foco no usuário e em situações ou necessidades de busca informacional. São estudos da mente, do intelecto e do modelo tradicional de recuperar a informação sob o prisma cognitivo individual. (ALMEIDA *et al.*, 2007).

O paradigma social ou sociocognitivo tem um olhar mais amplo sobre o enredo informacional e considera elementos subjetivos dos indivíduos, suas perspectivas, interesses e contextos sociais, distanciando-se dos modelos específicos de linguagem que representam o conhecimento (paradigma cognitivo) e de algoritmos considerados ideais (paradigma físico) para recuperação da informação. (CAPURRO, 2003). É nessa perspectiva que emergem os estudos analíticos de domínio na CI, com o propósito de contextualizar e esclarecer os princípios teóricos do paradigma sociocognitivo ou social, sob a ótica da construção coletiva de conhecimento.

Nos estudos de análise de domínio, as disciplinas, as profissões e o domínio são encarados como a junção de elementos que formam grupos de trabalho, comunidades discursivas ou colégios invisíveis com laços comuns entre seus participantes, e não mais como participações individuais. Para López-Huertas (2015,

p. 571), “a análise de domínio foi concebida para descobrir estruturas de conhecimento, dinâmica, padrões de linguagem e comunicação e comportamento de cooperação de domínios especializados.” Se reportarmos sobre o papel da mineração de texto como um instrumento para extração de regularidades, padrões ou tendências nos documentos em linguagem natural (ARANHA; PASSOS, 2006), encontra-se perceptível convergência entre os dois métodos, entretanto, um propõe a investigação e o outro fornece os meios para se realizarem as descobertas.

Consoante ao entendimento de que o conhecimento não surge da unicidade, Hjørland e Albrechtsen (1995, p. 400) destacam que “a melhor maneira de entender a informação é estudar os domínios do conhecimento como comunidades de pensamento ou discurso, que são partes da divisão de trabalho da sociedade.” Para os autores, a linguagem não é apenas “rótulos separados”, mas discursos que abarcam impressões históricas, sociais e culturais de seus escritores, nos quais a literatura produzida expressa a realidade (contextual – sociocultural), típica dos domínios de conhecimento.

No campo da Linguística, nessa mesma direção estão as teorias terminológicas de Gaudan (1993; 2014), Cabré (2005) e Faulstich (2006), que inferem que os discursos carregam as impressões sociológicas de seus autores. Novamente encontrou-se convergência teórica entre os assuntos desta tese, mas, agora, entre o que apregoam a análise de domínio e os preceitos da terminologia. Desse modo, entende-se que técnicas da mineração de texto podem auxiliar estudos em que se busque visões mais abrangentes e correlacionais de linguagem e discurso, com vistas a resultados mais holísticos e menos individualizados.

Portanto, diante do exposto, em relação ao aumento expressivo das comunicações científicas, alavancadas pelos estímulos aos programas de pós-graduação e incentivos institucionais efetuados pela CAPES; e ainda pelo grau de importância dado aos estudos da literatura científica, que amparada pela tecnologia, tem disseminação e acesso cada vez mais rápido, mas que, devido à abundância de informações, dificultam-se os processos investigativos, é que se propõe uma nova metodologia para desenvolvimento de estudos das comunicações científicas e pesquisas acadêmicas direcionada para o campo da CI, que abarca técnicas da área de linguagem, empregues por meio de algoritmos de mineração de texto. Doravante, as comunicações científicas poderão ser entendidas simplesmente como resumos, visto que esta pesquisa tratará especificamente dessa seção das publicações.

## 1.1 Questão de pesquisa

O panorama até aqui apresentado demonstra entrelaçamento das proposições, remetendo a algumas reflexões. Assim como a tecnologia influencia a sociedade em suas condutas (CASTELLS, 2005; SANT´ANA, 2020), como por exemplo, acesso e leitura de livros no formato digital, no meio científico não é diferente, pois, o contexto dos avanços tecnológicos também se faz presente na área acadêmica. Igualmente, a implementação de políticas de aperfeiçoamento de pessoal no país (NOBRE; FREITAS, 2017; BRASIL, 2022, n.p.) corroborou a explosão informacional no meio acadêmico, retrato de uma sociedade *online* e em rede (CASTELLS, 2005), que mesmo sem muito perceber, integra essas formalizações contemporâneas.

Contextos que podem ser encarados como campo fértil para a CI, dado seu objeto de estudo – informação registrada – agora também *online* e acessível. Vertentes da área que pesquisam publicações científicas sob vários aspectos podem aproveitar-se desse quadro para se desenvolverem. Assim como nas análises bibliométricas, outras abordagens dos estudos analíticos de domínio, que atuam sob a égide relacional de comunidades científicas ou pela dinâmica informacional do conhecimento presente nos discursos, podem utilizar os dados para verificar as interações entre os membros de um grupo (autores), ou entre um grupo e a informação por perspectivas contextuais, históricas e culturais, pressupostos do paradigma social e da análise de domínio.

Tem-se, portanto, um cenário promissor para execução de levantamentos terminológicos (HJØRLAND, 2002) em *corpora* textuais, com viés mais sociocognitivo. Cenário que está em conformidade com a análise de domínio (HJØRLAND; ALBRECHTSEN, 1995), com a socioterminologia (GAUDIN, 1993, 2014; FAULSTICH, 2006), além da visão do funcionamento dos termos e não apenas da presença dos vocábulos nos textos. (CABRÉ, 2005). Vale ressaltar que, esse levantamento pode ser estruturado com os métodos de dedução de frequência de termos e análise por classificação temática, aplicados com técnicas de mineração de texto. Nesse sentido, a Linguística de Corpus[12] (MARIAN, 2015) pode auxiliar estudos do campo da CI.

Destaque-se, dessa forma, os problemas até aqui observados: grande quantidade de documentos/informações sem um procedimento eficiente de

[12] Definição mais detalhada no fim da página 23.

tratamento cria sobrecarga de informações, dificultando encontrar padrões nos textos (DAVIES, 1989; GOLDSCHMIDT; PASSOS, 2005); capacidade limitada do ser humano em refinar, visualizar, explorar, registrar e correlacionar informações (WITTEN, 2004); manipulação e registro dos dados coletados feitos manualmente (SMIRAGLIA, 2013; ROSAS; GRÁCIO, 2015; MANHIQUE; CASARIN, 2018; HSU; LI, 2019; JOO; OH, 2019); dificuldade de estudos fora de bases bibliográficas reconhecidas, por não existirem ferramentas ou métodos adequados; comprometimento de pesquisas em países periféricos como o Brasil (OLIVEIRA; GRÁCIO, 2009); processos de análise da literatura científica sem automatização (GRÁCIO, 2020).

Desse modo, a questão da pesquisa pode ser anunciada ao relacionar-se as circunstâncias expressas. Ao considerar que procedimentos da mineração de texto proporcionam encontrar informações potencialmente úteis, tais como padrões de linguagem, características textuais e temáticas nos discursos de determinado grupo de autores, indaga-se: qual a contribuição que as técnicas de mineração de texto podem oferecer para otimizar investigações de análise de domínio na perspectiva da abordagem terminológica?

## 1.2 Relevância e justificativa

Dados complexos exigem mais do que simples avaliações e só podem ser entendidos com análises mais profundas, o que confere certo grau de dificuldade da percepção humana, como por exemplo, ligações informacionais subjacentes e suas relações em um conjunto de textos. Pesquisadores precisam confiar em suas anotações, documentos e memórias para encontrar fatos interessantes em inúmeros documentos (URBIZAGASTEGUI-ALVARADO, 2021). Tais alegações ilustram a necessidade de se encontrar meios mais eficientes e menos atribulados, que permitam realizar investigações mais abrangentes em *corpora* documentais.

Isto posto, tem-se como tese que as técnicas de mineração de texto podem contribuir de forma significativa para a abordagem terminológica da AD, ao propiciar instrumentalização de processos de limpeza, transformação e extração de padrões de escrita de forma automatizada, a partir de um *corpus* textual qualquer. Para subsidiar as análises de texto e de conteúdo por meio da abordagem terminológica, o fluxograma de implementação da metodologia proposta nesta pesquisa deverá valer-

se de algoritmos que permitam aplicar as técnicas de dedução de frequência de termos e de análise por categorização temática, advindos da área de Linguística. Entende-se que tais proposições viabilizariam investigações mais específicas do quantitativo de termos e das disposições e vínculos deles em relação as temáticas constantes no *corpus*, propiciando assim a operacionalização da abordagem terminológica para análises de domínios.

As hipóteses levantadas em relação aos problemas são de que os mecanismos da mineração de textos podem auxiliar a realização de pesquisas em corpus textuais. Ao automatizar os processos de manipulação dos dados, isso propiciaria maior agilidade e diminuiria a carga de trabalho do ser humano. Os métodos de dedução de frequência de termos e de análise por categorização temática podem ser alcançados com algoritmos da mineração de textos e proporcionar a descoberta de informações não explícitas, como padrões de linguagem e estruturas de conhecimento correlacionais presentes nos discursos.

Estudos da terminologia, de linguagens, de temáticas, padrões de escrita ou discurso, ou mesmo de sistemas de informação de um domínio, muitas vezes são feitos de forma manual (SMIRAGLIA, 2013; HSU; LI, 2019; JOO; OH, 2019; GRÁCIO, 2020) ou em conjunto com planilhas eletrônicas (SMIRAGLIA, 2013; ROSAS; GRÁCIO, 2015; MANHIQUE, CASARIN, 2018), gerando enorme volume de trabalho. A relevância social de um estudo está em repensar ou reformular procedimentos que contribuam com a sociedade.

Nesse caso, a presente proposta possibilita que os pesquisadores dispensem menos tempo para questões como o tratamento e transformação dos dados, e se dediquem mais à análise, permitindo-lhes um olhar mais cuidadoso em relação aos resultados, o que poderia representar um salto de qualidade aos estudos. Outra contribuição para a sociedade é que a metodologia pode servir como um modelo ou uma base canônica, ou seja, ser implementada em qualquer *software* que realize as tarefas demonstradas, não se prendendo ao aplicativo aqui utilizado. Outrossim, existe a possibilidade de interligação de novos módulos, possibilitando ajustar-se a outros domínios do conhecimento, podendo auxiliar estudos com diferentes demandas sociais.

Existem *software* para mineração e visualização de textos que fazem a varredura completa em documentos e fornecem uma visão geral do acervo. São *software* que realizam ordenação de padrões recorrentes por frequências ou por

comprimento de termos, busca de padrões de palavras múltiplas e, ainda, comparações de características textuais em busca de padrões terminológicos no *corpus* textual. Parte significativa da literatura científica está disponível na forma digital e *online,* indexada em bases bibliográficas ou em arquivos textos não estruturados. Diante das possibilidades oferecidas por *software* de mineração de texto e dos suportes em que as comunicações científicas se encontram, o surgimento de novos processos para abordagens sistemáticas e estudos da informação podem apresentar-se como aporte para diferentes áreas, tendo a mineração de texto como um instrumento de apoio de grande valia. (FELDMAN; DAGAN, 1995; GOLDSCHMIDT; PASSOS, 2005; URBIZAGASTEGUI-ALVARADO, 2021).

Portanto, do ponto de vista científico, esta pesquisa tem relevância por desvelar procedimentos de manipulação de textos e apontar características discursivas ante uma perspectiva mais abrangente e contextual, como preconizado pela análise de domínio. Conforme Witten (2004), mineradores de texto ou pesquisadores da informação podem descobrir novas hipóteses científicas apenas analisando a literatura. É relevante também pelo caráter de raridade e por propiciar um ângulo diferente de pesquisa na área, no qual é possível se extrair conteúdos profícuos dos textos sem inferências iniciais (das bases/documentos para os pesquisadores) e não apenas realizar a busca de termos pré-determinados (dos pesquisadores para as bases/documentos).

Ainda no campo científico, poderá contribuir para expansão de estudos em publicações que não estejam indexadas nas bases bibliográficas mais conhecidas. Por exemplo, Oliveira e Grácio (2009) mencionam a dificuldade de se obter dados da literatura brasileira sobre produção científica nas bases internacionais. A automatização pode assistir a área também em segmentos que ainda careçam de tecnologia. Observação feita por Grácio (2020, p. 227) assinala que “os *software*, atualmente disponíveis para as análises bibliométricas, não possuem procedimentos automatizados [...] considera-se bastante difícil desenvolver estudos análogos para grandes universos de análise, dada a necessidade de tratamento manual [...].”

As técnicas da metodologia proposta nesta tese ainda podem ser aplicadas conforme a necessidade do pesquisador, pois é possível realizar o levantamento terminológico nas palavras-chave, títulos, resumos ou até no texto na íntegra. Permite ainda o levantamento dos autores em um rol de publicações, caso a entrada de dados seja essa. Assim, a metodologia não se atém a realização de estudos a somente em

determinada seção dos documentos e não se prende a um tipo de coleta de dados, atrelada a padronizações de bases bibliográficas.

Segundo Hjørland (2002), ao menos duas das abordagens devem ser utilizadas nos estudos analíticos de domínio. O levantamento da terminologia em um *corpus* pode ser combinado com estudos bibliométricos, que já estão consolidados na área da CI. Assim, além de possibilitar estudos simultâneos, a pesquisa também se justifica por apontar meios apropriados para percepção, interpretação e principalmente integralização de características relevantes dos discursos. Justifica-se ainda ao consentir que a terminologia é o vocabulário associado a uma disciplina, profissão ou atividade “e, portanto, faz parte da linguagem especial do domínio. Conhecer a terminologia é uma parte importante da capacidade de comunicar e compreender o conhecimento em um determinado domínio.” (HJØRLAND, 2022, n.p.).

Guimarães (2014, p. 19), por exemplo, sugere a utilização da análise de domínio em pesquisas em organização do conhecimento com a realização de paralelo entre a “terminologia da literatura científica de um domínio (p.ex., palavras-chave de artigos científicos) e as linguagens de indexação da área, a análise do universo e das relações entre referentes e correntes teóricas de um domínio ou, ainda, a análise das relações entre temas, referentes teóricos e correntes teóricas em um domínio.”

A pesquisa pode ser empregue em contextos em que a mineração de texto se fizer pertinente para levantamento de termos, classificação ou verificação de relações temáticas entre documentos de um *corpus* textual qualquer. Além disso, pretende-se, contribuir com a literatura científica sobre o tema e colaborar com a linha de pesquisa Produção e Organização da Informação, do Programa de Pós-graduação em Ciência da Informação (PPGCI), da Universidade Estadual Paulista “Júlio de Mesquita Filho” (UNESP), campus de Marília-SP.

## 1.3 Objetivos

Considerando os argumentos apresentados, tem-se como objetivo geral, propor uma metodologia para realização das técnicas de dedução de frequência de termos e categorização temática por meio de algoritmos da mineração de texto. Tal metodologia visa instrumentalizar e automatizar a abordagem terminológica, uma das onze abordagens que compõem a análise de domínio na área de CI. A aplicação da metodologia busca encontrar padrões de conectividade nas estruturas de

conhecimento, linguagem e comunicação nos discursos de um determinado *corpus* (resumo de comunicações científicas), entre um período pré-estabelecido.

Para apresentar e ilustrar a aplicação da metodologia, entendeu-se como pertinente a realização de demonstração prática. Os procedimentos e técnicas foram aplicados em um *corpus* textual, que são os resumos das comunicações científicas do Grupo de Trabalho 7 (GT7) – Produção e Comunicação da Informação em Ciência, Tecnologia & Inovação, apresentadas nos ENANCIBs[13] realizados entre 2012 e 2018. Para se alcançar o que foi proposto de forma geral, pretende-se especificamente:

✓ Em relação ao desenvolvimento dos textos por seus autores, com o método de dedução de frequência de termos, pretendeu-se apontar as especificidades das sentenças dos resumos, a fim de detectar padrões de escrita e tipificar os laços comuns no que se refere à quantidade de termos utilizados nas comunicações do grupo de autores, propiciando caracterizar um domínio pela junção dos elementos e não mais de forma individual;

✓ Ainda com o método de dedução de frequência de termos, procura-se fazer o levantamento da ocorrência dos termos simples (copalavras)[14] e os padrões de coocorrências dos termos compostos mais utilizados pelos autores na composição dos textos, com vistas a demonstrar as dinâmicas de uso e de formação dos n-gramas (*cluster* de palavras), possibilitando observar a estrutura comunicacional dos resumos;

✓ Por meio da técnica de análise por categorização temática, indicar os conjuntos de termos que representam os principais tópicos abordados no *corpus* textual, apresentando as possíveis articulações entre os resumos e os tópicos levantados, correlacionando as comunicações por categorias para fazer o agrupamento dos textos com temas semelhantes;

✓ Por fim, tendo em vista uma alternativa para complementar a instrumentalização da abordagem terminológica para análise de domínio, pretendeu-se realizar o processo conhecido como *stem* (redução ao radical), que permite agrupar os termos com significados parecidos, com vistas a contribuir na elaboração de sistemas de recuperação da informação, ao padronizar e diminuir as variações dos vocábulos empregados nos processos de indexação.

---

[13] Encontros Nacionais de Pesquisa e Pós-Graduação (ENANCIB) são realizados pela da Associação Nacional de Pesquisa e Pós-graduação em Ciência da Informação (ANCIB).

[14] Nesta pesquisa, ocorrência dos termos e coocorrências dos termos são sinônimos de copalavras.

# 2 ESTUDOS CORRELATOS

Nesta seção, demonstram-se algumas pesquisas relacionadas à análise de domínio, mineração de texto e se evidencia o grande rol de estudos sobre a literatura científica. Apontamentos que servem para demonstrar os encaminhamentos de investigações realizados sobre os temas abordados nesta tese e os níveis de importância no cenário nacional e internacional. Não se pretendeu esgotar o assunto, mas sim apresentar brevemente um cenário relacionado às temáticas manifestas nesta pesquisa. Ao mesmo tempo, entende-se que a exposição do panorama encontrado nesses estudos pode servir para ratificar as argumentações sobre as contribuições a que esta pesquisa se propõe.

Lee, Kim e Kim (2010) analisaram o descritor biblioteca digital no âmbito da Biblioteconomia e da Ciência da Informação, com o objetivo de explorar o desenvolvimento e as características dessas áreas de pesquisa. A metodologia analítica de domínio foi empregada com apoio da mineração de texto. Os autores buscaram, na base de dados bibliográfica LISA, documentos publicados entre 1994 e 2008 que contivessem os termos biblioteca digital ou bibliotecas digitais e aplicaram as seguintes técnicas: criação de perfil (algoritmo TF + IDF), agrupamento por proximidade (*clustering* – algoritmo PNNC) e agrupamento por conexões (*cluster by network* – método *CBNet-Ward*).

Santarem (2011) desenvolveu pesquisa para caracterizar a comunidade científica brasileira em Tratamento Temático da Informação (TTI), apoiada em indicadores bibliométricos, um dos enfoques de estudos analíticos de domínio. A investigação se deu nos currículos de pesquisadores cadastrados na Plataforma Lattes, em que foram encontrados a expressão TTI nas palavras-chave, analisando-se coautorias, cocitações e demais parcerias entre os autores e as instituições.

Guimarães, González e Alencar (2012) também abordaram o TTI e analisaram a presença e a articulação do tema análise documental nos estudos apresentados pelo GT-2[15] nos ENANCIBs de 1994 a 2010. A pesquisa buscou caracterizar um domínio de conhecimento com auxílio de duas abordagens analíticas de domínio, a bibliométrica e a epistemológica. O levantamento temático se deu pela busca de ocorrência no título, subtítulo, resumo, palavras-chave ou títulos de seções, dos

[15] Grupo de Trabalho 2 (GT-2): Organização e Representação do Conhecimento.

seguintes termos: Análise Documental/Análise Documentária, Representação Documental/Representação Documentária, Linguagem Documental/Linguagem Documentária, Condensação Documental/Condensação Documentária, Leitura Documental/Leitura Documentária. Os levantamentos bibliométricos se deram por meio das análises de citação e cocitação[16].

Ainda nesse ano Lu e Wolfram (2012) desenvolveram estudo bastante interessante e com significado profícuo para esta tese, justamente pelas características metodológicas utilizadas. Os autores propõem novas abordagens para estudos bibliométricos, balizados pela mineração de texto. As abordagens assemelham-se com a proposta deste estudo, pois são baseadas na coocorrência de palavras e na modelagem de tópicos, com uso do algoritmo de Alocação Latente de Dirichlet (*Latent Dirichlet Allocation - LDA*).

Os autores tinham o objetivo de medir a relação dos autores com base no conteúdo (termos e temática) de suas produções. Utilizaram para isso os títulos, palavras-chave e resumos de 5527 documentos baixados da WoS. Foram identificados 6282 autores nas bibliografias, separando-se para a pesquisa os 50 mais prolíficos. Entendendo que os dados representam um tipo de medida de similaridade, para construir os mapas de relações e apresentação dos resultados os autores se valeram do *software* SPSS PROXSCAL.

Gandra e Duarte (2013) propuseram articulação entre a Análise de Domínio e como ela pode contribuir para a abordagem social dos estudos de usuários da informação. Buscaram verificar, por meio de revisão da literatura e análise dos conteúdos, se tais pressupostos são efetivamente utilizados e como são aplicados nos estudos de usuários; e concluíram que há poucas pesquisas sobre usuários que se baseiam na análise de domínio para serem desenvolvidas.

Também em 2013, Calvo Fuente, Cantos Mateo e Zulueta García verificaram, por meio da análise de coocorrência de palavras, as temáticas mais relevantes no campo da Fisioterapia, na Espanha, com dados provenientes das palavras-chave de artigos da área, pesquisados na *WoS* e sem corte temporal. Fez-se uso do *software* VOSviewer[17] para apresentação gráfica dos resultados.

---

[16] Análises de citação e cocitação são indicadores bibliométricos que fornecem a visualização dos autores basilares em um campo de conhecimento (MANHIQUE; CASARIN, 2018).

[17] VOSviewer é uma ferramenta de *software* projetada para analisar automaticamente registros bibliográficos.

Smiraglia (2013) utilizou a análise de domínio para verificar se o *Functional Requirements for Bibliographic Records (FRBR*[18]*)* poderia ser caraterizado como um domínio. Reuniram-se as citações de 91 artigos de profissionais e acadêmicos presentes no volume *The FRBR Family of Conceptual Models*, totalizando 1.511 referências que foram organizadas em uma planilha Excel. As citações exigiram limpeza manual, pois não estavam no formato autor-data, nem nomes invertidos, e algumas ainda apareciam em notas de rodapé e não nas referências bibliográficas, tornando o trabalho oneroso e demorado. Após a limpeza, restaram 1.499 citações e verificou-se no estudo indicadores como: análise citação, afiliação de autores, cocitação e copalavras (aplicada no título dos artigos).

Pesquisa apoiada em indicadores bibliométricos, mas com teor mais profundo, encontra-se em Grácio e Oliveira (2014), entendendo que a análise de domínio, por meio da bibliometria, permite identificar como o conhecimento científico é construído e socializado. As autoras utilizaram indicadores absolutos e relativos de cocitação, notadamente Cosseno de Salton, para comparar a contribuição dos indicadores na caracterização de um domínio, inferindo que a associação dos dois parâmetros para visualização das estruturas de uma especialidade científica seria o mais adequado.

Guimarães *et al.* (2014) buscaram categorizar os temas e subtemas, bem como os temas mais comuns e regionalizados e ainda as interlocuções entre autores e instituições. O estudo se deu nas atas dos congressos dos capítulos brasileiro, espanhol e norte-americano da ISKO[19], realizados em 2011 e 2013. A pesquisa centrou-se na abordagem bibliométrica identificando: temas, subtemas, assuntos, autoria e a procedência geográfica. Revelaram-se os autores mais produtivos, sua procedência geográfica e as colaborações científicas. Nas temáticas, as análises ocorreram em dois níveis: (a) em âmbito macro, a partir dos temas oficiais e subtemas (nesse caso, verificando os títulos dos artigos) de cada evento e as comparações entre eles; e, (b) em âmbito micro, a partir das palavras-chave, verificando os assuntos de maior predominância.

---

[18] Requisitos Funcionais para Registros Bibliográficos (FRBR) representa um grupo de modelos conceituais promulgados pela Federação Internacional de Associações e Instituições Bibliotecárias (IFLA), em 1998, e serve de base para reengenharia dos serviços bibliográficos de bibliotecas.

[19] ISKO – International Society for Knowledge Organization, criada na Alemanha, em 1989, por Ingetraut Dahlberg e Dagobert Soergel.

Suenaga e Cervantes (2014) buscaram compreender, na literatura das áreas de Organização do Conhecimento, Análise de Domínio e Arquivística, a estrutura de conhecimento oriundos dos arquivos. Buscou-se nas referências à fundamentação dos preceitos dos estudos analíticos de domínio no contexto da Organização do Conhecimento a possibilidade de entender a estrutura de conhecimento no domínio da Arquivística. Evidenciou-se que a Análise de Domínio é pertinente para estudos de Organização do Conhecimento, pois as autoras compreendem que a análise de domínio se interessa pela construção coletiva do conhecimento e memória social, aparentes nos processos comunicativos e na linguagem.

Em estudo direcionado a literatura da Zootecnia, a teoria analítica de domínio se fez presente em Rosas e Grácio (2015), que verificaram como a colaboração científica (coautoria) estrangeira impacta na produção científica da área. O indicador bibliométrico de coautoria no âmbito internacional foi aplicado em artigos brasileiros do domínio, publicados em periódicos Qualis A1 e A2 (de 2007 a 2009), complementados por estudos históricos e epistemológicos, amparados na proposta metodológica de Hjørland (2002) e Tennis (2003). A base bibliográfica utilizada foi a *Scopus* e os *software* Excel e Pajek auxiliaram na tabulação e construção da rede de coautoria respectivamente.

Nesse mesmo ano, Dante *et al.* (2015) aplicaram a análise de domínio e de rede para verificarem a questão das competências profissionais e sua relação com tecnologia e engenharia e concluíram que a análise de domínio facilita a estruturação e organização da informação, identificando-a como um campo de estudo incipiente, mas com potencial para análises de competências em diferentes áreas. Para os autores, a análise de domínio centra-se na verificação dos registros bibliográficos e sustenta que os desenvolvimentos científicos podem ser rastreados ao estudar indicadores presentes nas publicações acadêmicas. Os indicadores bibliográficos foram verificados por meio das análises de citação, cocitação e copalavras com auxílio dos *software* VOSviewer e o CiteSpace.

Guimarães *et al.* (2015) amparam-se na análise de domínio e utilizam a análise de conteúdo (abordagem epistemológica) e citações (abordagem bibliométrica) para comparar a conceituação da organização do conhecimento e os diferentes referenciais teóricos presentes nas atas dos congressos nacionais do ISKO, realizados pelos capítulos brasileiro, espanhol, norte-americano e francês nos anos de 2011 e 2013. Procurou-se o termo *knowledge organization* (e equivalentes em

francês, espanhol e português) nos títulos, palavras-chave, resumos e títulos de seções das comunicações publicadas. Partiu-se de 305 trabalhos, fazendo, em seguida, a leitura dos textos que revelaram incidência do termo, separando os que apresentavam conceitos, definições ou considerações teóricas da organização do conhecimento (refinamento), restando 48 artigos, para os quais construiu-se as redes de citação (bibliometria) com o *software* Pajek, com vistas a identificar os principais referentes teóricos e as interlocuções entre os capítulos.

Sérgio, Silva e Gonçalves (2016) ponderam sobre o grande volume de informações não estruturadas em formato textual na internet e propõem um modelo de descoberta de conhecimento baseado nas técnicas de correlação e associação temporal entre termos de um domínio em grandes coleções de documentos. Pesquisa descritiva e exploratória que utilizou a base bibliográfica *Science Direct* para coleta do material a ser trabalhado. Segundo eles, por esse tipo de informação, é possível extrair regras, padrões, tendências e redes, que auxiliam as tomadas de decisões nas instituições.

Amorim e Café (2016) versam sobre a definição dos conceitos de análise de domínio a partir dos artigos de Hjørland encontrados na *LISA, WoS, Scopus e AKO*[20], os quais foram estabelecidos em: comunidade discursiva, domínio e linguagem. O estudo apontou que esses termos fundamentam a análise de domínio, pois compreendem os aspectos sociais que implicam o uso da informação. Amorim e Bräscher (2016), embora enxerguem a metodologia analítica de domínio como promissora, destacam a falta de consenso dos conceitos e suas aplicações. Investigaram-se métodos da análise de domínio sob a perspectiva da cartografia deleuze-guattariana[21], com o intuito de propor uma metodologia cartográfica da análise de domínio e levantar debates sobre as metodologias. Esses dois últimos estudos apresentados possuem conteúdos bastante semelhantes ao de Amorim e Café (2014), inclusive em suas metodologias e bases de dados pesquisadas, porém, o estudo de 2014 focava apenas no pensamento do filósofo francês Gilles Deleuze.

Outra pesquisa desenvolvida no âmbito da Análise de Domínio (AD) está expressa em Guimarães *et al.* (2017), que analisou a presença do próprio tema análise

---

[20] LISA – Library and Information Science Abstracts, WoS – Web of Science e Scopus são bases de dados bibliográficas e AKO – Advances in Knowledge Organization – é um periódico.

[21] Deleuze e Guattari (2010) pretendiam entender o pensamento identificando seu funcionamento e não o definindo. A cartografia emerge como um novo modo de produzir conhecimento.

de domínio na literatura estrangeira nas bases da *Scopus, LISA e Web of Science*, entre os anos de 1995 e 2016, fazendo uso da abordagem bibliométrica. Utilizou-se o termo *domain analysis* nas bases para recuperar os documentos para as análises de citação, locais de publicação e períodos mais produtivos, a fim de apontar as comunidades epistêmicas do domínio. Destaca-se uma elite de 64 autores, com os temas predominantes em organização do conhecimento e estudos métricos em informação, tendo Hjørland como autor mais citado.

Manhique e Casarin (2018) definiram seu estudo como uma perspectiva analítica de domínio, por focarem em profissionais agrupados como comunidades discursivas, alinhados em sua linguagem e pensamento na área. Coletaram documentos na Scopus que continham a expressão *("information literacy") AND ("phenomenographic" OR "relational*")* com o objetivo de caracterizar a estrutura intelectual de competência informacional por meio da abordagem fenomenográfica ou relacional. Apontaram-se os autores mais produtivos, os mais citados, filiação institucional e países que mais produzem sobre a temática. O *software* Ucinet serviu para gerar e apresentar a rede de cocitação dos autores, provenientes dos dados trabalhados na planilha eletrônica Excel.

O artigo de Almeida e Dias (2018) procurou apresentar o estado da arte sobre análise de domínio no campo da CI no Brasil, ao analisarem 45 documentos entre artigos, teses, dissertações, comunicações orais do ENANCIB e outras publicações entre os anos de 2013 e 2018. Buscaram-se publicações contendo Análise de Domínio e Ciência da Informação em português, bem como seus correspondentes na língua inglesa (*Domain Analysis e Information Science*), nos resumos, títulos ou palavras-chave no Google Acadêmico, *LISA, LISTA*, BRAPCI, BDTD e BENANCIB[22].

Os autores explanam que fizeram uso das onze abordagens propostas por Hjørland (2002) para referida análise e entendimento dos fundamentos teórico-metodológicos da análise de domínio. O trabalho destacou como tendências de pesquisas os estudos bibliométricos (indicadores de análise de citação, cocitação, coocorrência de palavras e análise de redes) e técnicas de organização e representação do conhecimento.

---

[22] LISA – Library and Information Science Abstracts; LISTA – Library, Information Science & Tecnhnology Abstracts, BRAPCI – Base de Dados Referencial de Artigos de Periódicos em Ciência da Informação, BDTD – Biblioteca Digital Brasileira de Teses e Dissertações e o repositório BENANCIB.

Pesquisa que objetivou analisar as inter-relações; interdisciplinaridade e objetos comuns nos estudos do Design da Informação (DI), Ciência da Informação (CI) e Organização do Conhecimento (OC) foi encontrada em Nakano *et al.*, 2018, para parametrizar o lugar do DI no Brasil e sua relação com estudos sociais, além de verificar os programas de CI que contemplam o curso de design. Os autores destacaram o uso da análise de domínio como aporte metodológico, utilizando-se da abordagem histórica e o princípio da divisão social do trabalho sob a ótica da teoria da complexidade para caracterizar o domínio de DI.

Em pesquisa que combinou abordagens epistemológica, histórica e bibliométrica, Tognoli, Silva e Silva (2019) empregaram a análise de domínio na produção científica sobre Arquivologia na Organização do Conhecimento (OC). Faz-se a análise da temática arquivística recorrente em OC e análise de citação (indicador bibliométrico) nos periódicos *Knowledge Organization* (KO) e *Scire: organización y representacíon del conocimiento*. Analisaram-se 47 trabalhos publicados entre 1995 e 2019 nos dois periódicos, que foram recuperados com buscas pelos termos em português e espanhol (*archivo/archivistica*), na *Scire* (28 artigos), e em inglês (*archive/archival science*), na KO (19 artigos), presentes nos títulos, resumos e palavras-chave. O estudo apresenta a temática encontrada nos títulos, bem como o percentual de cada tema encontrado; o país de origem dos autores e ainda a análise bibliométrica que foram feitas e demonstradas com o *software* VOSviewer.

Damus e Acuña (2019) fazem uma abordagem teórica do paradigma social, além de verificar o alcance semântico da expressão, as possibilidades teóricas para pesquisas em biblioteconomia e as implicações práticas das abordagens sugeridas por Hjørland, apontado pelas autoras como criador do paradigma. O texto aponta o surgimento do tema e sua inserção na CI, utilizando principalmente o artigo seminal (HJØRLAND; ALBRECHTSEN, 1995) e outros de seu criador. Constatou-se que as abordagens contribuem para analisar os efeitos que o uso da terminologia tem nas comunidades discursivas e auxiliam no entendimento de: como diferentes conceitos nos discursos se relacionam; conceitos relevantes de um domínio e relações hierárquicas; ligações semânticas; variações de termos e conceitos ao longo do tempo e as influências socioculturais que os campos do conhecimento podem sofrer.

No mesmo ano, Hsu e Li (2019) objetivaram apontar tendências de pesquisa, identificar *clusters* das temáticas e como eles se relacionam, além de verificar as teorias mais usadas para processamento e interpretação dos dados e identificar

lacunas de pesquisa de *big data* médico. Utilizaram-se os indicadores de copalavras e coocorrência de termos nas palavras-chave de trabalhos na língua inglesa, coletados na base *Scopus*, entre 2000 e 2016, utilizando as seguintes expressões: *('big data' OR 'bigdata' OR 'mega data' OR 'megadata') E (‘*medic*’ OR ‘*health*’).*

Após excluírem-se trabalhos considerados inadequados (artigos de revisão, cartas ao editor, nota, editorial, capítulo de livro), restaram 1791 documentos, que manualmente foram manipulados executando-se: eliminação de trabalhos sem palavras-chave; agrupamento de termos idênticos; eliminação de palavras comuns (*stop words*) e remoção de diacrônicos (hifens, apóstrofos, asteriscos, barra, acentos), para geração dos indicadores de análise de redes e apresentá-los via *software* Gephi.

Pesquisadores das escolas de Biblioteconomia e Ciência da Informação da Universidade de Boston e do Kentucky/EUA (JOO; OH, 2019) apresentaram resultados preliminares de estudo no domínio de Organização do Conhecimento, ao compararem os títulos, palavras-chave e resumos de 914 artigos escritos por pesquisadores e bibliotecários entre os anos de 2008 e 2017. As bases para coleta do material foram: *Knowledge Organization; Cataloging & Classification Quarterly*; e *Library Resources & Technical Services*. Os autores identificaram de forma manual as afiliações e o cargo do autor principal para definirem se a autoria era de um pesquisador (449 documentos) ou de um bibliotecário (465 documentos). Utilizaram técnicas de mineração de texto para a comparação dos dados: análise de frequência de termos; modelagem de tópicos com o algoritmo de Alocação Latente de Dirichlet (LDA); e o processo de *stem* (derivação ou radicalização).

O estudo realizado por Joo (2020) utilizou a mineração de texto para analisar o domínio de usuários da informação, em 132 artigos coletados na Wikipedia. Vários métodos de análise textual foram aplicados para explorar conceitos e tópicos predominantes em relação ao tema. Os métodos aplicados incluem: análise de frequência de termos, modelos de tópicos de Alocação Latente de Dirichlet (LDA), agrupamento hierárquico, redução dimensional e análise de coocorrência de termos. A pesquisa demonstrou quatro conceitos fundamentais na temática – usuários, comportamentos, sistemas/ tecnologia e objetos de informação – e outras áreas no domínio – comportamento de busca, recuperação da informação, interação humano-máquina, experiência do usuário, fatores humanos e outros.

Nesse mesmo ano, Trabadela-Robles *et al.* (2020) intitularam seu artigo como Análise dos domínios científicos nacionais em Comunicação (Scopus, 2003-2018) e

explanaram sobre o crescimento de pesquisas no campo comunicacional, utilizando como método na pesquisa os indicadores bibliométricos como: número de documentos publicados em revistas científicas (Ndoc); porcentagem que os documentos representam em relação à produção global (%Ndoc); número de citações por documento; porcentagem de documentos citados; porcentagem das colaborações internacionais; liderança e porcentagem de liderança (autor do país atuou como líder – autor correspondente); impacto normalizado (NI – média de citação normalizada) e outros. Novamente a análise de domínio se deu por índices provenientes da bibliometria aplicados em artigos dos 27 países mais produtivos dentro da categoria Comunicação na base Scopus.

Mokhtarpour e Khasseh (2021) tinham como finalidade mapear e analisar a estrutura conceitual e temática de publicações indexadas na *WoS* entre 1990 e 2016, nos campos de Biblioteconomia e Ciência da Informação. Os métodos utilizados foram a análise de copalavras e o modelo de Kleiberg para encontrar pontos quentes[23] na rede de palavras. O *software* CiteSpace serviu para gerar as redes e as visualizações gráficas do estudo.

Rego-Piva, Casarin e Guimarães (2021) utilizaram a metodologia de análise de domínio como forma de atestar sua confiabilidade em processos de avaliação de revistas científicas, por meio de dados quantitativos e contextuais. O estudo foi aplicado no *Brazilian Journal of Information Science: research trends (*BRAJIS), do Programa de Pós-Graduação em Ciência da Informação (PPGCI) da UNESP, que tem formato eletrônico e livre acesso. Utilizou-se a abordagem bibliométrica no estudo diacrônico para analisar aspectos como: autores, vinculações institucionais, países, idiomas dos artigos, palavras-chave (copalavras/temas) e referências.

Outro estudo que se utiliza da análise de domínio (BARROS; LAIPELT, 2021), teve como objetivo mapear e fazer apontamentos temáticos, semânticos e discursivos em artigos da área de Organização do Conhecimento, em artigos publicados na Revista Em Questão, desde 2003, quando a revista passou a utilizar o sistema digital

[23] É um algoritmo de ranqueamento de ligações, baseadas em métricas de centralidade para verificar conexões em uma rede. A Ciência da Informação e a Ciência da Comunicação podem ajudar a entender e a extrair informações a partir da estrutura de coleções de páginas da web ou documentos.

SEER. Utilizando o *software* Sketch Engine[24], um *corpus* de 30 textos forneceu informações para organização do domínio a partir das abordagens temática, epistemológica, terminológica e discursiva. Para recuperação dos textos, utilizaram-se os termos: Indexação, Organização do Conhecimento, Organização da Informação, Sistemas de Organização do Conhecimento, Representação Documental, Representação do Conhecimento, Representação da Informação, Representação e Organização do Conhecimento, Taxonomia, Tesauro e Ontologia, nos títulos, resumos e palavras-chave dos artigos.

Marques, Marques e Maculan (2021) relatam a necessidade de transformar grandes volumes de dados em informação processável para uso humano, apontando que a mineração de texto pode definir relações entre elementos textuais internos ou externos a eles. Com técnicas da mineração aplicadas a dados abertos da CAPES[25], o artigo busca descobrir padrões de coocorrência de palavras-chave em teses e dissertações na área da CI. Os autores concluíram que a pesquisa se enquadra em estudos métricos, por ter-se gerado uma matriz de coocorrência de termos e a rede de conexões entre os trabalhos.

Ainda no mesmo ano, Ferreira e Correa (2021) realizaram estudo em artigos de periódicos brasileiros indexados na Brapci ou Scielo. Os termos utilizados para recuperação nas bases em setembro/2019 foram mineração de texto ou descoberta de conhecimento, buscados nos metadados dos trabalhos. Os autores empregaram procedimentos de estudos métricos da informação e análise de conteúdo. Após a exclusão das duplicidades, fez-se a leitura dos resumos e concluiu-se que, em 18 anos (primeiro artigo entrado em 2001 e último em 2018), somente 28 publicações tinham relação com a temática extração de conhecimento por meio da mineração de textos. Dentre estes, 13 tratam da mineração de textos científicos, dos quais 9 tem natureza aplicada.

Esse apanhado de quase 30 artigos, desde 2010 até 2021, que trataram ou fizeram uso de temáticas próximas as discutidas nesta pesquisa servem para apontar, mesmo que de maneira sutil, como se desenvolvem tais estudos. Alguns pontos são passíveis de observações e podem corroborar o julgamento da proposta da presente

---

[24] Sketch Engine: *software* de mineração de texto que pode identificar termos típicos na linguagem ou o que é raro, incomum ou emergente. Tem uso gratuito por 30 dias e está disponível em: https://www.sketchengine.eu/.

[25] Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES).

tese, pois parte considerável, 18 artigos, argumentam sobre o uso ou a definição da Análise de Domínio (AD), sendo que 12 deles apontam amparar-se nos estudos bibliométricos para obtenção dos resultados, demonstrando que estudos da produção científica são bastante fortes no campo da CI.

A teoria analítica de domínio propõe análises ante perspectivas mais abrangentes, enxergando o conhecimento como construção coletiva, tarefa bastante evidenciada pelos estudos bibliométricos e de coocorrência de termos (copalavras). Contudo, muitos desses estudos foram elaborados com os dados sendo manipulados de maneira manual, semiautomática, com uso de alguns *software* de forma concomitante, ou ainda foram efetuados cortes nas amostras para que ficassem com um tamanho possível de se trabalhar, retratando certas dificuldades.

Alguns dos *software* mencionados são Excel, Pajek, Gephi, CiteSpace, Ucinet e VOSviewer, utilizados como auxiliares para determinar relações e visualizações dos estudos. As bases bibliográficas LISA, Scopus e WoS foram as que mais apareceram como fonte de dados para as pesquisas e vários trabalhos se utilizaram de técnicas de mineração de textos, principalmente verificando frequência e coocorrência de palavras em diversas partes dos textos, como palavras-chave, títulos e resumos.

# 3 REFERENCIAL TEÓRICO

A ciência tem um papel importante dentro da sociedade por não se ater ao que se conhece como senso comum ou por achismos quanto a determinados assuntos. Não que o conhecimento empírico não deva ser levado em consideração, mas há que se terem meios para que as suposições ou as vivências sejam referendadas, aceitas e confirmadas. Aliás, à ciência cabe justamente o papel de confirmar ou refutar crenças, muitas vezes tomadas por ímpeto, desconhecimento ou apenas ideologias.

Para assimilar a Análise de Domínio (AD), entendeu-se como necessário retroceder, a fim de historicamente compreender a ligação da ciência com o que foi intitulado como domínio. Conforme Lloyd (1995), a racionalidade, a lógica, a objetividade, a coerência, a validade e peculiaridade é que dão às ciências o seu caráter de elucidação dos fenômenos observáveis, obedecendo a regras e métodos validados. Para o autor, um dos maiores avanços da ciência (compreensão da natureza da explicação) foi a "percepção geral de que metodologias, teorias e explicações se relacionam mutuamente através de constelações extralógicas e historicamente variáveis, e descritas como 'conhecimento', 'tradições', 'paradigmas', 'programas de pesquisa', 'campos' ou 'domínios'." (LLOYD, 1995, p. 48-49).

Os campos de conhecimento ou domínios científicos distinguem-se das não ciências por seus paradigmas[26] consensuais. A ciência tem sentido quando relacionada ao seu paradigma e por isso ela é historicamente específica e não universal. As investigações e explicações são constituídas de questões empíricas, o que as fazem também se diferenciar umas das outras. (LLOYD, 1995). O autor prossegue e explana que o delineamento claro e uma coleção de enunciados importantes de um objeto específico, podem ser chamados de domínio ou informações de fundo. Esse é o conjunto que desenvolve a racionalidade da ciência, que é o que torna a ciência algo singular, notório, baseado na razão.

> [...] a ciência visa tornar-se, tanto quanto possível, autônoma e autossuficiente em sua organização, em sua descrição e no tratamento de seu objeto – tornar-se capaz de delinear seus domínios de investigação e a informação de fundo relevante para isso, formular seus problemas, planejar os métodos de abordar esses problemas, determinar um leque de soluções possíveis e estabelecer critérios de reconhecimento das soluções aceitáveis, tudo exclusivamente em termos de domínio enfocado e de outras crenças

[26] Para Kuhn (2003) um paradigma é quando um conjunto de práticas científicas é reconhecida por determinada comunidade acadêmica durante algum tempo, com subsídios a problemas e modelos de soluções que podem ser verificados na sua área de atuação.

> aprovadas e fora de dúvida que tenham sido considerados relevantes para esse domínio, ou seja, tornar seu raciocínio inteiramente autossuficiente em todos os aspectos. (LLOYD, 1995, p. 52).

A CI se desenvolveu nesse contexto de diferenciação da ciência e não ciência, das especificidades de cada campo, clareza dos enunciados e delimitação do objeto de estudo, os quais direcionam para compreensão racional de fenômenos. Segundo Araújo (2014), nas décadas de 1920 a 1940, alguns cientistas começaram a trabalhar na elaboração de índices, resumos, canais de disseminação, provendo informações para facilitar o acesso e dar agilidade ao trabalho dos seus pares em suas áreas de atuação como química, física, engenharia, entre outras. Com o passar dos anos, esses profissionais se intitulariam cientistas da informação.

A institucionalização da Ciência da Informação começou na Inglaterra, em 1948, onde foi realizada a *Royal Society Scientific Information Conference*, e onde, dez anos mais tarde, seria criado o *Institute of Information Scientist*. Nessa mesma época, vinculado à Academia de Ciências o Viniti, *Vserossiisky Institut Nauchnoi i Tekhnicheskoi Informatsii* também teve origem na extinta União Soviética e, ainda em 1958, viria a acontecer nos Estados Unidos a *International Conference on Scientific Information*. A CI foi obtendo aceitação como uma ciência voltada à informação em ciência e tecnologia, que se destacaria principalmente pela preocupação da circulação da informação, seu fluxo e acesso e não mais pela posse dos documentos.

Nesse mesmo período, há também a preocupação de transferir o conhecimento acadêmico para a sociedade, e vários estudiosos passam a verificar o processo de comunicação da informação científica, analisando os registros das pesquisas como: relatórios, seminários, artigos, livros, características dos trabalhos, além de como se obtinha os dados para sos estudos. Assim, a CI consolidou-se como campo que estuda a origem, armazenamento, materialização e fluxos da informação nos diversos suportes, proposta que seria então disseminada pela Organização das Nações Unidas para a Educação, a Ciência e a Cultura (UNESCO), nos anos 1970, e adotada pelo IBICT no Brasil. (BORKO, 1968; ARAÚJO, 2014).

## 3.1 Paradigmas epistemológicos em Ciência da Informação

A literatura na Ciência da Informação (CI) demonstra que três paradigmas epistemológicos (físico, cognitivo e social) fizeram ou fazem parte dos estudos da área, e embora caracterizados de forma distinta, são complementares e relacionais.

O paradigma físico situa-se ante uma epistemologia material, baseado nos processos computacionais, inferindo à informação um caráter puramente técnico, algo mensurável, quantificável, processável e transmissível por sistemas informatizados, sem preocupar-se com significados semânticos. Tal entendimento estende-se para a recuperação da informação em sistemas computadorizados, os quais então mediriam a precisão dos resultados obtidos nos processos de busca informacional. (SARACEVIC, 1995; CAPURRO, 2003).

O segundo paradigma, o cognitivo, surgiu por volta dos anos 1970 e direciona-se para os conhecimentos e comportamentos individuais do ser humano. Conforme Capurro (2003), algumas teorias fizeram parte desse paradigma, as quais defendiam uma visão cognitiva centrada no usuário, alegando que a busca por informação se origina de uma necessidade ou situação problemática, além da teoria popperiana de mundo físico, mundo da mente/consciência e o mundo das ideias/registros intelectuais; e ainda do modelo tradicional de recuperação da informação, mas com base no estado cognitivo do indivíduo. Se preocupava com o comportamento do usuário perante a informação, mas não ainda com os contextos coletivos e sociais.

Diante desse cenário, perceberam-se a necessidade de pensar em sistemas de recuperação da informação que levassem em conta elementos subjetivos do ser humano, bem como suas perspectivas, interesses e relações coletivas. Dessa forma, o paradigma social começa a ganhar força, por não se ater somente à estudos que buscam um modelo de linguagem para representar o conhecimento (paradigma cognitivo) e tampouco em algoritmos de recuperação da informação (paradigma físico) considerados ideais. (CAPURRO, 2003).

O principal direcionamento do paradigma social é olhar para os cenários informacionais de forma mais abrangente, e embora se entenda que os paradigmas físico e cognitivo amparam os estudos, as análises no paradigma social visam fugir do trivial, para ir além dos cálculos, pois, espera-se que os resultados não sejam simplesmente números de representatividades individuais, mas que retratem individualidades coadunadas, perspectivas e cenários informacionais coletivos.

## 3.2 O paradigma social na Ciência da Informação

A Análise de Domínio tem sua origem na Ciência da Computação, especificamente na Engenharia de *Software*, tendo como pioneiro Neighbors, o qual

utilizava a análise de domínio e reutilização de componentes em níveis mais altos de abstração[27] do que propriamente na codificação[28] do *software.* (PRIETO-DÍAZ, 1990; TAYLOR, 2018). Destaque-se que, mesmo surgindo na área de computação, a análise de domínio valeu-se da Ciência da Informação (CI) para se aprimorar.

Prieto-Díaz (1990, p. 48), atuante no campo da computação, propôs um modelo mais compreensível que o de Neighbors, baseado em uma estrutura que deriva esquemas de categorização especializados advindos da biblioteconomia. “Ao derivar um esquema de classificação facetada, o objetivo é criar e estruturar um vocabulário controlado que seja padrão não apenas para classificar, mas também para descrever títulos em uma coleção específica de domínio.” O autor pondera que existem várias pesquisas em muitas outras disciplinas como engenharia de conhecimento e modelagem conceitual, que lidam com problemas e resultados semelhantes, também conhecidos como análise de domínio.

> Outra aplicação da análise de domínio ocorre em Biblioteconomia ao derivar esquemas de classificação especializados. Esquemas de classificação de facetas especializados são derivados por meio de um processo manual que consiste em agrupar termos relacionados de uma amostra de títulos selecionados, definir nomes de facetas desses grupos, ordenar os termos em cada faceta e especificar regras para sintetizar classes compostas. O esquema de classificação resultante torna-se um modelo conceitual para o domínio da coleção. Agrupamento de termos, de títulos é equivalente a localizar objetos e operações em um domínio de aplicativo. A nomeação de facetas e a definição de regras de classificação são equivalentes a derivar um modelo de domínio ou criar uma linguagem de domínio. (PRIETO-DÍAZ, 1990, p. 50).

Embora existam apontamentos de ligação da análise de domínios com a CI por autores de outras áreas, ela só se torna conhecida nesse campo de estudo em 1995, quando Hjørland e Albrechtsen publicaram um artigo seminal sobre o tema, e a destacam como uma nova abordagem, fornecendo uma visão introdutória do assunto. Conforme os autores, para a CI, seria mais coerente analisar os domínios de conhecimento como “comunidades de pensamento ou comunidades de discurso que constituem a divisão social do trabalho.” (HJØRLAND; ALBRECHTSEN, 1995, p. 401).

---

[27] Engloba selecionar os dados e sua classificação, omitindo elementos não necessários para concentrar-se nos que realmente importam. Usada para criar uma representatividade do que se busca solucionar. A essência é selecionar a informação a ser ignorada para compreender o problema sem perder informações importantes. Capacidade de filtrar informações essenciais e descartar informações desnecessárias em um determinado contexto.

[28] Para que uma máquina execute uma tarefa, é preciso criar um algoritmo (sequência) de passos por meio do qual a tarefa será executada e esse algoritmo deve ser traduzido para uma linguagem de programação, que é compreendida pela máquina.

Entretanto, como apontado por Guimarães (2014, p. 17), precursor do tema no Brasil, “apenas sete anos mais tarde que a dimensão conceitual da análise de domínio tornou-se mais nítida – e operacional – para a ciência da informação.” Tal entendimento advém das 11 abordagens propostas por Hjørland (2002), as quais definiriam de forma mais profícua o tema.

Hjørland e Albrechtsen (1995, p. 400), descreveram a análise de domínio como “paradigma social, postulando a CI como uma das ciências sociais, destacando uma psicologia social, uma sociolinguística, uma ciência do conhecimento e uma perspectiva sociológica da ciência na CI.” A análise de domínio foi ainda caracterizada pelos autores como uma abordagem filosófico-realista, que busca encontrar a base para CI nas atividades sociais, em percepções subjetivas dos usuários, contrapondo os paradigmas físico e cognitivo. Assim, pelo prisma do paradigma social, um sistema de recuperação consideraria elementos inerentes ao usuário, seu contexto social, sua visão de mundo e da própria estrutura para recuperação da informação.

Embora Hjørland e Albrechtsen (1995) reconhecessem desconfiança ante seus apontamentos sobre análise de domínio na CI ser realmente algo novo, e mesmo observando que abordagens anteriores e contemporâneas continham princípios básicos da análise de domínio, eles entendiam que ainda não havia nenhuma definição clara do assunto. Não obstante, o principal objetivo dos autores não era apresentar uma nova teoria em si, mas oferecer uma base conceitual, uma visão mais esclarecedora dos princípios teóricos, estimulando pesquisas para continuidade do desenvolvimento da área da ciência da informação.

Até aquele momento, poucos autores discutiam sobre a unidade de estudo de CI não ser o indivíduo e sim os domínios (especialidades, disciplinas, ambientes). Hjørland e Albrechtsen (1995) apontam alguns pesquisadores como Patrick Wilson, Robert S. Taylor e Rosenbaum como sendo contribuintes latentes do tema que igualmente compartilhavam de suas perspectivas sobre AD. Para Wilson (1993), a análise de domínio procura entender como uma especialidade é afetada, e não como o indivíduo é afetado. O grupo deveria entender que a informação tem um estado lógico ou evidente: espera-se que a situação cognitiva individual sobre a informação seja adequada à especialidade, importante para a situação coletiva e não apenas para o indivíduo.

Taylor (1991) apresentou abordagem que se preocupava com a visão coletiva, baseada no conceito de ambientes de uso da informação, um tipo de pesquisa

também orientada para o domínio. Essa abordagem inspirou Rosenbaum (1993) a fazer uma integração teórica com a teoria sociológica moderna. Visões que representam uma linha de pesquisa contemporânea, consoante ao entendimento sobre análise de domínio postulado no artigo seminal de Hjørland e Albrechtsen (1995). Embora esses três pesquisadores apresentassem contribuições para a área de CI, não havia descrições detalhadas da abordagem teórica social confrontando com outras teorias. (HJØRLAND; ALBRECHTSEN, 1995).

## 3.3 O enfoque da análise de domínio

A Ciência da Informação recebe influências de várias outras áreas como: Pesquisa Educacional, Ciência Cognitiva (Linguística – Inteligência Artificial – Psicologia – Filosofia) e Sociologia. Os estudos sobre o conhecimento em diversos campos e as tendências transdisciplinares sobre o tema contribuíram nos apontamentos sobre a análise de domínio, pois o ponto de vista de Hjørland e Albrechtsen (1995, p. 405) alinhava-se com novas e importantes perspectivas em disciplinas adjacentes à CI:

> É claro que é perigoso generalizar sobre os desenvolvimentos em campos com muitas teorias diferentes, mas encontramos evidências de um desenvolvimento em que a psicologia moderna, linguística e o novo campo de estudo composicional estão todos olhando para a linguagem e outros processos cognitivos no contexto de um desenvolvimento sociocultural – em vez de uma estrutura intrapsicológica, em que a linguística é vista como parte da psicologia cognitiva.

O enfoque das ações cognitivas em um domínio representa uma tendência muito forte, em contraponto ao longo tempo que a mente foi entendida apenas como uma calculadora universal[29]. Teorias da área de psicologia sobre domínios específicos em processos cognitivos os percebem como um mecanismo ecológico[30] adaptativo da mente, ajustado a tarefas determinadas pelo conteúdo. Visão muito próxima do funcionalismo e do pragmatismo filosófico, que enxergam o conhecimento como fenômeno adaptável. (HJØRLAND; ALBRECHTSEN, 1995).

Os autores observam que, na área da Linguística, também aconteceu o deslocamento da visão estruturalista para uma mais funcionalista e sociolinguística (GAUDIN, 1993, 2003, 2014; PÊCHEUX, 1997; FAULSTICH, 1999, 2006; CABRÉ,

[29] Percepção reducionista do raciocínio humano a cálculos.
[30] Aqui entendida como as relações que os seres estabelecem entre si e com o meio em que vivem.

2005), com ênfase no uso da linguagem em comunidades discursivas. Mesmo com o estruturalismo influente na área, o surgimento de abordagens de domínios específicos no campo instigou novas pesquisas. Estudos que se baseavam em textos como isolados de estruturas sociais, culturais e históricas, passaram a destacar o conhecimento com relação dialética entre uma comunidade científica e seus integrantes, mediada pela linguagem e influências históricas do domínio.

Sobre a importância do ambiente na análise de domínio, Hjørland e Albrechtsen (1995) complementam que os grupos são compostos por indivíduos experientes e outros iniciantes com concepções diferentes sobre uma mesma obra. Um texto considerado claro não é o que explana tudo, e sim o que expressa o necessário. “O que é necessário e relevante para ser dito depende mais do que o propósito do escritor; também depende do que os leitores sabem e não sabem, e daí o que o escritor pode validamente assumir ou não.” (HJØRLAND; ALBRECHTSEN, 1995, p. 408). Dessa forma, a questão da clareza depende da correlação entre leitor e escritor, mediada pelo conteúdo expresso, e não da intenção do autor ali representada. Essa teoria aponta que o assunto de um documento deveria ser definido pelo potencial epistemológico da obra.

As observações dos autores sobre a influência do convívio social ante os entendimentos semânticos expõem que o padrão estruturalista/cognitivista (funcionamento do cérebro) não é capaz de desvendar as essências do conhecimento individual. As concepções de mundo e o modo de expressão de cada sujeito, devem ser explicadas pela divisão social do trabalho, antes de tudo, o que igualmente se aplica à comunicação científica. Para eles, ter a linguagem como meros rótulos reduziria o conhecimento ao entendimento individual, e por isso deveriam preocupar-se com visões mais holísticas, percebendo a linguagem como expressão da realidade, com impressões históricas, culturais e sociais.

Hjørland e Albrechtsen (1995) apresentam a análise de domínio como uma alternativa ao individualismo metodológico, o qual se fazia bastante presente nas ciências comportamentais, cognitivas e sociais, inclusive na Ciência da Informação (CI). No individualismo metodológico, os estudos sobre conhecimento são encarados como um processo cognitivo mental individual, isolados do contexto social e histórico de onde o conhecimento é produzido/alcançado. Na análise de domínio os discursos englobam características individuais (conhecimentos, preconceitos, estilos etc.), mas detêm nuances das relações sociais.

Desse modo, o entendimento de Hjørland e Albrechtsen (1995) é que, dentro da CI, o ponto principal são os domínios de conhecimento, as disciplinas, e não mais indivíduos biológicos, fisiológicos ou psicológicos. Cada sujeito integra grupos de trabalho, comunidades epistêmicas ou comunidades discursivas, um ser e uma ciência social em vez de intrapsíquica, do sujeito isolado, abstrato, menos mecanicista e mais contextual, sociocultural, própria de domínios.

Conforme exposto, a análise de domínio surge da preocupação de Hjørland e Albrechtsen (1995) em dar continuidade ao desenvolvimento da CI. Por exemplo, em relação ao paradigma físico, que se atém a parte técnica e não em quão bem informar o usuário, os autores se preocupam com a questão de o paradigma desprezar os profissionais da área, sob a ótica de que esses apenas abastecerão a tecnologia com informações e o que realmente interessaria seriam os aparatos tecnológicos da computação, tornando a CI mera coadjuvante de outro campo.

Outra preocupação dos autores refere-se às disciplinas de conteúdo ou assunto, pois diversos centros de informação e bibliotecas especializadas recrutam pessoas formadas em áreas específicas, como Direito, Química ou Medicina, para trabalhar com questões informacionais (relevância; indexação; recuperação da informação etc.), o que levaria a crer que tudo que se precisa para atuar como profissional da informação é ter conhecimento da área de formação. Para eles, abordagens com inclinações individualistas, que se abstêm dos aspectos coletivos ou domínios do conhecimento não contribuem com a CI.

Hjørland e Albrechtsen (1995) complementam que, no campo da ciência da informação, as abordagens inclinadas para estudos da mente (comportamento; pensamentos) do usuário podem revelar rotinas e padrões capazes de auxiliar no desenvolvimento de sistemas de informação. Porém, a CI deveria construir esses princípios como ferramentas de aperfeiçoamento de práticas sociais informacionais, e não apenas estudá-las, pois os princípios para construção de um sistema de informação devem vir dos cientistas e não dos usuários. Segundo os autores, os usuários pouco conhecem, por exemplo, sobre fontes de informação ou estratégias de busca, o que acaba por comprometer os estudos individualistas.

Para Hjørland e Albrechtsen (1995, p. 411), o cognitivismo implica que os sistemas informacionais devem “refletir a percepção subjetiva dos usuários sobre conhecimento e informação, não uma realidade objetiva (o que poderia contribuir para o desenvolvimento do conhecimento).” Para os autores, teorias sobre conhecimento

são mais importantes do que sobre usuários e sistemas de informação, pois, os usuários não conseguem expor necessidades de assuntos que não conhecem.

Teorias sobre o conhecimento são complexas e, ao isolar o organismo e a mente dos ambientes, dificulta-se mais o entendimento da realidade, tornando o estudo mais subjetivo ainda. A exclusão das percepções individuais exclui também das metodologias científicas as percepções de valores, portanto, para Hjørland e Albrechtsen (1995) a CI deveria aprofundar-se em teorias mais amplas e considerar a cultura na qual os sistemas informacionais estão inseridos, o que significaria abrir mão, ou pelo menos aprimorar as concepções sobre comportamento e cognitivismo.

> Quando defendemos os estudos de domínio como uma nova abordagem em CI, estamos sugerindo uma integração teórica de duas linhas contemporâneas de pesquisa que carecem de tal integração: estudos cognitivos e estudos bibliométricos, que chamamos de predecessores dos estudos de domínio. Essa integração é parcialmente alcançada mudando os pressupostos teóricos sobre a cognição. Dessa forma, a análise de domínio, construída em bases mais socioculturais, teorias pragmáticas e realistas representam uma teoria alternativa aos fenômenos cognitivos, e a este respeito, análise de domínio e cognitivismo não são dois pontos de vista complementares, mas dois pontos de vista teóricos mutuamente exclusivos. (HJØRLAND; ALBRECHTSEN, 1995, p. 413).

Observa-se que, no desenvolvimento de estudos analíticos de domínio, os pesquisadores têm um grande desafio a superar, no qual o conhecimento seja visto como processo de concepção coletiva, que reflita as características do domínio. Encontrar minuciosidades correlacionais presentes nos discursos que representem a especialidade sob alguns aspectos, como no caso da terminologia, apresenta-se como tarefa extremamente custosa, mas, para a análise de domínio, nesse ponto se concentra uma das contribuições para a área de CI, em que o cognitivismo puramente individual perde espaço para composições relacionais, capazes de revelar práticas informacionais e auxiliar no desenvolvimento de sistemas de recuperação da informação ou caracterizar um domínio.

## 3.4 Dimensão metodológica da análise de domínio em onze abordagens

Após Hjørland e Albrechtsen (1995) destacaram a análise de domínio e os motivos que os levam a crer na sua relevância, a dimensão metodológica da proposta começa a ganhar forma. Como já apresentado[31], os autores explanam sobre a

[31] Verificar nas páginas 46 e 47.

importância do ambiente, da influência social no entendimento semântico, na linguagem expressa, das visões de mundo e que cada indivíduo é carregado de impressões históricas, culturais e sociais. Tais acepções devem ser explicadas pela divisão social do trabalho e igualmente aplicadas à comunicação científica.

A abordagem de análise de domínio ficaria mais clara e funcional para a área da ciência da informação (GUIMARÃES, 2014) somente em 2002[32]. Nesse ano, Hjørland publicaria o artigo *Domain analysis in information science: Eleven approaches traditional as well as innovative*, no qual apresenta onze perspectivas para estudos analíticos de domínio. O entendimento é de que nas bibliotecas especializadas, os recursos informacionais devem ser identificados, descritos, organizados e disponibilizados para os objetivos específicos da área. Assim, não é possível abordar todos os domínios como sendo fundamentalmente semelhantes, de modo que uma abordagem teórica para a Ciência da Informação (CI) deve atentar-se às diferentes comunidades de discurso.

As onze abordagens propostas inicialmente por Hjørland (2002), tentam resolver esse problema, possibilitando analisar e caracterizar um domínio por meio delas: (1) produção de guias de literatura; (2) elaboração de classificações especiais e tesauros; (3) indexação e recuperação da informação; (4) estudos empíricos de usuários; (5) estudos bibliométricos; (6) estudos históricos; (7) estudos de documentos e gêneros; (8) estudos epistemológicos e críticos; (9) estudos terminológicos; (10) estruturas de instituições da comunicação científica; (11) cognição, conhecimento e inteligência artificial. As descrições compactadas das abordagens explicitadas abaixo são baseadas no próprio Hjørland (2002; 2017) e em Oliveira (2013), assim entendidas:

(1) Guias de literatura: também denominados de guias de fontes de informação ou guias de materiais de referência. São documentos que mostram e descrevem as principais publicações dentro de áreas específicas. É uma espécie de bibliografia dos documentos principais de um campo, que se concentra nas referências literárias primárias do domínio. Orienta os usuários nas escolhas, informando sobre pontos fortes e fracos das publicações, auxiliando também no processo operacional de pesquisa nas bases de dados informacionais.

---

[32] Verificar na página 47.

(2) Elaboração de classificações especiais e tesauros: apresentam os vocabulários específicos utilizados em diferentes domínios. A metodologia de construção desses instrumentos pode ser considerada um tipo de análise de domínio, mesmo implícita. São dicionários de sinônimos elaborados com base na abordagem facetada, organizados em uma estrutura de acordo com as relações semânticas entre os conceitos.

(3) Indexação e recuperação da informação: procuram organizar os documentos de forma a facilitar a visualização do possível conteúdo teórico e epistemológico da publicação. A distribuição das coleções deve propiciar maior facilidade na recuperação da informação, por isso, não se pode ignorar as diferentes demandas nos diferentes domínios.

(4) Estudos empíricos de usuários: verificação do comportamento dos usuários na busca da informação, mostrando peculiaridades das necessidades informacionais, bem como dados empíricos sobre o uso dos vários elementos de um sistema em comunidades distintas. Servem para direcionar os profissionais da informação na organização dos domínios, conforme as impressões e condutas desses usuários na utilização dos serviços. Muitos estudos em Ciência da Informação e nas ciências cognitivas buscam verificar a relação generalizada de pessoas com processos informacionais, esperando uma reação mecânica do indivíduo, sem considerar seus preceitos socioculturais, objetivos, valores e significado frente aos documentos.

(5) Estudos bibliométricos: vistos como campo fértil para pesquisas na área de CI e valiosa ferramenta que fornece importantes insumos para estudos de análise de domínio. Possibilita se encontrar conexões entre documentos, fatores de dependência ou ligação entre artigos, pesquisadores, campos, abordagens ou regiões geográficas. A bibliometria propicia estudos consistentes, aprofundados e relevantes para o mapeamento e visualização de domínios científicos, evidenciando correntes teóricas por meio de estudos de citação, cocitação ou colaboração científica.

(6) Estudos históricos: demonstram as tradições, os paradigmas, os documentos e formas de expressão, bem como suas correlações e influências mútuas no domínio. Embora não sejam muito utilizadas, pesquisas com perspectiva histórica permitem uma compreensão mais profunda e coerente e menos mecanicista, para se entender os documentos, a organização, os sistemas, o conhecimento e a informação. Não são estudos históricos comuns de domínios, mas estudos históricos que enfatizam o desenvolvimento da terminologia, categorias, literaturas, gêneros,

sistemas de informação etc., que podem ser encarados como abordagens para análise de domínio.

(7) Estudos de documentos e gêneros: verificam como estão organizados e estruturados os vários tipos de publicações em um domínio. Estudos quantitativos ou qualitativos dos diferentes gêneros de documentos em comunidades distintas têm informações para adequação dos serviços informacionais. Diferentes disciplinas ou comunidades de discurso desenvolvem tipos especiais de documentos, com adaptações às suas necessidades específicas (na Música: partituras; em Geografia: mapas e atlas; na lei: códigos; corpos de lei; em Astronomia: almanaques; etc.). A importância relativa dada a um tipo de documento varia, pois, dependendo do domínio, livros e periódicos podem ser considerados publicações formais ou não.

(8) Estudos epistemológicos e críticos: buscam analisar todos os pressupostos, explícitos ou implícitos, que embasam os paradigmas prevalecentes do domínio, procurando explicar claramente o arcabouço teórico da área. Revelam os fundamentos de um domínio, além de uma avaliação crítica sobre os conhecimentos específicos, segundo as suposições básicas sobre o conhecimento e a realidade. A epistemologia é entendida como a interpretação de toda experiência científica, produzida e coletada pelos pesquisadores. São estudos de base que fornecem diretrizes para a seleção, organização e recuperação da informação. É considerada a faceta principal, já que sem ela as outras abordagens são consideradas superficiais.

(9) Estudos terminológicos, linguagens para fins específicos, semântica do banco de dados e análise de discurso: são presentes e vistos na Ciência da Informação como temas complexos, porém importantes. Podem refletir o pensamento e a linguagem dos autores, tanto em bancos de dados quanto em textos livres. Possibilitam vislumbrar um sistema de recuperação da informação eficiente, sejam por via natural ou em ambientes com linguagem controlada. Destaque-se que esta pesquisa tem por finalidade uma metodologia para o levantamento terminológico por meio da mineração de texto, por isso, receberá especial atenção mais à frente.

(10) Estruturas e instituições da comunicação científica: a Ciência da Informação deve concentrar-se em pesquisas que investiguem as causas essenciais que tornam diferentes as estruturas, instituições e serviços nos domínios. O sistema de comunicação científica provavelmente será visto como mais estruturado e formalizado se comparado à área de humanidades, pois há critérios mais objetivos de análise. Estudos bibliométricos ganham destaque nessa abordagem por fornecer

informações sobre: maiores produtores; quanto publicam e em quais canais de distribuição; quais correntes epistêmicas são presentes no domínio; dados em relação às influências regionais; quais agentes e instituições se destacam etc. Entender o arranjo da divisão interna do trabalho nos domínios, as relações e compartilhamentos informacionais entre especialidades, permite compreender a função de documentos específicos e serviços de informação, possibilitando a elaboração de guias de literatura.

(11) Cognição profissional e representação do conhecimento em ciência da computação e inteligência artificial (IA): estudos analíticos de domínio são uma ferramenta utilizada na área de Ciência da Computação, de forma mais específica pela Engenharia de *Software*, que buscava reutilizar componentes comuns em programas considerados similares. Pesquisas nos campos da IA e ciências cognitivas com visões individualistas visam desenvolver sistemas especialistas, simulando mecanicamente o raciocínio humano, com auxílio da inteligência artificial.

Segundo Hjørland (2017, p. 437), “essas onze abordagens enfatizam que os objetos de estudo dos pesquisadores da informação são sociais e teóricos.” Nesse artigo, o autor ainda reconhece mais três abordagens, expressas como contribuições de Smiraglia (2015), no qual propõe uma taxonomia revisada, deixando de fora a indexação e recuperação de informações (abordagem 3) e a de estudos de estruturas e instituições na comunicação científica (abordagem 10); e sugerindo o acréscimo da semântica de banco de dados e análise do discurso.

O autor ainda julgou relevante os apontamentos de Guimarães e Tognoli (2015), que relatam a importância do conhecimento sobre procedência. Hjørland (2017) passou então a determinar que os estudos analíticos de domínio em Ciência da Informação podem se valer das quatorze abordagens (onze mais três). Para ele, as abordagens podem ser aplicadas também na área de organização do conhecimento, notadamente em sistemas de organização e processos de organização do conhecimento, ante uma perspectiva sociológica e epistemológica combinada, e enfatiza a importância do conhecimento do assunto.

## 3.5 Preceitos da abordagem terminológica

Embora trabalhe-se bastante com o termo análise de domínio, a metodologia levantada por Hjørland (2002), baseada em onze abordagens mais três, é bastante

ampla, necessitando de maior aprofundamento para ser discutida como um todo. Como esse não é o escopo principal do estudo, atém-se aos preceitos da linguagem presentes nos discursos, notadamente para a abordagem terminológica, a qual será investigada por meio de técnicas da mineração de textos, portanto, discorre-se a seguir um apanhado moderado dos princípios que regem os constructos desse tema. A parte histórica ficará de fora, por se entender que não há necessidade de escrutínio desse tópico, já que também não é desígnio desta pesquisa.

Sager (1990, p. 2) define que terminologia é estudo e o

> campo de atividade relacionado com a coleta, descrição, processamento e apresentação de termos, ou seja, itens lexicais pertencentes a áreas especializadas de uso em um ou mais idiomas. Em seus objetivos, é semelhante à lexicografia, que combina o duplo objetivo de coletar dados gerais sobre o léxico de um idioma com o fornecimento de um serviço informativo e, às vezes, até consultivo aos usuários do idioma.

Pêcheux (1997) ilustra que, antes da introdução da ciência linguística, os estudos da língua, na maioria das vezes, eram feitos por meio de questões de compreensão do texto e da gramática nas propriedades normativas ou descritivas, desenvolvendo atividades normalmente direcionadas para indagações como: sobre o que este texto fala? Quais as ideias principais do texto? O texto está em conformidade com as normas da língua em que foi escrito? Essas questões eram colocadas de forma simultânea porque se entrelaçavam, contudo, pontos de abrangência sintática e semântica evidenciadas no texto também ajudavam nas resoluções dos itens a respeito do sentido do texto: O que o autor quis dizer?

O autor remete a Saussure e sua obra póstuma, Curso de Linguística Geral, de 1916, entendendo que as ideias e os temas apresentados na obra introduzem a Linguística como ciência e apontam a linguagem como componente crucial da comunicação humana. Segundo Pêcheux (1997), a linguagem deve ser vista como um sistema e não como tendo a função de exprimir sentido, objeto em que a ciência se insere para descrever seu funcionamento. Em suma, trata-se de analisar a linguagem não pela sua função, mas pelo seu funcionamento. Encaminhamentos que ignoram o significado individual de cada parte nas implicações linguísticas e priorizam os princípios subjacentes que tornam cada parte possível, quer ela se realize ou não.

Mesmo com a dimensão científica da Linguística, alguns pontos não foram abarcados e carecem dos estudos da linguagem e suas expressões, o que direciona novamente a diferentes formas de uma mesma questão: o que esse texto quer dizer?

Que significação contêm esse texto? A essência desse texto difere-se da essência daquele outro texto? Questões em que a análise de conteúdo ou análise de texto se adequam. (PÊCHEUX, 1997). Os dois métodos de estudo da linguagem que seguem são considerados não-linguísticos e derivam de metodologias psicológicas ou sociológicas direcionados à linguagem.

Um dos procedimentos para análises de conteúdo e de texto é o método de dedução de frequência. Definição encontrada em Pêcheux (1997) caracteriza-o como o processo de contar as ocorrências de um mesmo signo linguístico (palavra ou letra) dentro de uma sequência de comprimento fixo e de determinar uma frequência que possa ser comparada a outras, para fornecer um teste de comparação entre vários elementos da sequência ou entre sequências paralelas para o mesmo item. Embora considerado de nível mais baixo, por ater-se puramente à existência do elemento linguístico e não ao funcionamento do sistema, é visto como importante para a área por contribuir no desenvolvimento de ferramentas estatísticas para tratamento da informação.

O segundo método é a análise por categorias temáticas. Se o primeiro método é estatístico, esse segundo baseia-se na análise de conteúdo e pode responder questões de sentido do texto ou das diferenças entre um texto e outro. É um nível mais alto que o primeiro, em que uma série de significações são detectas por meio de indicadores interligados: relação funcional; expressão da significação/meios de expressão. Como o procedimento era feito de forma manual para categorização de determinado segmento, um codificador humano fazia o julgamento através das significações presentes em um quadro de análise (presença, ausência e intensidade de termos) e apontava uma das classes de equivalência já definidas.

Os métodos apresentados são baseados em levantamento de termos (metodologia proposta nesta pesquisa), pois os pressupostos da Socioterminologia de Gaudin (1993, 2014) e Faulstich (2006) e a Teoria Comunicativa da Terminologia (TCT) de Cabré (2005), reiteram a relevância dos levantamentos terminológicos, visto que tomam o texto como objeto central das investigações. Tais procedimentos são confluentes por basearem suas análises em unidades de significação nos textos sem se esquecer do contexto sociocultural, inclinado para o descritivismo (aceita mais de

uma definição), distanciando-se da terminologia clássica[33]. Percebe-se aqui similaridade substancial com o paradigma social da Ciência da Informação (CI) e a teoria da Análise de Domínio (AD), no que se refere a estudos mais abrangentes e tendências socioculturais.

Três aspectos são basilares no modelo sociolinguístico da terminologia segundo Gaudin (2003): a dimensão social, o funcionamento discursivo dos termos (esses dois aspectos servem a esta pesquisa) e o recorte diacrônico em terminologia. Creditando o termo como um signo dinâmico nas linguagens de especialidades, Faulstich (1999) aponta, então, para a teoria da variação terminológica, a qual possibilita a mudança no uso real dos termos por meio de análises terminológicas em contextos linguísticos e contextos discursivos e afasta-se da delimitação unívoca de termo-conceito-significado.

Área correlata que vem sendo bastante estudada e que enxerga os processos computacionais como indispensáveis é a Linguística de Corpus, a qual compreende a linguagem sob perspectivas probabilísticas, inferindo que as ocorrências não acontecem de forma aleatória, tornando possível evidenciar e quantificar padrões com uso de ferramentas estatísticas. Tais avanços no campo da Linguística se deram por meio de recursos tecnológicos, que permitem analisar milhões de dados de forma quantitativa e qualitativa. (MARIAN, 2015).

Para Hjørland (2022, n.p.), “o termo terminologia tem dois sentidos: o corpo de termos usados dentro de uma disciplina e o campo de estudo dedicado ao estudo da terminologia no primeiro sentido.” Para o autor, a terminologia como campo de estudo se relaciona com outros campos menores, como tradução, linguagem para fins especiais, lexicografia etc., bem como a campos principais, como Linguística, Estudos Cognitivos, Sociologia e Filosofia (da ciência). O autor ainda define terminologia como o vocabulário associado a um determinado domínio, considerando-a uma espécie de linguagem especial do domínio. Portanto, levantamentos terminológicos são importantes sob diversos aspectos em diferentes áreas

A presente pesquisa faz uso desse elemento, uma vez que as investigações se darão nos discursos dos resumos de comunicações científicas, principalmente em relação ao levantamento dos termos mais recorrentes (sem pormenorizá-los) que

[33] Metodologia prescritivista, que tem como propósito estabelecer relações monorreferenciais e unívocas, resultando em termos normalizados e padronizados e que não admitem variações.

também serve de subsídio para definição e relação temática entre os textos. Análises de conteúdo e de texto baseada no método de dedução de frequência e coocorrência de termos e a análise por categorias temáticas (PÊCHEUX, 1997) são amparados pela importância de levantamentos terminológicos, que enxerga os textos como elementos nucleares nas pesquisas (GAUDIN, 1993, 2014; CABRÉ, 2005; FAULSTICH, 2006).

## 3.6 A mineração de texto

A Mineração de Texto (MT) (*Text Mining* – TM) tem seu surgimento ancorado na Mineração de Dados (MD) (*Data Mining* – DM), é como se a mineração de texto fosse uma etapa da mineração de dados (FAYYAD *et al.* 1996; WEISS *et al.* 2005). Não é difícil perceber que os dois sistemas apresentam várias semelhanças, por isso, alguns conceitos são apresentados em paralelo, já que muitos métodos para descoberta de conhecimento em texto provêm da mineração de dados, diferenciando-se apenas em procedimentos peculiares do tipo de dado. A maioria dos sistemas possuem rotinas de pré-processamento, algoritmos de descoberta de padrões e elementos de apresentação, como ferramentas para navegação e visualização dos resultados. (FELDMAN; SANGER, 2007).

A mineração em ambientes digitais teve como base o aprendizado de máquina e a estatística; e segundo Witten (2004), por ter surgido primeiro, há mais literatura produzida sobre mineração de dados do que sobre mineração de texto. Enquanto a mineração de dados acompanhou alguns avanços tecnológicos na década de 1990, estabelecendo-se de vez como uma tecnologia prática e utilizável, a mineração de texto realizou seus *workshops* em momentos menos propícios da história, como em julho de 1999, na *International Machine Learning Conference*, e em agosto do mesmo ano, na *International Joint Conference on Artificial Intelligence*, ocorridos pouco antes do *crash* das empresas “ponto.com”[34].

Na mineração de dados, pressupõe-se que eles estão armazenados em um formato estruturado, por isso, grande parte do trabalho de pré-processamento está baseado em duas tarefas: limpeza e normalização dos dados e junções das tabelas.

---

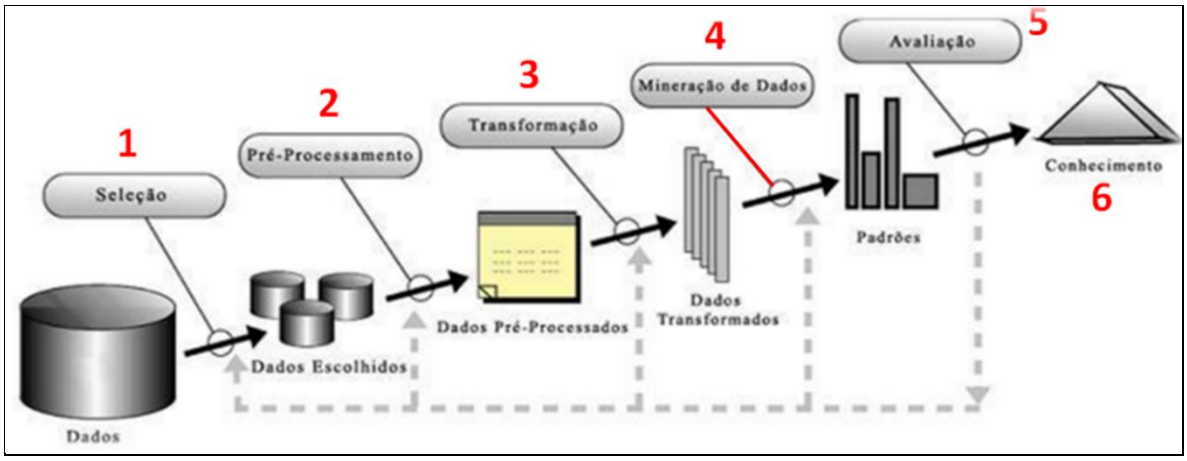
[34] A bolha das companhias de informática ocorreu entre os anos de 1995 e 2000. Anos em que as bolsas de valores dos países industrializados tiveram forte alta nos preços de empresas do setor. Investidores despreparados, avaliações erradas, corrupção corporativa e outros fatores econômicos levaram muitas empresas a quebrar, e março de 2000 decretou-se o “*crash* das ponto.com”.

Nos sistemas de mineração de texto, as operações de pré-processamento focam em identificar e extrair características relevantes para documentos em linguagem natural. São operações responsáveis por transformar dados não estruturados de coleções de documentos em um formato intermediário, algo irrelevante nos sistemas de mineração de dados. (FELDMAN; SANGER, 2007).

Tanto a mineração de dados quanto a mineração de texto são empregadas em um processo conhecido como Descoberta de Conhecimento (*Knowledge Discovery – KD*), que, de forma simplificada, é entendido como um método de extração de padrões válidos e conhecimentos ocultos em grandes bases de dados ou de textos. Fayyad, Piatetsky-Shapiro e Smyth (1996) explicam que tais padrões devem ser confiáveis, compreensíveis e úteis, para que o conhecimento obtido possa ter seu uso, científico ou comercial, aproveitados. São estabelecidas métricas com estimativas estatísticas para definir a utilidade desses padrões, tais como: níveis de confiança, compreensão e utilidade, as quais não serão aqui discutidas, por não se tratar do foco da pesquisa.

Demonstram-se, na Figura 1, as fases de composição dos processos de descoberta de conhecimento: a escolha de uma amostra de dados, o pré-processamento, a transformação dessa amostra, a mineração através da manipulação com algoritmos (MD ou MT), a devida interpretação e avaliação dos resultados, culminando no conhecimento de informações antes desconhecidas. Segundo Fayyad, Piatetsky-Shapiro e Smyth (1996), antes de se iniciar o processo de KD convém inteirar-se do domínio.

Figura 1 – Etapas do fluxo de processos de KD (D ou T)




Fonte: Adaptado de Fayyad *et al.*, (1996, p. 84).

A Etapa 1 apresentada na figura corresponde ao processo de seleção dos dados, em que se define o tipo de mineração que se pretende utilizar (dados ou texto).

Compreende ainda a questão de selecionar o conjunto de dados ou focar em um subconjunto de variáveis ou amostras, deixando-os prontos para submissão aos processos. Na Etapa 2 executam-se a limpeza e o pré-processamento de dados, que normalmente incluem remover distorções, coletar as informações necessárias para modelar ou contabilizar tais distorções e decidir sobre estratégias de como lidar, por exemplo, com ausência de dados. (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

Na fase seguinte (Etapa 3), acontece a redução e projeção de dados, como por exemplo, a remoção de palavras sem relevância, elementos tipográficos (diacríticos, pontuação, números, caracteres especiais, sublinhados, espaçamentos etc.), além de padronização do texto (maiúsculo/minúsculo) [35]. A Etapa 4 é o momento em que se combina os objetivos a serem alcançados com os métodos de mineração específicos, por exemplo, sumarização, classificação ou agrupamento. Realiza-se então os processos exploratórios, aplicando os algoritmos de mineração com os métodos selecionados para a busca de padrões[36].

Ainda conforme Fayyad, Piatetsky-Shapiro e Smyth (1996), na Etapa 5 são feitas as interpretações dos padrões extraídos, e, se necessário, retorna-se a qualquer uma das etapas de 1 a 4 para mais iterações. A Etapa 6 é a fase que ocorre a ação sobre o conhecimento descoberto, em que é possível usá-lo diretamente, incorporá-lo a outro sistema para ação futura ou simplesmente documentá-lo e relatá-lo às partes interessadas.

Nesse conjunto de tarefas, a Etapa 4 (mineração) tem papel de destaque, pois é considerada a parte mais importante do processo, uma vez que nessa fase definem-se os métodos para se aplicação dos algoritmos escolhidos. (FAYYAD *et al.*, 1996; BERRY; LINOFF, 1997; HAN; KAMBER, 2006). Berry e Linoff (1997, p. 7) definem que a mineração de dados “é a exploração e análise de grandes quantidades de dados para descobrir padrões e regras significativos.” São empregues processos emprestados da Estatística, Ciência da Computação e do aprendizado de máquina (*machine learning)*[37]; e a escolha de técnicas específicas dependem da natureza do trabalho, do tipo de dados disponíveis e das preferências do profissional.

---

[35] As fases do processo de 1 a 3 serão detalhadas no item 3.7.1 deste estudo.
[36] Mais detalhes nos itens de 3.7.2 a 3.7.7.
[37] Um dos recursos da inteligência artificial que utiliza algoritmos e pode aprender e alterar seu comportamento de forma autônoma. Visa melhorar o desempenho de tarefas ou tomar decisões apropriadas em diversos contextos. Essas regras são geradas com base na própria experiência adquirida e do reconhecimento de padrões dos dados analisados nos treinamentos.

Os procedimentos da mineração de dados ou de textos podem ser usados para aperfeiçoar o desempenho de processos industriais, ou de negócios que possuem um volume muito grande de informações. Da mesma forma, são empregues em pesquisas científicas, como instrumento de análise e descoberta de conhecimento, resultante das observações dos dados ou dos ensaios. Encontrar padrões nos conjuntos de dados ou de textos tem uma longa tradição na área acadêmica, inicialmente, no campo da estatística e, mais recentemente, na inteligência artificial. Em teoria, a mineração pode ser aplicada a qualquer tipo ou quantidade de dados, mas, normalmente, é utilizada em grandes volumes (FAYYAD, *et al.*, 1996; BERRY; LINOFF, 1997; WEIS *et al.*, 2005; HERRERA VARELA, 2006).

Segundo Herrera Varela (2006), o estabelecimento do mundo tecnológico, permitiu às instituições realizarem a coleta de qualquer tipo de informação presente na rede (Word Wide Web), por se acreditar na possibilidade de se encontrar algo valioso entre esses *bits*. A tecnologia facilitou também o processo de transformação das informações, tornando-as digitais e digitalizáveis, e igualmente acarretou a queda gradativa no custo de armazenamento.

> É por isso que, embora a mineração de dados possa ser aplicada a qualquer tipo de informação, variando apenas as técnicas a serem usadas em cada tipo de estrutura de dados analisada, principalmente a extração de dados em bancos de dados relacionais, bancos de dados espaciais, banco de dados temporais, bancos de dados documentais e bancos de dados multimídia**,** também tem havido uma forte tendência desde o advento da Internet para extrair informações especialmente da World Wide Web**.** (HERRERA VARELA, 2006, p. 125).

Para Lobaina e Suárez (2018), a mineração no ambiente digital é um processo revolucionário, constituindo-se na maneira mais rápida de se analisar grandes volumes de dados para encontrar padrões e novos conhecimentos. É uma técnica já utilizada em muitas outras áreas, mas ainda incipiente na área das bibliotecas. Pode-se encontrar ainda na literatura o termo bibliomineração, utilizado pela primeira vez por Nicholson e Stanton (NICHOLSON, 2004), os quais definem o termo como a mineração de dados aplicada à biblioteca.

A maioria dos trabalhos que contêm os termos biblioteca e mineração de dados não se referem a dados de bibliotecas tradicionais, mas sim a bibliotecas especializadas no contexto dos *software*. Segundo Langridge (1977, p. 81) bibliotecas especializadas “enfatizam uma área de conhecimento ou servem a um grupo especial de pessoas (frequentemente com o mesmo interesse).” O termo *bibliomining* foi criado

a fim de torná-lo mais propício para profissionais preocupados com a mineração de dados em ambientes de biblioteca, os quais buscam trabalhos, pesquisas e autores em um campo em que os padrões de comunicação científica se encaixem na bibliometria. (NICHOLSON, 2004).

Consoante aos procedimentos de mineração em banco de dados e técnicas de mineração aplicadas em bibliotecas especializadas, a Mineração de Texto (MT) tem se destacado por propiciar extração de informações de elementos textuais, presentes de maneira massiva no universo digital. (HERRERA VARELA, 2006). Surgiu especificamente para tratar dados e informações textuais (não-estruturadas ou semiestruturados), pois o nível de complexidade nesse tipo de representação de informação é considerado alto. (FELDMAN; DAGAN, 1995).

Dada a amplitude de criação e disseminação de textos ou de arquivos textuais no ambiente Web, compreende-se que estudos desses dados na atualidade é algo salutar. Normalmente, são dados que não possuem estruturação definida em termos de armazenamento ou recuperação, e que geralmente estão nas próprias páginas de internet, ou em repositórios de acesso livre ou pagos. Aranha e Passos (2006, p. 1) complementam que a mineração de texto foi “inspirada pelo *data mining* ou mineração de dados, que procura descobrir padrões emergentes de banco de dados estruturados, a mineração de textos pretende extrair conhecimentos úteis de dados não estruturados ou semiestruturados.”

Para os autores a mineração de texto é um conjunto de técnicas que buscam encontrar regularidades, padrões ou tendências em textos dispostos em linguagem natural, usadas para navegação, organização e descoberta de informação em bases textuais, vista como extensão da mineração de dados. É também chamada de Mineração de dados de texto (*Text Data Mining – TDM*) e Descoberta de Conhecimento em Textos (*Knowledge Discovery in Texts - KDT*). (FELDMAN; DAGAN, 1995).

Sob a égide da busca de informações relevantes nos meios digitais, minerar dados em e-mail, textos livres, arquivos eletrônicos, páginas web, redes sociais, arquivos digitalizados ou fóruns de debates têm forte apelo e destaque. A mineração de texto possui procedimentos que vêm ganhando visibilidade e interesse, e “tem sido escolhida pelo mundo acadêmico como uma linha útil de pesquisa e há um interesse crescente da comunidade acadêmica no processamento automatizado de linguagem natural.” (URBIZAGASTEGUI-ALVARADO, 2021, p. 3). Segundo o autor, técnicas de

mineração de texto combinadas com processos de análises bibliométricas podem ser exploradas para descobrir padrões desconhecidos em diversos domínios.

Nessa mesma linha, Salazar-Lópes *et al.,* (2020, p. 3) assinala que a mineração de texto “faz análises rápidas de grandes volumes de informação textual, assim, em estudos científicos é possível mostrar por meio de gráficos o que se tem produzido em determinado tema, mesmo que este novo conhecimento nem sempre esteja expresso no resumo ou nas palavras-chave.” Os autores acreditam que, ao utilizar ferramentas computacionais para mineração nos textos, é possível encontrar conceitos e hipóteses sobre a temática pesquisada, e continuam argumentando sobre o uso da computação para auxiliar as análises, complementar ou “fazer análises preliminares em estudos bibliométricos e em revisão sistemática, que precisam ser analisadas às vezes em inúmeros artigos, além de equiparar técnicas de análise de mineração de texto com técnicas de análise estatística.”

Embora tanto a mineração de dados quanto a de textos sejam façam parte do processo de descoberta de conhecimento e busquem informações potencialmente úteis, Witten (2004) destaca diferença substancial entre elas. Na mineração de dados, ocorre a extração de informações implícitas, previamente desconhecidas, pois a informação está: oculta, desconhecida e dificilmente seria extraída sem recorrer a técnicas automáticas. Na mineração de texto, a informação está explicitada no discurso, mas as restrições de recursos humanos a tornam inviável para as pessoas lerem o texto por si mesmas, porque as informações não possuem uma estrutura passível de processamento automático. Apesar da diferença filosófica, do ponto de vista computacional, os problemas se equivalem.

Outra equivalência entre mineração de dados e de texto é que as informações extraídas tenham caráter de utilidade. Isso significa que as descobertas devem possibilitar ações a serem tomadas.

> No caso de mineração de dados, essa noção pode ser expressa de maneira relativamente independente de domínio: padrões acionáveis são aqueles que permitem que previsões não triviais sejam feitas em novos dados da mesma fonte. O desempenho pode ser medido pela contagem de sucessos e fracassos, técnicas estatísticas podem ser aplicadas para comparar diferentes métodos de mineração de dados no mesmo problema e assim por diante. No entanto, em muitas situações de mineração de texto, é muito mais difícil caracterizar o que “acionável” significa de maneira independente do domínio específico em questão. Isso torna difícil encontrar medidas justas e objetivas de sucesso. (WITTEN, 2004, p. 2).

Portando, compreende-se que a mineração de dados é direcionada para dados estruturados e a mineração de texto é inclinada para os dados textuais (textos, frases, palavras) não estruturados. Do mesmo modo, entende-se que a essência é a mesma: extrair conhecimento que normalmente não seriam recuperadas utilizando métodos tradicionais de consulta. Segundo Witten (2004) sistemas de mineração de texto normalmente executam seus algoritmos de descoberta de conhecimento em coleções de documentos preparadas, por isso, uma fase importante no processo de mineração deve ser dedicada ao que comumente refere-se como operações de pré-processamento. Tais operações incluem diferentes técnicas, selecionadas e adaptadas da recuperação de informações, extração de informações e pesquisa em Linguística Computacional, que transformam conteúdo bruto, não estruturado e de formato original em um formato intermediário.

Conforme Aranha e Passos (2006), 80% do conteúdo digital está em formato de texto, e 80% das informações armazenadas nas empresas não são estruturadas. A mineração de texto pode ser empregada em qualquer domínio em que se tenha acesso aos conteúdos textuais, e que se busquem informações específicas, permitindo tanto análises qualitativas quanto quantitativas, as quais, segundo Moura (2009), podem ser feitas obedecendo a regras semânticas ou por abordagens estatísticas.

Análises semânticas verificam os termos em relação ao contexto e normalmente utilizam técnicas de Processamento de Linguagem Natural (PNL), inferindo caráter qualitativo aos resultados. Estudos estatísticos podem revelar informações conforme presença ou ausência dos termos nos textos. No aspecto qualitativo, os termos são considerados por algoritmos de base booleana ou binária para essa verificação e, na análise quantitativa, os algoritmos fazem estimativas de frequência relativa ou normalizada em relação ao texto e ao corpus.

## 3.7 Técnicas para descoberta de conhecimento em texto

Conforme Wives (2004), existem diversas técnicas para descoberta de conhecimento em texto, como: detecção de sentenças; análise da sintaxe e da estrutura das sentenças; análise de agrupamento ou conglomerados (*clustering*); classificação; sumarização; análises qualitativas, quantitativas; e identificação de regras de associação e padrões. Como existem muitas técnicas e ainda inúmeras

variações de métodos dentro delas, buscou-se destacar apenas os procedimentos significativos para este estudo, sem muito se aprofundar a eles, uma vez que o modo de funcionamento e os cálculos realizados pelos algoritmos não fazem parte dos objetivos da pesquisa.

Os algoritmos aqui utilizados estão bem estabelecidos e são amplamente aplicados na área de mineração de texto, portanto, pretendeu-se realizar breve apresentação para obter-se uma visão geral dos procedimentos, sem estender-se aos assuntos ou explorá-los minuciosamente. Em suma, os algoritmos pertinentes a este estudo procuram, por meio de cálculos estatísticos e probabilísticos, encontrar padrões nos discursos como: características textuais das sentenças; frequência dos termos, coocorrências de palavras para composição dos n-gramas; bem como verificar as principais temáticas e articulações entre os resumos, além de realização de processo de *stem* (redução ao radical), que pode ser utilizado em outras em pesquisas.

### 3.7.1 Etapas de pré-processamento e transformação

A principal diferença entre a mineração de dados e mineração de textos encontra-se na etapa de pré-processamento, pois um banco de dados possui uma estrutura de armazenamento. Na mineração de texto, os dados são considerados não estruturados e a literatura aponta para a necessidade de se adequar minimamente a os documentos escritos em linguagem natural, a fim de possibilitar a limpeza e padronização dos textos e, consequentemente, viabilizar a extração de conhecimento. Nesse sentido, os textos devem passar por dois processos vistos como essenciais, no qual o primeiro é o processo de conversão do texto a um formato intermediário e o segundo é o que encontra os padrões ou conhecimento (destilação do conhecimento) nos documentos. (TAN, 1999).

Para que os textos sejam minimamente estruturados e estejam na forma intermediária, primeiramente, se faz seu redimensionamento, reduzindo-o às frases, e posteriormente, as frases são reduzidas a palavras (*tokens*). As sentenças podem demonstrar um padrão de escrita do grupo de autores, pois é possível verificar a quantidade de termos em cada frase que constituem os discursos. Na mineração de texto, as palavras são conhecidas como *tokens* e, assim como na escrita, podem ter formato simples ou composto, de modo que cada termo pode ser representado por um ou mais *tokens* e um *token* pode fazer parte de um ou vários termos.

Na etapa de transformação dos textos, algumas alterações são requeridas para facilitar a semiestruturação e uma delas é o processo que deixa as letras em um só formato, todas em maiúsculo ou todas em minúsculo. Esse procedimento ajuda nos processos de comparação dos textos. Outra transformação comum é a remoção de palavras irrelevantes para as análises, conhecidas como *stop words*. É normalmente uma lista de palavras auxiliares ou conectivas como interjeições, preposições, artigo, pronomes e outras que não possuem significado em si e não fornecem nada de discriminativo do texto. (FRAKES; BAESA-YATES, 1992; YANG; PEDERSEN, 1997). Destaque-se que, dada a diferença entre as línguas, as listas que contém essas palavras também se diferem dentro das ferramentas de mineração.

Do mesmo modo, existem particularidades em cada domínio, por isso, em algumas especialidades é comum a existência de uma outra lista com palavras específicas, que também podem ser desconsideradas. Na ferramenta utilizada neste estudo (Knime[38]), é possível criar uma tabela com essas palavras pertinentes ao domínio, e realizar a remoção delas junto com a lista de *stop words,* que é própria (nativa) da ferramenta. Esse processo de eliminação de palavras consideradas sem importância, tem como propósito refinar os termos à um conjunto sintético, mas representativo na coleção de documentos, diminuindo significativamente a quantidade de termos e o custo computacional[39] nas etapas seguintes. (MANNING; RAGHAVAN; SCHÜTZE, 2008).

### 3.7.2 Frequência e coocorrência de termos

Se um termo possui apenas um *token*, ele recebe a denominação de unigrama e é considerado um termo simples. Os termos compostos recebem o nome de n-gramas e podem ser constituídos por 2 *tokens* (bigrama), 3 *tokens* (trigamas) e assim consecutivamente. Com o processo de tokenização, o texto é reduzido a palavras ou *tokens* sem interpretação semântica, mas as ferramentas estatísticas conseguem

---

[38] KNIME – Analytics Platform – é um *software* de código aberto, modular e escalável que abrange vários módulos de carregamento de dados, transformação, análise e exploração visual. A primeira versão foi lançada em julho de 2006, mas o projeto nasceu no início de 2004, na Universidade de Konstanz, no sul da Alemanha, quando uma equipe de desenvolvedores de uma empresa de *software* do Vale do Silício, especializada em aplicações farmacêuticas, começou a trabalhar em uma nova plataforma como ferramenta de colaboração e pesquisa. O software pode ser baixado diretamente no site: https://www.knime.com/downloads e seu uso é livre. A versão do Knime utilizada neste estudo foi a 4.6.2, atualizada em 21 de setembro de 2022.

[39] Custo computacional tem relação ao tempo de processamento de determinada tarefa, que por sua vez está ligado ao uso de recursos do equipamento.

analisar e inferir significado mesmo quando examinados isoladamente. (LOPES, 2004).

Uma das maneiras de se fazer o levantamento e cálculo de frequência dos unigramas e ainda verificar o padrão de coocorrência de bigramas ou trigramas no *corpus* se dá por meio de medidas estatística simples. Alguns algoritmos baseiam-se em medidas como TF (*term frequency*) ou DF (*document frequency*). A medida *term frequency* determina a frequência absoluta de presença de um termo específico na coleção das publicações, enquanto a *document frequency* calcula o número de artigos em que um termo distinto está presente. O TF faz a contagem das ocorrências de determinada expressão independentemente para cada documento e utiliza esses algarismos como uma medida numérica, que são normalizados para valores no intervalo entre 0 e 1. (WITTEN, 2004).

### 3.7.3 Representação e visualização de dados textuais

Depois da realização do pré-processamento e extração de termos que representam a coleção documental, busca-se então pelos padrões que podem revelar conhecimentos úteis ainda desconhecidos. Como não serão utilizados algoritmos com poder de varredura semântica, pode-se representar os dados extraídos pelos algoritmos estatísticos no formato *bag of words*, na qual os *tokens* são tidos como independentes, gerando um conglomerado desarrumado, em que a ordem de ocorrência dos termos é irrelevante.

O modelo espaço vetorial é um dos mais utilizado para representação de dados textuais e recuperação da informação. Nesse modelo, cada documento é visto como um vetor em um espaço multidimensional e cada dimensão é um termo do *corpus*. Os documentos são representados como um vetor que tem suas dimensões definidas pelos termos presentes no conjunto inicial de documentos. Cada posição do vetor é um *token* com um valor numérico (atribuição de peso ou medida) que demonstra sua representatividade para o documento de modo a destacar os termos mais importantes. (FELDMAN; SANGER, 2007; LU; WOLFRAN, 2012). Quanto maior o peso associado à coordenada do vetor, mais relevante a expressão é para o texto. A atribuição de pesos aos *tokens* serve, por exemplo, para calcular o grau de aproximação entre o que se indica para a busca e um possível documento encontrado. (LOPES, 2004).

### 3.7.4 Composição dos n-gramas

A técnica de mineração denominada *N-Grams* ou N-Gramas é bastante utilizada, e tem como finalidade fazer o agrupamento de palavras que aparecem juntas no texto, ou seja, são termos ou *tokens* compostos que expandem as investigações das ocorrências para essas combinações conjuntas (coocorrências) das palavras e não apenas aos *tokens* simples. (WITTEN, 2004). Representar os dados utilizando uma *bag of words* (pacote de palavras) apresenta apenas as palavras presentes nos textos ou suas combinações sem considerar o enquadramento semântico ou a sequência em que ocorrem. Essas palavras ou combinações são chamadas de n-gramas (*n-grams*) (MOURA *et al.*, 2010), os quais podem ser um unigrama como “pesquisa”, um bigrama como “pesquisa ciência” ou ainda uma trigrama “pesquisa ciência informação.”

### 3.7.5 Modelagem de tópicos

Uma das maneiras de se verificar as similaridades textuais nos discursos é por meio de algoritmos que analisam temáticas ou tópicos correlacionais do conjunto de documentos. A técnica de extração de assuntos conhecida como Modelagem de Tópicos baseia-se em algoritmos que entendem cada documento como uma mistura de tópicos, e que cada tópico é um conjunto probabilístico de palavras significativas (STEYVERS; GRIFFITHS, 2007; LU; WOLFRAN, 2012), em que, ao se identificar esses tópicos ocultos, é possível captar o significado dos textos.

Um dos algoritmos largamente utilizado na mineração de texto para este fim é o de Alocação Latente de Dirichlet (*Latent Dirichlet Allocation* – LDA), que “é um modelo probabilístico generativo de um *corpus*. A ideia básica é que os documentos sejam representados como misturas aleatórias sobre tópicos latentes, onde cada tópico é caracterizado por uma distribuição sobre palavras” (BLEI; NG; JORDAN, 2003, p. 996), logo, o algoritmo realiza os cálculos e faz o agrupamento dos tópicos encontrados no conjunto de documentos.

Segundo os autores, o LDA encontra grupos de palavras relacionadas de duas formas: na primeira o algoritmo atribui cada palavra a um tópico de forma aleatória e o usuário define o número de tópicos que quer encontrar. O algoritmo mapeia os documentos do *corpus* de forma que as palavras em cada texto sejam capturadas por tópicos criados por fórmulas matemáticas. Na segunda, o algoritmo percorre cada palavra iterativamente e reatribui a palavra a um tópico, considerando a probabilidade

de que a palavra pertença a um tópico e a probabilidade de que o documento seja gerado por um tópico. Os cálculos são feitos inúmeras vezes até que todas as premissas do algoritmo sejam satisfeitas.

### 3.7.6 Agrupamento ou *clustering*

O processo de *clustering* tem ganho bastante destaque na criação de taxonomias, estudo de ontologias e na recuperação da informação. Quando um usuário realiza a busca por um documento específico, hipoteticamente ele pode estar interessado em outros documentos que tratam do mesmo assunto ou que façam parte de um mesmo contexto. O *clustering* tem como finalidade fazer o agrupamento de objetos, que são reunidos por meio de uma medida de proximidade. Os grupos são formados por elementos com características altamente idênticas entre si, mas opostas em relação aos elementos dos outros conjuntos. Em outros termos, ele maximiza a paridade intragrupos (relação dentro dos grupos) e minimiza a paridade intergrupos (relação entre diferentes grupos). (EVERITT *et al.*, 2011).

Em grandes quantidades de dados, os cluster são conjuntos que fornecem uma visão mais fracionada em relação ao todo, permitindo entendimentos e usos mais específicos e eficientes em determinadas situações. É considerada uma técnica interessante, pois, possibilita encontrar traços relacionais entre documentos que aparentemente não tinham conexão; e organiza-os em *clusters* conforme apontamentos probabilísticos de semelhança, apenas pelas regularidades dos dados, sem uso de conhecimento ou interferência externa. Assim, os tópicos ou temáticas descobertas pelo algoritmo conseguem estabelecer interligações e apontar documentos correlatos. (ZONG; GOSH, 2003).

O algoritmo LDA serve também a esse processo, visto que os cálculos e ações executadas por ele permitem indicar uma categorização de documentos baseada nos tópicos extraídos do *corpus*. Essa função viabiliza a verificação das possíveis correlações temáticas e interlocuções presentes nos discursos, apresentando os grupos de artigos pertencentes a cada conjunto de assunto detectado.

### 3.7.7 Processo de *stemming*

É comum encontrar na literatura os termos lematização, stematização ou *stemming* como processos equivalentes, entretanto, há que se tomar certo cuidado, pois existem algumas diferenças entre eles. A técnica de lematização utiliza-se de

comparações semânticas, portanto é mais complexa. Como no Knime ainda não há suporte de lematização para termos na língua portuguesa e por não ser intenção deste estudo comparar as técnicas, a lematização não será aqui explorada. De forma simplificada, essas técnicas são utilizadas para redução de um subconjunto de termos com proximidade gramatical a uma forma canônica inicial comum (derivação ou radicalização do termo) que possa representá-los.

O *stemming* é uma técnica bastante utilizada na descoberta de conhecimento em dados não estruturados, pois é possível mapear palavras semântica ou morfologicamente relacionadas, agrupando termos de mesmo sentido, inferindo que palavras de mesmo radical possuem significado semelhante, já que podem variar conforme seus afixos (prefixos e sufixos). É aplicado sobre cada palavra separadamente e o objetivo é reduzir as expressões para uma representação mais simples, conhecida como *stem* (raiz). (PORTER, 1980; FRAKES; BAEZA-YATES, 1992; BERRY, 2004).

Por exemplo, as palavras: escrevi, escreveu, escrever, escrevemos e escrevendo seriam reduzidas e agrupadas apenas no radical “escrev”. A técnica de *stemming* permite encontrar os vocábulos primários e identificar a forma raiz das palavras, anteriores às suas variações. É possível também verificar a formação de termos compostos (n-gramas) que possuem significado semântico igual. (MANNING; RAGHAVAN; SCHÜTZE, 2008). Tais processos podem melhorar os resultados de recuperação da informação, padronizando e diminuindo as variações dos termos tanto na indexação quanto nos processos de busca. (PORTER, 1980; BERRY, 2004).

Como o stemming é baseado na Linguística, ele é totalmente dependente do idioma e, talvez pela complexidade característica da língua portuguesa, poucos algoritmos dessa técnica são direcionados a ela. Existem diversos algoritmos de radicalização (stemming) de palavras, mas o mais conhecido e considerado principal é o de Porter (1980), uma vez que a maioria dos algoritmos de redução de radical que surgiram baseiam-se nele. O *stemming* limita a quantidade de termos a serem processados e com isso pode beneficiar tanto as investigações quanto o custo computacional. (VIEIRA; VIRGIL, 2007).

# 4 PROCEDIMENTOS METODOLÓGICOS

Essa seção trata de como se encaminharam as estratégias e os mecanismos utilizados no desenvolvimento para a nova proposta metodológica. Para demonstrar a aplicação das técnicas de dedução de frequência de termos e categorização temática por meio de algoritmos da mineração de texto, utilizou-se um *corpus* textual, aqui delimitado nos resumos das comunicações de determinado evento científico (ENANCIBs), ocorridos entre os anos de 2012 e 2018.

Este estudo apresenta-se como de natureza básica, porque pretende-se contribuir com novos olhares sobre o tema para crescimento do campo científico, pois, embora se aplique a metodologia, tal realização tem caráter demonstrativo. Por ser básica e constituir-se ainda de teorias, quanto aos procedimentos, a pesquisa se pauta em revisões bibliográficas, pois, foi elaborada com base na literatura já publicada. Em relação aos objetivos, pode-se considerar o estudo como exploratório, visto que o experimento metodológico aqui proposto, serve para o levantamento terminológico e poderá ser explorado em diferentes coleções documentais.

Em relação à forma de abordagem do problema, o estudo tem característica quali-quantitativa, em razão do ponto de vista dos mecanismos para sua resolução. O estudo também se vale de técnicas da mineração de texto, que utilizam fórmulas estatísticas e probabilísticas para tratamento, manipulação e extração dos elementos de interpretação e apresentação dos resultados.

Serão utilizados algoritmos para descoberta de conhecimento em caráter qualitativo (estimativas booleanas) e quantitativo (frequência do termo em relação ao texto e ao *corpus*), no entanto, baseado apenas em algoritmos estatísticos. Portanto, como não faz parte do propósito desta pesquisa apontamentos semântico, não se discorreu sobre Processamento de Linguagem Natural (PNL), extração de padrões com algoritmos de aprendizado de máquina (supervisionado ou não supervisionado).

## 4.1 Etapa 1: escolha do *corpus* textual para materialização do estudo

A Associação Nacional de Pesquisa e Pós-graduação em Ciência da Informação (ANCIB) é uma sociedade civil, sem fins lucrativos, fundada em junho de 1989, oriunda de esforços de alguns Cursos e Programas de Pós-graduação da área no país. Desde o início, a associação admite sócios institucionais (os Programas de Pós-graduação em Ciência da Informação) e sócios individuais que são os

professores, pesquisadores, estudantes de pós-graduação e profissionais egressos dos programas.

O foco da associação é fazer o acompanhamento e incentivar atividades de ensino de pós-graduação e de pesquisa na área de CI no país. Tem atuação relevante no meio acadêmico e vem se destacando tanto nacional como internacionalmente pela representatividade científica e política, debatendo assuntos imprescindíveis para a área de ciência da informação. Suas atividades são constituídas por duas frentes: os Programas de Pós-graduação *stricto sensu*, representados por seus coordenadores, e os Encontros Nacionais de Pesquisa em Ciência da Informação (ENANCIBs), organizados em Grupos de Trabalho (GTs), o qual é considerado um dos principais fóruns de debates e reflexões, reunindo pesquisadores interessados nas temáticas especializadas da CI. (ANCIB, 2022, n.p.).

A Etapa 1 no processo de descoberta de conhecimento corresponde à seleção dos dados e, entendendo quão expressivo tem sido a representação da ANCIB para a área, definiu-se que o conjunto de dados a ser explorado neste estudo deveria originar-se dos encontros de pesquisas impulsionados pela entidade, com os trabalhos provenientes e desenvolvidos por pesquisadores atuantes nos programas de mestrado e doutorado sob sua tutela, portanto, comunicações científicas bastante expressivas dentro da área.

Por reconhecer que a literatura produzida nos encontros promovidos pela ANCIB é extremamente significativa, serão utilizados os resumos das comunicações científicas do GT7 – Produção e Comunicação da Informação em Ciência, Tecnologia & Inovação, publicados entre os anos de 2012 e 2018, nos ENANCIBs. A coleta dos dados foi feita a partir dos anais dos eventos disponíveis nos respectivos *sites*, totalizando 287 publicações, incluindo apresentações orais e pôsteres, sem distinção da língua em que foram escritos. Os quantitativos das comunicações científicas estão demonstrados no Quadro 1, apresentados por ano e local de realização dos encontros.

Quadro 1 – Trabalhos do GT-7 para mineração de texto

| Edição | Ano | Local | Quant. Trabalhos |
|---|---|---|---|
| XIII ENANCIB | 2012 | Rio de Janeiro – RJ | 30 |
| XIV ENANCIB | 2013 | Florianópolis – SC | 37 |
| XV ENANCIB | 2014 | Belo Horizonte – MG | 42 |
| XVI ENANCIB | 2015 | João Pessoa – PB | 29 |

| | | | |
|---|---|---|---|
| XVII ENANCIB | 2016 | Salvador – BA | 48 |
| XVIII ENANCIB | 2017 | Marília – SP | 56 |
| XIX ENANCIB | 2018 | Londrina – PR | 45 |
| | | Total | 287 |

Fonte: Elaborado pelo autor.

Ainda que essa quantidade de documentos (287) seja considerada pequena para processos de mineração de texto, pois, as definições direcionam para grandes quantidades (BERRY; LINOFF, 1997; ARANHA; PASSOS, 2006; HAN; KAMBER, 2006; LOBAINA; SUÁREZ, 2018; SALAZAR-LÓPES *et al.*, 2020), entende-se que minerar dados para levantar-se os principais termos ou mapear as correlações dos documentos, mesmo nessa quantidade, se fosse feita de maneira manual, seria uma atividade difícil e cansativa.

A mineração será realizada nos resumos dos trabalhos, isso porque, embora a metodologia aqui proposta viabilize o levantamento em qualquer componente do texto - título, autor, palavras-chave - as técnicas contemporâneas de recuperação automática de documentos e análise textual, entendem tais componentes como primários e preferem trabalhar com textos, mesmo que pequenos. Outrossim, informações extraídas dos resumos servem mais às pessoas do que aos computadores e, por entender que expressam um retrato mais fiel dos documentos, adequando-se a estudos mais aprofundados. (WITTEN, 2004).

A presente pesquisa está baseada em uma coleção documental estática, visto que a quantidade inicial de documentos não será alterada. Em coleções dinâmicas, pode ocorrer a inclusão de documentos novos ou atualizados ao longo do tempo (FELDMAN; SANGER, 2007), o que poderia causar distorções nos resultados. Para aplicação das técnicas de mineração de texto, aproveitou-se para delimitar um intervalo temporal que não simplesmente determinasse as datas de início e fim, mas que também pudesse colaborar com outros projetos.

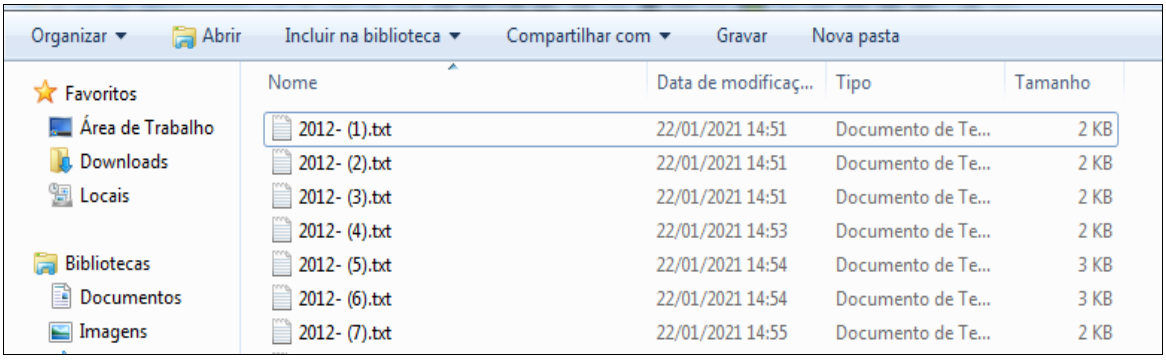
Do mesmo modo, teve-se a intenção de oportunizar novas investigações de estudos analíticos de domínio, pois, tais investigações sugerem como adequada a realização de no mínimo duas perspectivas em simultâneo para se alcançar definições mais apropriadas de especialidades de área. (HJØRLAND, 2002; URBIZAGASTEGUI-ALVARADO, 2021). Por exemplo, como tem-se conhecimento de estudo bibliométrico já realizado por Nogueira, Reis e Oliveira (2019), o qual apresenta análise no mesmo *corpus* e período, entende-se que a junção das duas pesquisas

poderia criar possibilidades de novos olhares sobre o prisma analítico de domínio, já que se teriam os estudos e resultados de duas abordagens (bibliométrica e terminológica).

## 4.2 Etapas 2 e 3: pré-processamento e transformação dos dados

Todos os arquivos do *corpus* utilizado foram baixados no formato de arquivo "pdf' (*Portable Document File*), comum para a maioria dos usuários de computadores e internet, por ser um formato padrão de armazenamento de arquivos nas plataformas digitais. Os arquivos coletados foram transformados do formato inicial para o formato "txt". Isso se faz necessário porque os arquivos em padrão txt são aceitos na maioria dos *software*, além de estarem limpos de formatações e marcações que outras aplicações possam deixar. Nessa etapa, utilizou-se como ferramenta o *software AntFileConverter*[40], que tem capacidade de realizar a conversão (pdf para txt) em um conjunto de arquivos em um mesmo processo. A Figura 2 mostra como os arquivos foram salvos, convertidos e armazenados dentro do computador.

Figura 2 – Arquivos transformados de pdf para txt




Fonte: Captura de tela da pesquisa (2021).

Após o procedimento de conversão, cada arquivo foi salvo nas respectivas pastas criadas e separadas por ano (2012, 2013, 2014, 2015, 2016, 2017 e 2018), para que se tivesse um controle mais preciso na manipulação dos dados. Em seguida, executou-se a limpeza das partes dos textos que não seriam utilizadas e ficaram armazenados nos arquivos apenas o conteúdo dos resumos de cada documento, organizados em pastas específicas, identificados por ano e sequência em que foram convertidos.

[40] É uma ferramenta de uso livre para converter arquivos PDF e Word (DOCX) em texto simples (txt). Pode ser baixado no site: https://www.laurenceanthony.net/software/antfileconverter/

O *software* Knime será a ferramenta utilizada para as próximas fases do KD, especificamente para a realização da mineração textual (pré-processamento; transformação; manipulação; mineração com uso de algoritmos; resultados). O *software* funciona com nós ou módulos interligados, que executam tarefas específicas a cada módulo. O primeiro passo para se iniciar um fluxo é carregar no *software* a base de dados ou *corpus* (textos, sites, redes sociais, arquivos etc.) que contenham os materiais a serem trabalhados, no caso desta pesquisa, os resumos das comunicações científicas, já armazenados em arquivos no formato txt.

E assim se dará a continuidade desta pesquisa, visto que, independentemente do tipo de arquivo (doc, docx, PDF, xls, txt etc.), os processos se iniciam com o carregamento do acervo para dentro do *software* Knime. O ponto de partida, então, é escolher um nó que execute a leitura dos arquivos específicos e como se optou pela manipulação de arquivos de textos puro (txt), se utilizou para dar início ao fluxograma o módulo *Flat File Document Parser*, que carrega este tipo específico de arquivo.

Na sequência do fluxograma, pode-se executar a transformação de todas as palavras para minúsculo, o que pode ser feito com o nó *Case Converter.* Em seguida, com o módulo *Stop Word Filter*, é feito a remoção das palavras consideradas comuns no vocabulário português (Brasil) e sem relevância em processos de busca de informação (as, e, os, de, para, com, sem, foi etc.). Ainda nessa sequência de limpeza, serão feitos ajustes nos textos como a retirada dos diacríticos, das pontuações e a remoção de números, utilizando-se para isso respectivamente os módulos *Diacritic Remover*, *Punctuaction Erasure* e *Number Filter*.

Alguns módulos podem ser posicionados independentemente do módulo anterior, caso os resultados dos dados processados no antecessor não sejam utilizados especificamente no módulo seguinte. Por exemplo, na preparação dos dados, utilizar um módulo que converte todas as letras para minúsculo/maiúsculo ou um nó que retira palavras consideradas irrelevantes, independem da sequência de posicionamento no fluxograma, porque não faz diferença para a continuidade do processo deixar as palavras minúsculo/maiúsculo antes ou depois de se excluírem as palavras. São módulos importantes para a limpeza dos dados, que fazem parte do fluxograma, mas podem funcionar de maneira independente um do outro.

## 4.3 Etapa 4: mineração e busca de padrões

Depois de cumpridas as fases 1, 2 e 3, na fase 4, com as técnicas da mineração de texto, se inicia a busca por padrões ainda não conhecidos nos textos. Após a limpeza e preparação dos dados, eles passarão a ser manipulados para obtenção e apresentação dos resultados. Assim, o próximo módulo utilizado é o *Sentence Extractor*, que separa o texto de cada resumo em sentenças e faz a contagem das palavras que compõem as frases. O nó *Column Filter* é um nó somente de manipulação e serve para armazenar apenas a coluna com os dados do módulo anterior a serem aproveitados posteriormente.

O mesmo ocorre com o módulo *Numeric Row Spliter,* que tem como finalidade fazer o filtro da quantidade de termos que se pretende minerar, pois sua configuração permite colocar um intervalo de, por exemplo, 5 a 10 palavras. Desse modo, as sentenças com menos de 5 e mais de 10 termos não serão computadas. Igualmente, há a possibilidade de definição do intervalo entre 0 e 500, o que faria com que não fosse descartada nenhuma sentença, a não ser que nos textos existam frases com mais de 500 palavras.

Para complementar os módulos anteriores, utiliza-se o nó *Color Manager,* que prepara a saída de dados para a próxima tarefa, que seria a apresentação dos dados em modo gráfico, executado pelo nó *Interactive Histogram*. Esse encadeamento dos nós apresentados até aqui serve para apontar algumas características textuais dos discursos, que corresponde a um dos objetivos do estudo.

Em seguida, um outro objetivo específico tem como mote fazer o levantamento da frequência de uso e dimensionar a coocorrências dos principais termos presentes no *corpus*. Para isso, aproveita-se a limpeza e transformação dos dados (fases 1, 2 e 3), interligando os módulos que realizaram essas tarefas com o módulo *Bag of Words*, que separa cada palavra (unigrama – pós-graduação) que compõe as sentenças em uma célula distinta dentro do *software*. O conjunto de módulos a seguir serve para preparar a apresentação dos resultados posteriormente, de forma inicial como nuvem de palavras, mas que poderá ser aproveitada em outras demonstrações gráficas.

Segue o percurso sequencial dos módulos utilizados para realização desse objetivo: *Value Couter* (faz a contagem das coocorrências dos termos); *Rank* (classifica e ordena os termos conforme configurado, da maior ocorrência para a menor); *Row Filter* (define a quantidade de termos que se deseja filtrar); *RowID*

(transforma os identificadores dos termos em dados possíveis de serem tratados); *String to Term* (prepara o dado para ser trabalhado pelo nó seguinte); *Domain Calculator* (permite escolher a quantidade de termos a serem apresentados nos gráficos); *Color Manager* (configura-se os termos com cores diferentes); e *Tag Cloud*, (apresenta uma nuvem de palavras, diferenciando o tamanho das letras - maiores e menores - conforme ranqueamento).

Para buscar o que se propõe na sequência, em que se procura localizar os n-gramas (*cluster* de palavras) mais utilizados nos discursos pelos autores, utiliza-se um algoritmo denominado n-gram ou n-gramas. Esse algoritmo possibilita encontrar *clusters* de palavras e quantidade de ocorrências dos *clusters* no *corpus* textual, partindo-se de bigramas (pós-graduação ciência), trigramas (pós-graduação ciência informação) e assim sucessivamente. Utilizaremos neste estudo bigramas e trigramas, por entender que com esse conjunto de palavras já é possível encontrar padrões nos textos (o módulo pode ser configurado conforme preferência do pesquisador, por exemplo com cluster de 2, 3, 4, 5 palavras, ou mais).

Compreende-se que, até aqui, as etapas 1, 2, e 3 devem ter sido entendidas como basilares para a aplicação dos algoritmos, por serem processos que fazem a preparação e manipulação para a execução da mineração de texto. Após as etapas iniciais, os módulos seguintes para obtenção dos *clusters* serão os seguintes: *Ngram Creator* - módulo com algoritmo que executa todos os procedimentos de agrupamento e verificação de frequência dos *clusters* presentes no *corpus*, nos documentos individualmente e nas sentenças de cada documento; e o nó *Tag Cloud* - que demonstra em modo gráfico os *clusters* de palavras que mais se reptem nos discursos*.* Esta pesquisa, utilizará os resultados de frequência obtidos no *corpus* por representar as ocorrências dos *clusters* de forma mais abrangente e menos individual, em concordância com apontamentos das abordagens de análise de domínio.

O próximo objetivo tem como finalidade indicar os conjuntos de termos que representam os principais assuntos abordados nos textos estudados. A técnica a ser utilizada para realização desse item é a de modelagem de tópicos com aplicação do algoritmo LDA, o qual utiliza como base a acepção de que os documentos são formados por um conjunto de tópicos e que cada tópico é um conjunto de palavras significativas obtidas por meio de cálculos probabilísticos.

Esse algoritmo também aproveita os dados já processados até a etapa 3 para realização dos cálculos, para fazer o agrupamento dos tópicos, e ainda para classificar

os documentos por *clusters* (com base nos tópicos extraídos), inferindo o significado dos textos. Essa última função de classificação viabiliza também a verificação das possíveis correlações temáticas e interlocuções dos discursos presentes nos resumos, e servirá ao objetivo de organizar os artigos em *cluster* por assuntos detectados pelo algoritmo.

Por último, pretende-se realizar o *stemming* (derivação ou radicalização) para agrupar os termos com mesmo radical. Como essa técnica possibilita a melhora em resultados de busca e recuperação da informação, ela pode contribuir nos processos de indexação e desenvolvimento de sistemas de organização do conhecimento. Nesta pesquisa, utilizar-se-á o módulo de *stemming*, disponível na ferramenta Knime denominado *Snowball*, que faz uso de uma biblioteca de mesmo nome[41], pois, embora existam outros algoritmos no *software*, ele é o único que faz o processo de radicalização na língua portuguesa.

Quadro 2 – Encadeamento esperado do fluxograma

| Módulos das etapas de tratamento e transformação de dados – (1, 2 e 3) | | | |
|---|---|---|---|
| Flat File Document Parser – Case Converter – Stop Word Filter – Diacritic Remover – Punctuaction Erasure – Number Filter | | | |
| Módulos da etapa de mineração – (4) | | | |
| Objetivo (1) | Objetivo (2) | Objetivo (3) | Objetivo (4) |
| Sentence Extractor – Column Filter - Numeric Row Spliter - Color Manager – Interactive Histogram | Bag of Words – Value Couter – Rank – Row Filter – RowID – String to Term – Domain Calculator – Color Manager – Tag Cloud – Ngram Creator – Tag Cloud | LDA – Value Counter – Column Filter – Sorter – Objetct Inserter – Network Viewer | Snowball Stemmer – Bag of Words – Value Couter – Rank – Row Filter – RowID – String to Term – Domain Calculator – Color Manager – Tag Cloud |

Fonte: Elaborado pelo autor.

No Quadro 2, demonstra-se o esboço de representação de como os processos de mineração de texto serão feitos para se alcançar os objetivos desta pesquisa, e de como, teoricamente, os módulos serão posicionados para elaboração do fluxograma. Não se descarta a utilização de mais alguns módulos no decorrer do estudo para

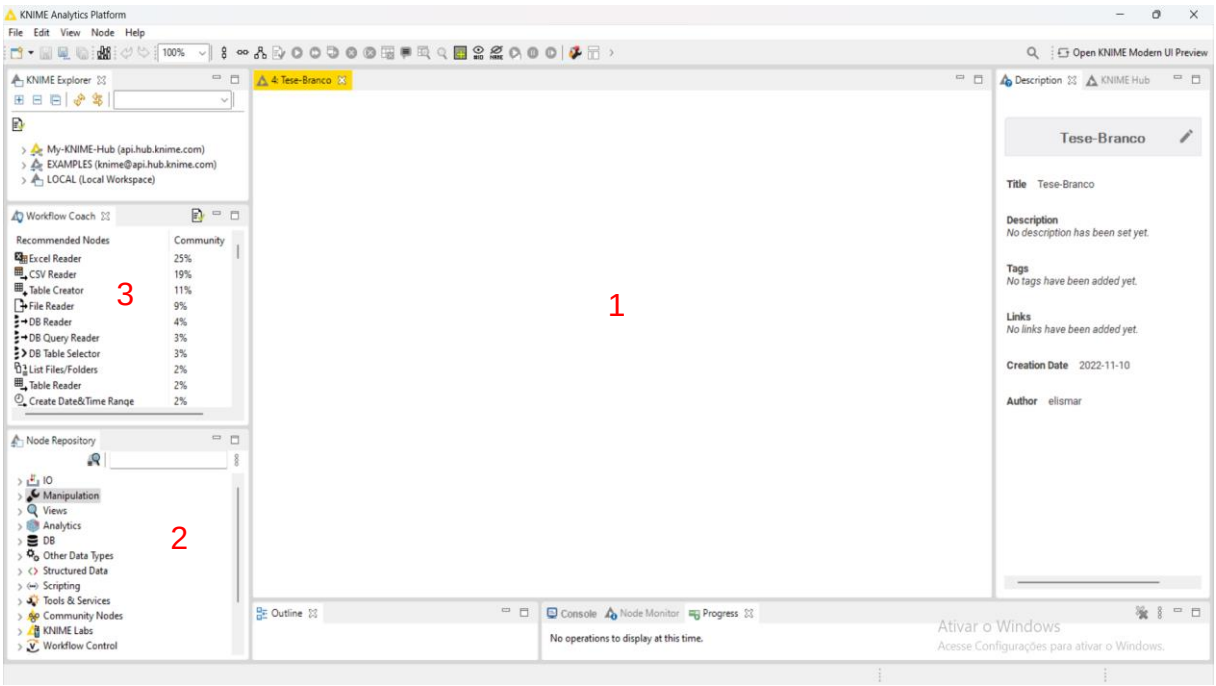
[41] Detalhes sobre a biblioteca de derivação do Snowball estão em: http://snowball.tartarus.org/.

transformação e adequação dos dados para alguma saída para outro nó, conforme as necessidades forem aparecendo, entretanto, esses módulos serão utilizados apenas para preparar a saída de dados para melhor demonstração gráfica, como, por exemplo, diferenciação nos resultados (cores ou tamanhos) dos elementos a serem apresentados. Esse encadeamento de módulos apresentado no quadro, serviu como esboço dos parâmetros necessários para elaboração de um fluxograma canônico, com capacidade para realizar tratamento, manipulação e mineração de texto para levantamento temático em *corpus* textuais.

# 5 APLICAÇÃO DA METODOLOGIA E RESULTADOS

Com base no encaminhamento metodológico parametrizado no Quadro 2 e com o *software* Knime em execução, iniciaram-se os procedimentos de criação do fluxograma para mineração de texto. Na Figura 3, apresenta-se a tela da ferramenta (Knime), utilizada nesta pesquisa para consecução da metodologia delineada. Não se tem pretensão de explorar o *software* além do que esteja no escopo desta tese, nem de se aprofundar nas explicações sobre ele, por isso, as explanações aqui postas servirão apenas aos objetivos específicos deste estudo. Outrossim, a metodologia pode ser implantada em qualquer *software,* desde que realize as funções para o levantamento terminológico proposto na pesquisa.

Figura 3 – Tela inicial do *software* Knime




Fonte: Captura de tela da pesquisa (2021).

O *software* foi instalado no sistema operacional Windows, mas ele pode ser utilizado também no Linux ou no MacOS; e sua usabilidade é bastante semelhante à de outros *software* presente no cotidiano de usuários de computador, portanto, entende-se que não há necessidade de apresentações específicas em relação ao comportamento da ferramenta. Na Figura 3, a parte ilustrada com o número 1 (um) é a área de trabalho, onde os módulos são arrastados para que sejam feitas as interligações pertinentes e eles executem suas funções específicas.

Na parte em que se visualiza o número 2 (dois – *Node Repository*) é o local em que estão os nós propriamente ditos, os quais devem ser selecionados e arrastados para a área de trabalho. A seção 3 (três – *Workflow Coach*) baseia-se em estatísticas de utilização de sequências de módulos, captadas em nível mundial e serve para recomendar os nós que podem ser utilizados à frente de um módulo qualquer que esteja selecionado na área de trabalho. As estatísticas de indicação se dão pelo percentual de uso de um módulo específico à frente do outro, baseado nas sequências recorrentes verificadas na comunidade de usuários do *software*.

## 5.1 Criação de fluxograma para descoberta de conhecimento em texto

Para a elaboração do fluxograma, conforme o Quadro 2, a fase de pré-processamento se iniciaria com o nó *Flat File Document Parser,* entretanto, há um detalhe importante desse módulo que merece ser observado em caso de utilização futura. O módulo funciona perfeitamente e faz o carregamento dos arquivos em texto puro (txt) para o Knime de maneira correta, mas, a cada vez que se executa essa ação no módulo, os arquivos são ordenados em sequências diferentes. Como cada sequência no carregamento dos artigos seria diferente, para que os resultados fossem apresentados de modo correto, seria necessário carregar os arquivos e realizar todos os outros processos de uma só vez, pois, em uma nova ação de inserção dos arquivos a sequência mudaria, comprometendo a repetição dos processos quando necessário, e até mesmo para possível comprovação e replicação da metodologia.

Isso não seria problema para a realização da mineração, pois, se o processo todo for feito, os resultados serão extraídos, já que a sequência de carregamento pouco importa. Contudo, quando há necessidade de se fazer a “captura das telas” para visualização posterior, ou a realização de novos testes, em uma eventual repetição de rotinas dos módulos, os textos seriam carregados de maneira diferente, alterando a apresentação dos resultados a cada procedimento.

Por exemplo, as capturas de telas da sequência/carregamento 1 (um) seriam diferentes da sequência/carregamento 2 (dois) e, assim, consecutivamente. Embora em qualquer momento pudesse acontecer de o carregamento apresentar a mesma sequência do carregamento 1 (um), isso dependeria de muitas variáveis, o que causaria transtornos e insegurança quanto aos resultados, pois, o próprio ato de se

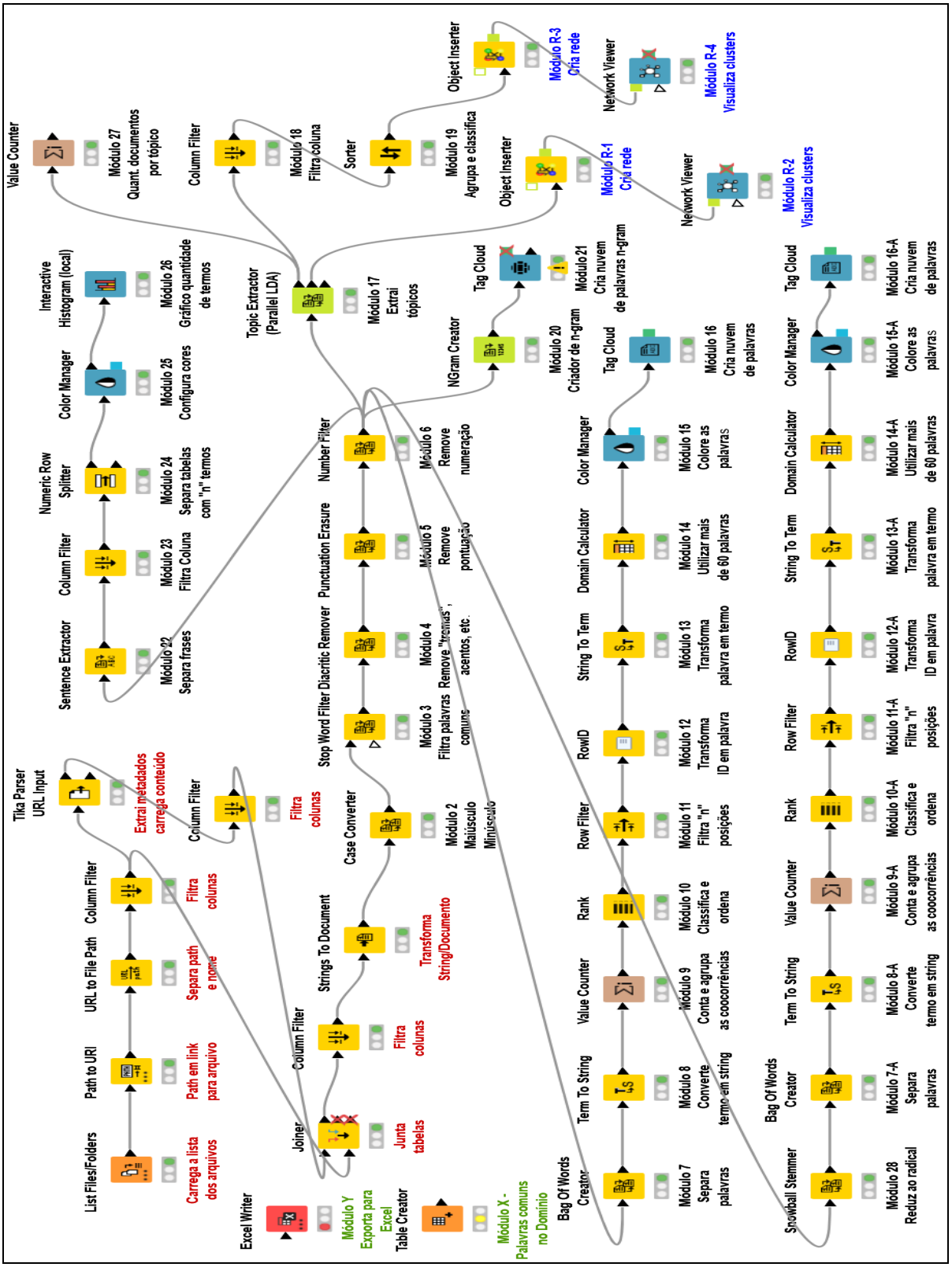
verificar a igualdade dos resumos em duas sequências seria um processo manual e bastante custoso.

Por se tratar de uma proposta metodológica, essa situação torna-se inapropriada, porque, ao se fazer uso de um itinerário para replicação de sistematização e consecução de tarefas, espera-se que o recurso auxilie a pesquisa e que cause o mínimo de problemas. Outrossim, faz-se necessário observar a que público a metodologia poderá servir, para que seja demandado o mínimo de conhecimento e esforço técnico em relação aos *software* utilizados.

Outra questão que causa insatisfação no módulo *Flat File Document Parser* é a dificuldade em se executar a manipulação dos dados e obter o nome e o conteúdo dos arquivos ao mesmo tempo. O nó em questão faz o carregamento do conteúdo de forma adequada, entretanto, as linhas no Knime que identificam os documentos, ou mostram o *path* (caminho do local em que o arquivo está armazenado no computador), ou mostram o conteúdo dos resumos, podendo atrapalhar o desenvolvimento dos estudos nos itens que necessitam a visualização desses dois elementos ao mesmo tempo.

Novamente, argumente-se que, em alguns casos da mineração de texto, esse cenário não causaria obstáculo, mas em situações em que fosse preciso analisar o conteúdo dos documentos e depois demonstrar as eventuais interligações entre eles, a apresentação dos resultados ficaria prejudicada, pois, não seria possível apresentar o nome dos artigos e suas interligações. Portando, fez-se necessário uma solução para obtenção do nome/identificação e o seu conteúdo correspondente, e ainda, que os arquivos fossem carregados para o *software* sempre na mesma ordem. A Figura 4 apresenta o fluxograma já com os módulos iniciais que resolvem o problema detectado, permitindo a manipulação dos dados de forma mais apropriada. Os nós que substituíram o *Flat File Document Parser* estão descrito no fluxograma com letras na cor vermelha.

Figura 4 – Fluxograma de descoberta de conhecimento em texto

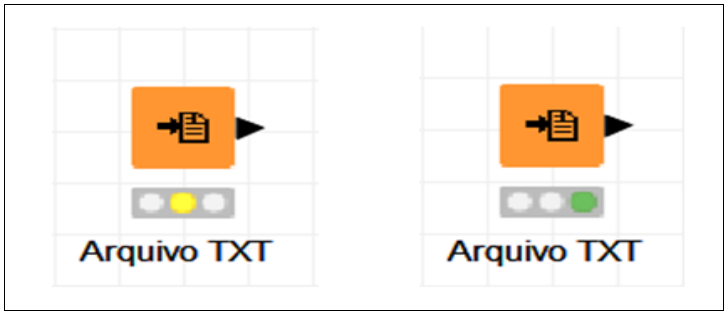



Fonte: Elaborado pelo autor.[42]

[42] Um tutorial de instalação do Knime e este fluxograma podem ser baixados no *link* - https://drive.google.com/drive/folders/1tzbR5X9-I4xJ0159fr77HUP1_OBQB26x?usp=drive_link

Quanto ao funcionamento da ferramenta Knime, ao se realizar o comando de execução das ações nos módulos e não for encontrado nenhum erro de configuração, então, cada módulo realiza a tarefa que lhe condiz, prepara os dados para serem executados pelo nó seguinte e alteram seu *status* de amarelo para verde (mudança que ocorre em todos os módulos). A mudança de *status* indica que as ações foram executadas com sucesso, possibilitando visualizar nos próprios módulos o resultado de cada etapa. A Figura 5 apenas demonstra como os módulos se comportam em relação ao seu estado inicial e após a execução de uma tarefa realizada com sucesso.

Figura 5 – Exemplo de nó do Knime com mudança de status

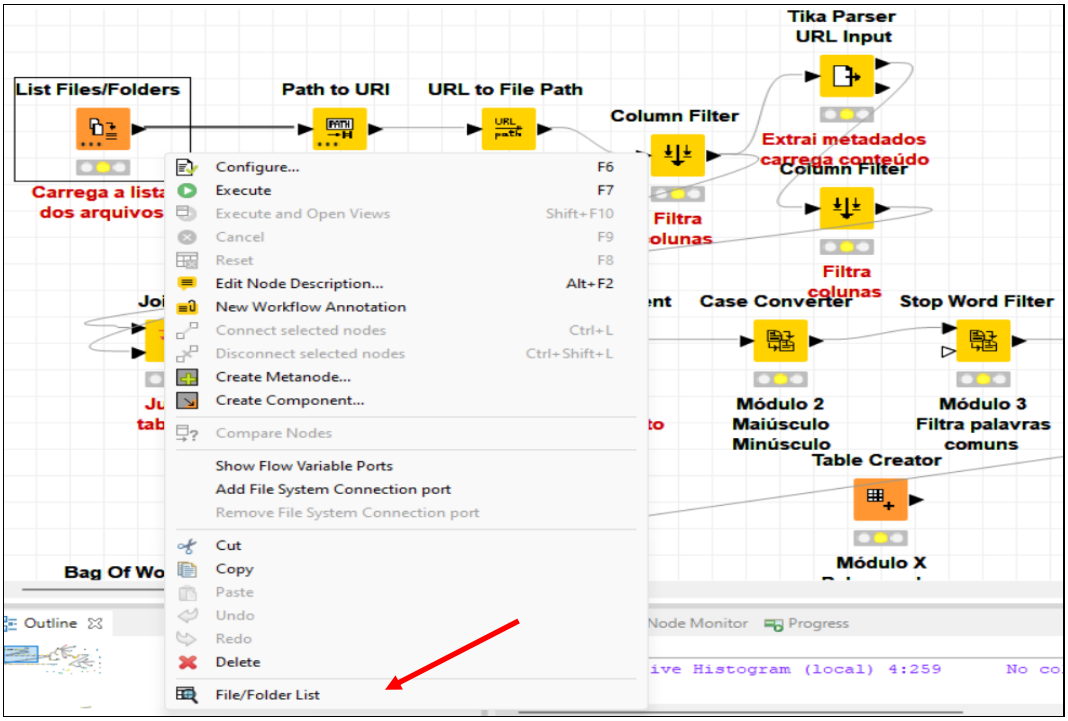



Fonte: Captura de tela da pesquisa (2021).

Antes de o usuário efetuar o comando para a execução da ação propriamente, a maioria dos módulos precisam ser configurados, visto que esse procedimento direciona o módulo para a realização correta da tarefa. O comportamento do *software* Knime é idêntico à maioria de outros *software* aplicativos, o que também pode facilitar seu uso. Ao efetuar um clique com o botão direito do *mouse* sobre o módulo que se pretende configurar, todos eles apresentarão uma lista de opões, muito parecidas às que podem ser observadas na Figura 6.

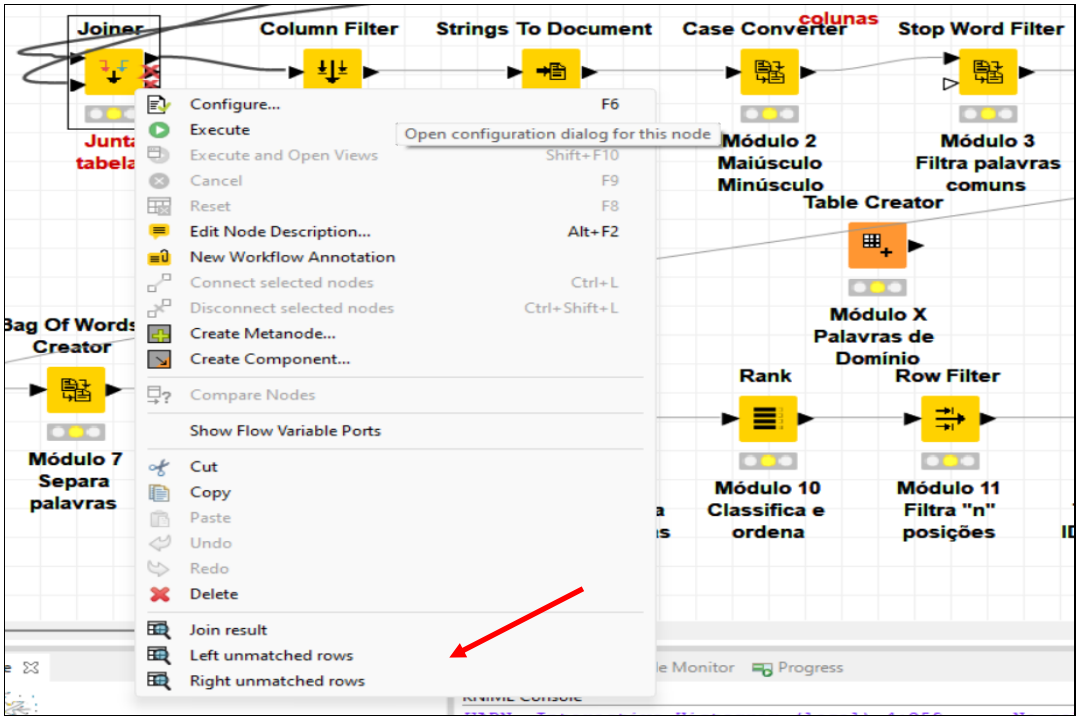
Em particular, na referida figura destaque-se: *Configure, Execute, Reset* e *File/Folder List*. A opção *Configure* serve para acessar as possibilidades e preferências de configuração de cada módulo em específico, ao passo que o *Execute* realiza a ação conforme configurada e muda o status do módulo para a cor verde. *Reset* faz a ação ao contrário do *Execute*, caso seja necessário, pois limpa os dados, desfaz a ação executada e volta o *status* do módulo para amarelo. O *Reset* é utilizado em casos de mudança de configuração do módulo por exemplo, porque, ao fazer as alterações nos comandos, é necessário refazer a ação de execução, por isso é preciso zerar os dados carregados com a configuração anterior e executá-la novamente. Assim os dados podem ser trabalhados e carregados com as novas definições de configuração.

Figura 6 – Menu de acesso aos comandos dos módulos




Fonte: Captura de tela da pesquisa (2022).

Figura 7 – Menu com mais opções de apresentação de dados



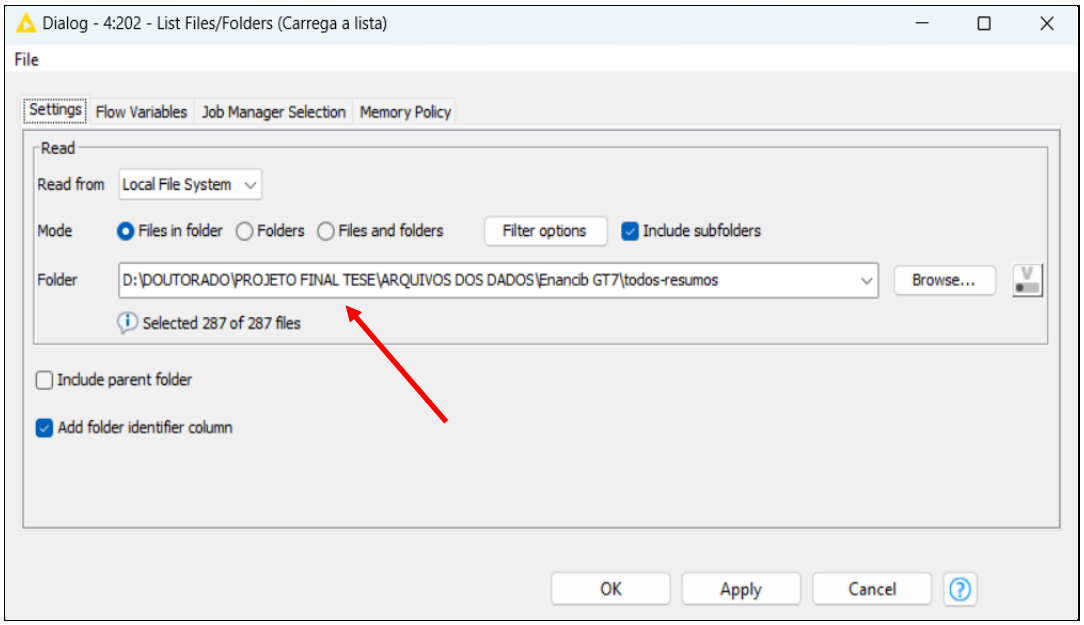

Fonte: Captura de tela da pesquisa (2022).

Especial atenção deve ser dispensada para a(s) última(s) seção(ções) dos menus, pois, nessa seção estão as opções de apresentação dos dados, de acordo com cada módulo utilizado, os quais podem ser configurados conforme necessidade

ou preferência dos usuários. Como alguns nós possuem mais de uma opção de saída de dados, eles podem apresentar mais possibilidades de apresentação, como mostrado na Figura 7. Cada módulo possui sua especificidade, tanto para as configurações de entrada de dados quanto para a apresentação deles.

Por exemplo, na Figura 8, demonstra-se as opções de configuração do *Módulo List Files/Folders*, o qual serve para carregar para o Knime os arquivos (*corpus* textual) que serão utilizados na mineração de textos. É possível perceber na figura que há opções de definição do local em que os arquivos estão armazenados (*Local File System* – computador local) e serão lidos, do modo como eles estão gravados, de acordo com a preferência do pesquisador (*Files in folder – Folder – Files and folders*) e ainda, a possibilidade de apontamento da pasta no computador local, que pode ser selecionada utilizando o botão *Browse*.

Figura 8 – Tela de configuração do módulo *List Files/Folders*



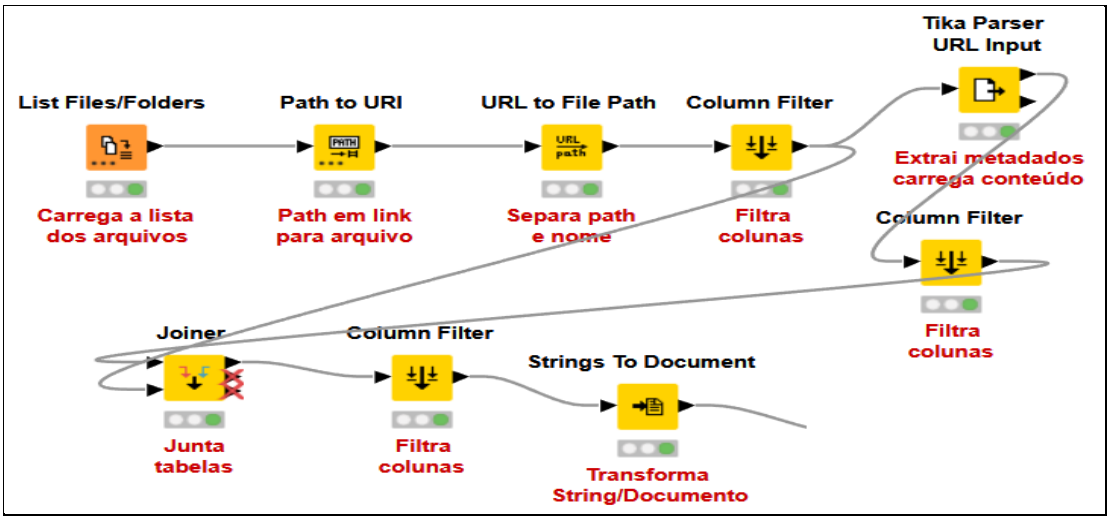

Fonte: Captura de tela da pesquisa (2022).

Tais métodos de configuração devem ser de conhecimento dos usuários contemporâneos de computador, dado que são ações recorrentes em vários aplicativos utilitários de dispositivos digitais. Saliente-se novamente que, cada módulo possui suas configurações específicas, por isso, ao acessar a opção de ajustes em outro módulo qualquer, telas diferentes da Figura 8 serão apresentadas e terão as indicações de configuração exclusivas do módulo selecionado.

## 5.2 Etapa 1: detalhamento do novo módulo 1

Após a exposição básica de utilização de algumas funções de configuração dos módulos no Knime, na Figura 9, expõem-se os módulos que compõem o que se convencionou chamar nesta pesquisa de Novo Módulo 1, devido ao problema encontrado e já explicado com o módulo *Flat File Document Parser* (antigo Módulo 1). O antigo Módulo 1 foi substituído pelos seguintes nós: *List Fil/Folders*, *Path to URL*, *URL To file Path*, *Column Filter*, *Tika Parser URL Input*, *Column Filter*, *Joiner*, *Column Filter* e *Strings to Document*.

Figura 9 – Módulos que compõem o novo módulo 1

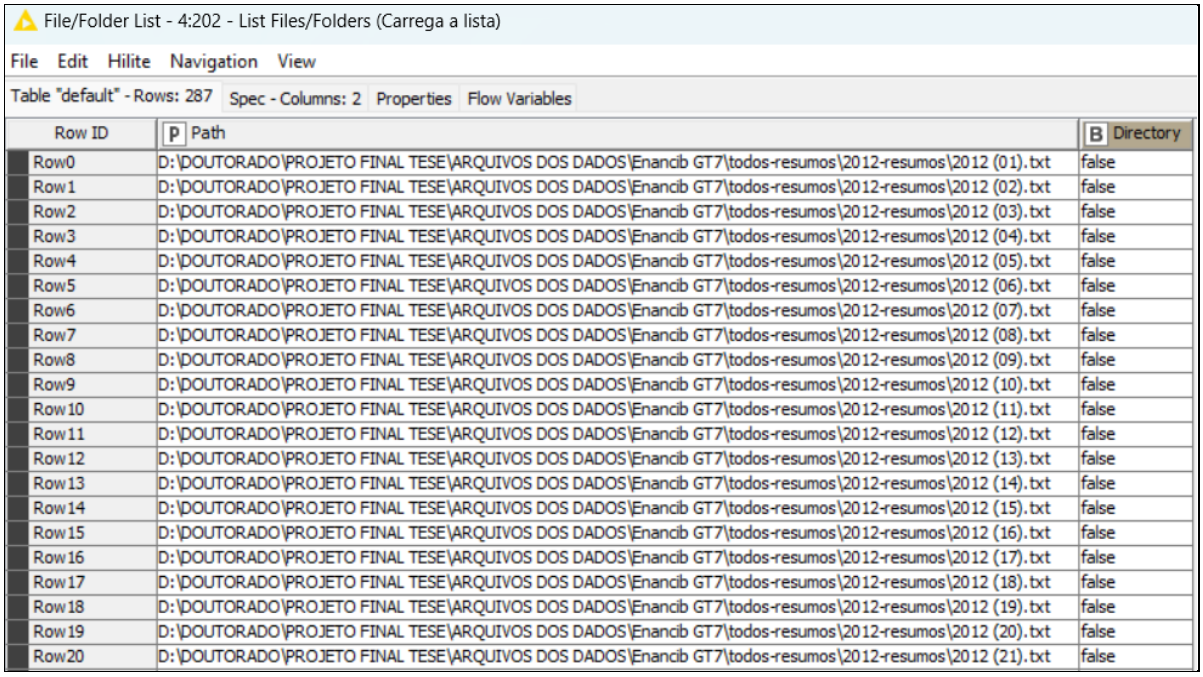



Fonte: Captura de tela da pesquisa (2022).

A seguir, apresenta-se uma sequência de figuras que demonstrarão o modo de armazenamento e saída dos dados desses módulos. As figuras normalmente foram capturadas depois de já terem sido executadas as ações (opção *Execute*) nos módulos e ao escolher a linha/*link*/botão que gera a apresentação dos dados (ver Figura 6). A Figura 10 apresenta o resultado do módulo *List File/Folders*, os quais mostram o caminho de cada arquivo com o resumo das comunicações científicas, e ainda, se esse caminho corresponde a uma pasta (*Directory*) ou não.

O módulo *List Files/Folders* apenas mostra ao *software* a localização do arquivo. Já o módulo *Path to URL* cria um *link* para o referido arquivo no computador local, resultado que é demonstrado na Figura 11. Em algumas aplicações, apenas a localização do arquivo seria suficiente, entretanto, para esta pesquisa, o conteúdo do arquivo também será utilizado, por isso, há a necessidade de criação do *link* para conseguir acessá-lo e, consequentemente, obtenção do seu conteúdo.

Figura 10 – Saída de dados do módulo *List File/Folders*



File/Folder List - 4:202 - List Files/Folders (Carrega a lista)

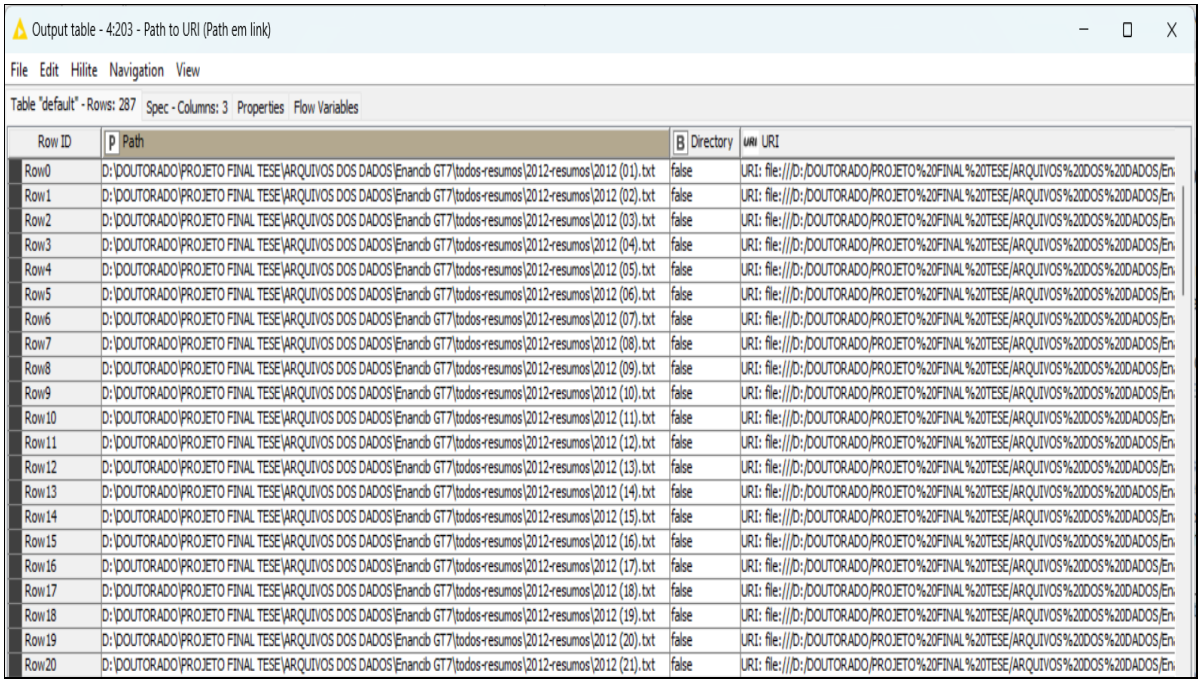
File Edit Hilite Navigation View

Table "default" - Rows: 287 Spec - Columns: 2 Properties Flow Variables

| Row ID | Path | Directory |
|---|---|---|
| Row0 | D:\DOUTORADO\PROJETO FINAL TESE\ARQUIVOS DOS DADOS\Enancib GT7\todos-resumos\2012-resumos\2012 (01).txt | false |
| Row1 | D:\DOUTORADO\PROJETO FINAL TESE\ARQUIVOS DOS DADOS\Enancib GT7\todos-resumos\2012-resumos\2012 (02).txt | false |
| Row2 | D:\DOUTORADO\PROJETO FINAL TESE\ARQUIVOS DOS DADOS\Enancib GT7\todos-resumos\2012-resumos\2012 (03).txt | false |
| Row3 | D:\DOUTORADO\PROJETO FINAL TESE\ARQUIVOS DOS DADOS\Enancib GT7\todos-resumos\2012-resumos\2012 (04).txt | false |
| Row4 | D:\DOUTORADO\PROJETO FINAL TESE\ARQUIVOS DOS DADOS\Enancib GT7\todos-resumos\2012-resumos\2012 (05).txt | false |
| Row5 | D:\DOUTORADO\PROJETO FINAL TESE\ARQUIVOS DOS DADOS\Enancib GT7\todos-resumos\2012-resumos\2012 (06).txt | false |
| Row6 | D:\DOUTORADO\PROJETO FINAL TESE\ARQUIVOS DOS DADOS\Enancib GT7\todos-resumos\2012-resumos\2012 (07).txt | false |
| Row7 | D:\DOUTORADO\PROJETO FINAL TESE\ARQUIVOS DOS DADOS\Enancib GT7\todos-resumos\2012-resumos\2012 (08).txt | false |
| Row8 | D:\DOUTORADO\PROJETO FINAL TESE\ARQUIVOS DOS DADOS\Enancib GT7\todos-resumos\2012-resumos\2012 (09).txt | false |
| Row9 | D:\DOUTORADO\PROJETO FINAL TESE\ARQUIVOS DOS DADOS\Enancib GT7\todos-resumos\2012-resumos\2012 (10).txt | false |
| Row10 | D:\DOUTORADO\PROJETO FINAL TESE\ARQUIVOS DOS DADOS\Enancib GT7\todos-resumos\2012-resumos\2012 (11).txt | false |
| Row11 | D:\DOUTORADO\PROJETO FINAL TESE\ARQUIVOS DOS DADOS\Enancib GT7\todos-resumos\2012-resumos\2012 (12).txt | false |
| Row12 | D:\DOUTORADO\PROJETO FINAL TESE\ARQUIVOS DOS DADOS\Enancib GT7\todos-resumos\2012-resumos\2012 (13).txt | false |
| Row13 | D:\DOUTORADO\PROJETO FINAL TESE\ARQUIVOS DOS DADOS\Enancib GT7\todos-resumos\2012-resumos\2012 (14).txt | false |
| Row14 | D:\DOUTORADO\PROJETO FINAL TESE\ARQUIVOS DOS DADOS\Enancib GT7\todos-resumos\2012-resumos\2012 (15).txt | false |
| Row15 | D:\DOUTORADO\PROJETO FINAL TESE\ARQUIVOS DOS DADOS\Enancib GT7\todos-resumos\2012-resumos\2012 (16).txt | false |
| Row16 | D:\DOUTORADO\PROJETO FINAL TESE\ARQUIVOS DOS DADOS\Enancib GT7\todos-resumos\2012-resumos\2012 (17).txt | false |
| Row17 | D:\DOUTORADO\PROJETO FINAL TESE\ARQUIVOS DOS DADOS\Enancib GT7\todos-resumos\2012-resumos\2012 (18).txt | false |
| Row18 | D:\DOUTORADO\PROJETO FINAL TESE\ARQUIVOS DOS DADOS\Enancib GT7\todos-resumos\2012-resumos\2012 (19).txt | false |
| Row19 | D:\DOUTORADO\PROJETO FINAL TESE\ARQUIVOS DOS DADOS\Enancib GT7\todos-resumos\2012-resumos\2012 (20).txt | false |
| Row20 | D:\DOUTORADO\PROJETO FINAL TESE\ARQUIVOS DOS DADOS\Enancib GT7\todos-resumos\2012-resumos\2012 (21).txt | false |

Fonte: Captura de tela da pesquisa (2022).

Figura 11 – Saída de dados do módulo *Path to URL*



Output table - 4:203 - Path to URI (Path em link)

File Edit Hilite Navigation View

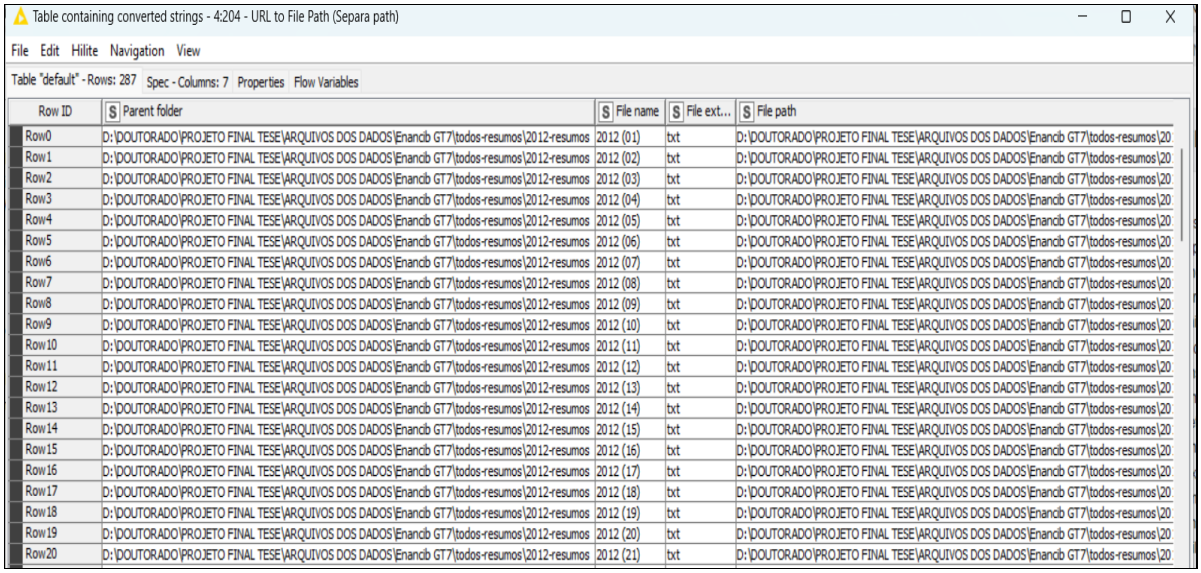
Table "default" - Rows: 287 Spec - Columns: 3 Properties Flow Variables

| Row ID | Path | Directory | URI |
|---|---|---|---|
| Row0 | D:\DOUTORADO\PROJETO FINAL TESE\ARQUIVOS DOS DADOS\Enancib GT7\todos-resumos\2012-resumos\2012 (01).txt | false | URI: file:///D:/DOUTORADO/PROJETO%20FINAL%20TESE/ARQUIVOS%20DOS%20DADOS/En |
| Row1 | D:\DOUTORADO\PROJETO FINAL TESE\ARQUIVOS DOS DADOS\Enancib GT7\todos-resumos\2012-resumos\2012 (02).txt | false | URI: file:///D:/DOUTORADO/PROJETO%20FINAL%20TESE/ARQUIVOS%20DOS%20DADOS/En |
| Row2 | D:\DOUTORADO\PROJETO FINAL TESE\ARQUIVOS DOS DADOS\Enancib GT7\todos-resumos\2012-resumos\2012 (03).txt | false | URI: file:///D:/DOUTORADO/PROJETO%20FINAL%20TESE/ARQUIVOS%20DOS%20DADOS/En |
| Row3 | D:\DOUTORADO\PROJETO FINAL TESE\ARQUIVOS DOS DADOS\Enancib GT7\todos-resumos\2012-resumos\2012 (04).txt | false | URI: file:///D:/DOUTORADO/PROJETO%20FINAL%20TESE/ARQUIVOS%20DOS%20DADOS/En |
| Row4 | D:\DOUTORADO\PROJETO FINAL TESE\ARQUIVOS DOS DADOS\Enancib GT7\todos-resumos\2012-resumos\2012 (05).txt | false | URI: file:///D:/DOUTORADO/PROJETO%20FINAL%20TESE/ARQUIVOS%20DOS%20DADOS/En |
| Row5 | D:\DOUTORADO\PROJETO FINAL TESE\ARQUIVOS DOS DADOS\Enancib GT7\todos-resumos\2012-resumos\2012 (06).txt | false | URI: file:///D:/DOUTORADO/PROJETO%20FINAL%20TESE/ARQUIVOS%20DOS%20DADOS/En |
| Row6 | D:\DOUTORADO\PROJETO FINAL TESE\ARQUIVOS DOS DADOS\Enancib GT7\todos-resumos\2012-resumos\2012 (07).txt | false | URI: file:///D:/DOUTORADO/PROJETO%20FINAL%20TESE/ARQUIVOS%20DOS%20DADOS/En |
| Row7 | D:\DOUTORADO\PROJETO FINAL TESE\ARQUIVOS DOS DADOS\Enancib GT7\todos-resumos\2012-resumos\2012 (08).txt | false | URI: file:///D:/DOUTORADO/PROJETO%20FINAL%20TESE/ARQUIVOS%20DOS%20DADOS/En |
| Row8 | D:\DOUTORADO\PROJETO FINAL TESE\ARQUIVOS DOS DADOS\Enancib GT7\todos-resumos\2012-resumos\2012 (09).txt | false | URI: file:///D:/DOUTORADO/PROJETO%20FINAL%20TESE/ARQUIVOS%20DOS%20DADOS/En |
| Row9 | D:\DOUTORADO\PROJETO FINAL TESE\ARQUIVOS DOS DADOS\Enancib GT7\todos-resumos\2012-resumos\2012 (10).txt | false | URI: file:///D:/DOUTORADO/PROJETO%20FINAL%20TESE/ARQUIVOS%20DOS%20DADOS/En |
| Row10 | D:\DOUTORADO\PROJETO FINAL TESE\ARQUIVOS DOS DADOS\Enancib GT7\todos-resumos\2012-resumos\2012 (11).txt | false | URI: file:///D:/DOUTORADO/PROJETO%20FINAL%20TESE/ARQUIVOS%20DOS%20DADOS/En |
| Row11 | D:\DOUTORADO\PROJETO FINAL TESE\ARQUIVOS DOS DADOS\Enancib GT7\todos-resumos\2012-resumos\2012 (12).txt | false | URI: file:///D:/DOUTORADO/PROJETO%20FINAL%20TESE/ARQUIVOS%20DOS%20DADOS/En |
| Row12 | D:\DOUTORADO\PROJETO FINAL TESE\ARQUIVOS DOS DADOS\Enancib GT7\todos-resumos\2012-resumos\2012 (13).txt | false | URI: file:///D:/DOUTORADO/PROJETO%20FINAL%20TESE/ARQUIVOS%20DOS%20DADOS/En |
| Row13 | D:\DOUTORADO\PROJETO FINAL TESE\ARQUIVOS DOS DADOS\Enancib GT7\todos-resumos\2012-resumos\2012 (14).txt | false | URI: file:///D:/DOUTORADO/PROJETO%20FINAL%20TESE/ARQUIVOS%20DOS%20DADOS/En |
| Row14 | D:\DOUTORADO\PROJETO FINAL TESE\ARQUIVOS DOS DADOS\Enancib GT7\todos-resumos\2012-resumos\2012 (15).txt | false | URI: file:///D:/DOUTORADO/PROJETO%20FINAL%20TESE/ARQUIVOS%20DOS%20DADOS/En |
| Row15 | D:\DOUTORADO\PROJETO FINAL TESE\ARQUIVOS DOS DADOS\Enancib GT7\todos-resumos\2012-resumos\2012 (16).txt | false | URI: file:///D:/DOUTORADO/PROJETO%20FINAL%20TESE/ARQUIVOS%20DOS%20DADOS/En |
| Row16 | D:\DOUTORADO\PROJETO FINAL TESE\ARQUIVOS DOS DADOS\Enancib GT7\todos-resumos\2012-resumos\2012 (17).txt | false | URI: file:///D:/DOUTORADO/PROJETO%20FINAL%20TESE/ARQUIVOS%20DOS%20DADOS/En |
| Row17 | D:\DOUTORADO\PROJETO FINAL TESE\ARQUIVOS DOS DADOS\Enancib GT7\todos-resumos\2012-resumos\2012 (18).txt | false | URI: file:///D:/DOUTORADO/PROJETO%20FINAL%20TESE/ARQUIVOS%20DOS%20DADOS/En |
| Row18 | D:\DOUTORADO\PROJETO FINAL TESE\ARQUIVOS DOS DADOS\Enancib GT7\todos-resumos\2012-resumos\2012 (19).txt | false | URI: file:///D:/DOUTORADO/PROJETO%20FINAL%20TESE/ARQUIVOS%20DOS%20DADOS/En |
| Row19 | D:\DOUTORADO\PROJETO FINAL TESE\ARQUIVOS DOS DADOS\Enancib GT7\todos-resumos\2012-resumos\2012 (20).txt | false | URI: file:///D:/DOUTORADO/PROJETO%20FINAL%20TESE/ARQUIVOS%20DOS%20DADOS/En |
| Row20 | D:\DOUTORADO\PROJETO FINAL TESE\ARQUIVOS DOS DADOS\Enancib GT7\todos-resumos\2012-resumos\2012 (21).txt | false | URI: file:///D:/DOUTORADO/PROJETO%20FINAL%20TESE/ARQUIVOS%20DOS%20DADOS/En |

Fonte: Captura de tela da pesquisa (2022).

A Figura 12 apresenta a saída de dados do *Módulo URL to Path*, que separa o nome do arquivo para identificação (coluna *File Name*), o tipo de arquivo que está sendo carregado (coluna *File extension*) e determina o caminho que o *link* dos arquivos criados no módulo anterior deve apontar (*File path*), transformando-a ainda em uma *string,* para que possa ser manipulada pelo *software* de mineração.

Figura 12 – Saída de dados do módulo *URL to Path*

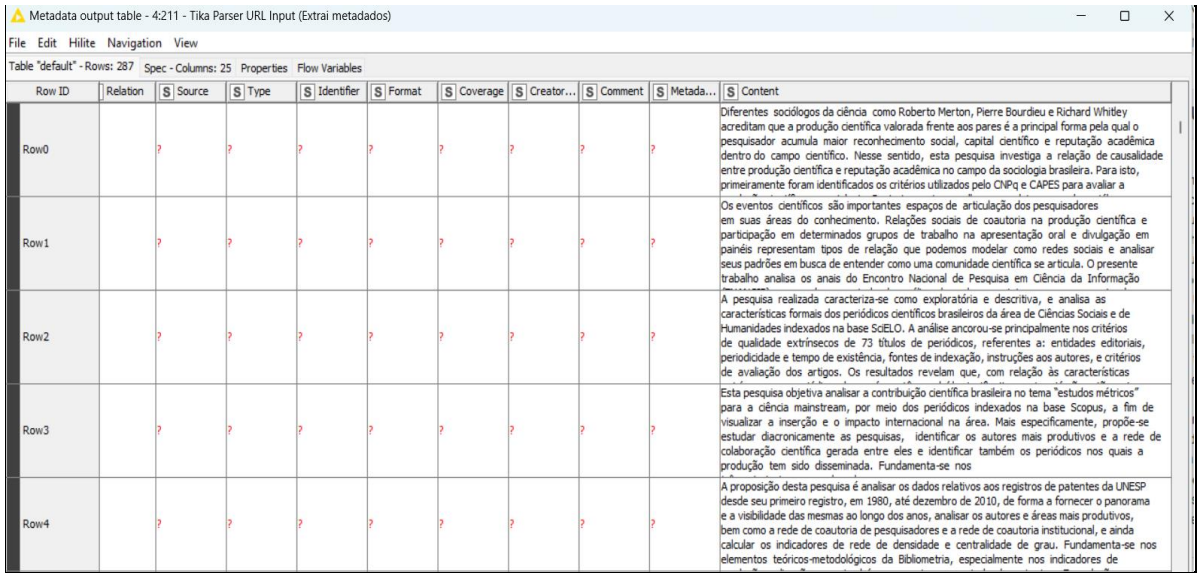



| Row ID | Parent folder | File name | File ext... | File path |
|---|---|---|---|---|
| Row0 | D:\DOUTORADO\PROJETO FINAL TESE\ARQUIVOS DOS DADOS\Enancib GT7\todos-resumos\2012-resumos | 2012 (01) | txt | D:\DOUTORADO\PROJETO FINAL TESE\ARQUIVOS DOS DADOS\Enancib GT7\todos-resumos\20 |
| Row1 | D:\DOUTORADO\PROJETO FINAL TESE\ARQUIVOS DOS DADOS\Enancib GT7\todos-resumos\2012-resumos | 2012 (02) | txt | D:\DOUTORADO\PROJETO FINAL TESE\ARQUIVOS DOS DADOS\Enancib GT7\todos-resumos\20 |
| Row2 | D:\DOUTORADO\PROJETO FINAL TESE\ARQUIVOS DOS DADOS\Enancib GT7\todos-resumos\2012-resumos | 2012 (03) | txt | D:\DOUTORADO\PROJETO FINAL TESE\ARQUIVOS DOS DADOS\Enancib GT7\todos-resumos\20 |
| Row3 | D:\DOUTORADO\PROJETO FINAL TESE\ARQUIVOS DOS DADOS\Enancib GT7\todos-resumos\2012-resumos | 2012 (04) | txt | D:\DOUTORADO\PROJETO FINAL TESE\ARQUIVOS DOS DADOS\Enancib GT7\todos-resumos\20 |
| Row4 | D:\DOUTORADO\PROJETO FINAL TESE\ARQUIVOS DOS DADOS\Enancib GT7\todos-resumos\2012-resumos | 2012 (05) | txt | D:\DOUTORADO\PROJETO FINAL TESE\ARQUIVOS DOS DADOS\Enancib GT7\todos-resumos\20 |
| Row5 | D:\DOUTORADO\PROJETO FINAL TESE\ARQUIVOS DOS DADOS\Enancib GT7\todos-resumos\2012-resumos | 2012 (06) | txt | D:\DOUTORADO\PROJETO FINAL TESE\ARQUIVOS DOS DADOS\Enancib GT7\todos-resumos\20 |
| Row6 | D:\DOUTORADO\PROJETO FINAL TESE\ARQUIVOS DOS DADOS\Enancib GT7\todos-resumos\2012-resumos | 2012 (07) | txt | D:\DOUTORADO\PROJETO FINAL TESE\ARQUIVOS DOS DADOS\Enancib GT7\todos-resumos\20 |
| Row7 | D:\DOUTORADO\PROJETO FINAL TESE\ARQUIVOS DOS DADOS\Enancib GT7\todos-resumos\2012-resumos | 2012 (08) | txt | D:\DOUTORADO\PROJETO FINAL TESE\ARQUIVOS DOS DADOS\Enancib GT7\todos-resumos\20 |
| Row8 | D:\DOUTORADO\PROJETO FINAL TESE\ARQUIVOS DOS DADOS\Enancib GT7\todos-resumos\2012-resumos | 2012 (09) | txt | D:\DOUTORADO\PROJETO FINAL TESE\ARQUIVOS DOS DADOS\Enancib GT7\todos-resumos\20 |
| Row9 | D:\DOUTORADO\PROJETO FINAL TESE\ARQUIVOS DOS DADOS\Enancib GT7\todos-resumos\2012-resumos | 2012 (10) | txt | D:\DOUTORADO\PROJETO FINAL TESE\ARQUIVOS DOS DADOS\Enancib GT7\todos-resumos\20 |
| Row10 | D:\DOUTORADO\PROJETO FINAL TESE\ARQUIVOS DOS DADOS\Enancib GT7\todos-resumos\2012-resumos | 2012 (11) | txt | D:\DOUTORADO\PROJETO FINAL TESE\ARQUIVOS DOS DADOS\Enancib GT7\todos-resumos\20 |
| Row11 | D:\DOUTORADO\PROJETO FINAL TESE\ARQUIVOS DOS DADOS\Enancib GT7\todos-resumos\2012-resumos | 2012 (12) | txt | D:\DOUTORADO\PROJETO FINAL TESE\ARQUIVOS DOS DADOS\Enancib GT7\todos-resumos\20 |
| Row12 | D:\DOUTORADO\PROJETO FINAL TESE\ARQUIVOS DOS DADOS\Enancib GT7\todos-resumos\2012-resumos | 2012 (13) | txt | D:\DOUTORADO\PROJETO FINAL TESE\ARQUIVOS DOS DADOS\Enancib GT7\todos-resumos\20 |
| Row13 | D:\DOUTORADO\PROJETO FINAL TESE\ARQUIVOS DOS DADOS\Enancib GT7\todos-resumos\2012-resumos | 2012 (14) | txt | D:\DOUTORADO\PROJETO FINAL TESE\ARQUIVOS DOS DADOS\Enancib GT7\todos-resumos\20 |
| Row14 | D:\DOUTORADO\PROJETO FINAL TESE\ARQUIVOS DOS DADOS\Enancib GT7\todos-resumos\2012-resumos | 2012 (15) | txt | D:\DOUTORADO\PROJETO FINAL TESE\ARQUIVOS DOS DADOS\Enancib GT7\todos-resumos\20 |
| Row15 | D:\DOUTORADO\PROJETO FINAL TESE\ARQUIVOS DOS DADOS\Enancib GT7\todos-resumos\2012-resumos | 2012 (16) | txt | D:\DOUTORADO\PROJETO FINAL TESE\ARQUIVOS DOS DADOS\Enancib GT7\todos-resumos\20 |
| Row16 | D:\DOUTORADO\PROJETO FINAL TESE\ARQUIVOS DOS DADOS\Enancib GT7\todos-resumos\2012-resumos | 2012 (17) | txt | D:\DOUTORADO\PROJETO FINAL TESE\ARQUIVOS DOS DADOS\Enancib GT7\todos-resumos\20 |
| Row17 | D:\DOUTORADO\PROJETO FINAL TESE\ARQUIVOS DOS DADOS\Enancib GT7\todos-resumos\2012-resumos | 2012 (18) | txt | D:\DOUTORADO\PROJETO FINAL TESE\ARQUIVOS DOS DADOS\Enancib GT7\todos-resumos\20 |
| Row18 | D:\DOUTORADO\PROJETO FINAL TESE\ARQUIVOS DOS DADOS\Enancib GT7\todos-resumos\2012-resumos | 2012 (19) | txt | D:\DOUTORADO\PROJETO FINAL TESE\ARQUIVOS DOS DADOS\Enancib GT7\todos-resumos\20 |
| Row19 | D:\DOUTORADO\PROJETO FINAL TESE\ARQUIVOS DOS DADOS\Enancib GT7\todos-resumos\2012-resumos | 2012 (20) | txt | D:\DOUTORADO\PROJETO FINAL TESE\ARQUIVOS DOS DADOS\Enancib GT7\todos-resumos\20 |
| Row20 | D:\DOUTORADO\PROJETO FINAL TESE\ARQUIVOS DOS DADOS\Enancib GT7\todos-resumos\2012-resumos | 2012 (21) | txt | D:\DOUTORADO\PROJETO FINAL TESE\ARQUIVOS DOS DADOS\Enancib GT7\todos-resumos\20 |

Fonte: Captura de tela da pesquisa (2022).

O módulo *Column Filter* (mostrado na Figura 9) serve para filtrar somente as colunas que o pesquisador deseja utilizar. Se voltarmos à Figura 12, por exemplo, podemos utilizar esse módulo para filtrar apenas as colunas *File name* e *File extension*, de modo que apenas essas duas ficarão armazenadas no módulo *Column Filter*. Portanto, entende-se que não há necessidade de apresentação dos dados desse módulo. Já o módulo seguinte (pode ser conferido na Figura 9), *Tika Parser URL Input*, tem papel bastante importante. Caso os arquivos contivessem metadados delimitados, ele conseguiria extrair essas informações e ainda carregar o conteúdo do arquivo, de acordo com o que é apresentado na Figura 13.

Figura 13 – Saída de dados do módulo *Tika Parser URL Input*



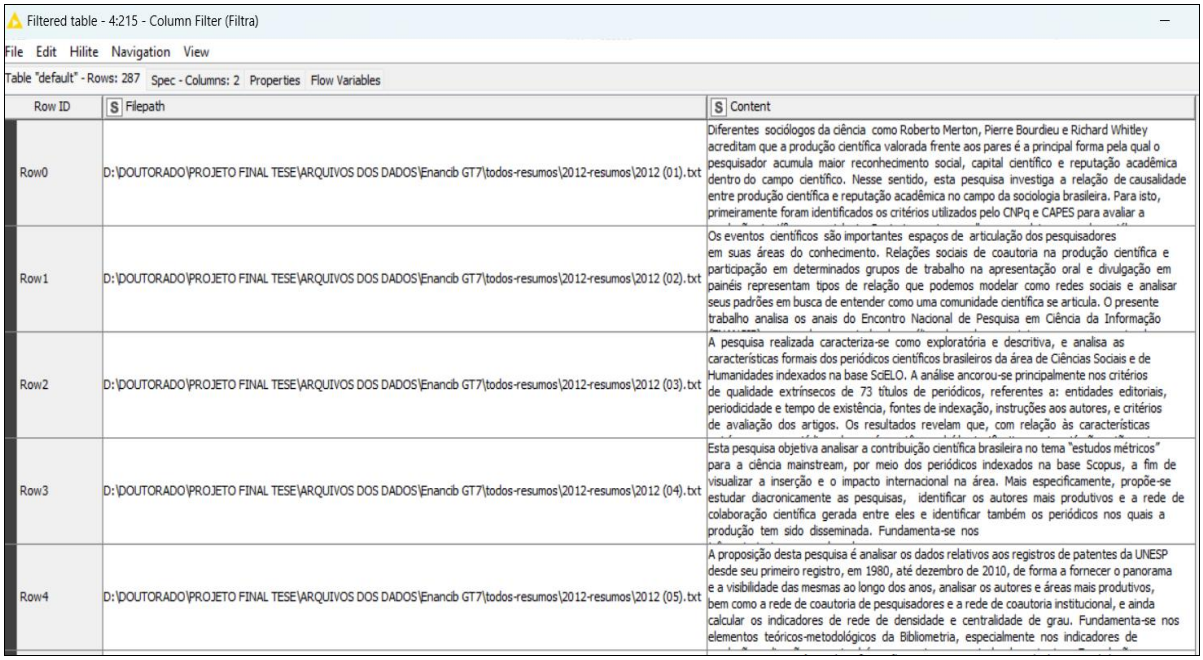

| Row ID | Relation | Source | Type | Identifier | Format | Coverage | Creator... | Comment | Metada... | Content |
|---|---|---|---|---|---|---|---|---|---|---|
| Row0 | | ? | ? | ? | ? | ? | ? | ? | ? | Diferentes sociólogos da ciência como Roberto Merton, Pierre Bourdieu e Richard Whitley acreditam que a produção científica valorada frente aos pares é a principal forma pela qual o pesquisador acumula maior reconhecimento social, capital científico e reputação acadêmica dentro do campo científico. Nesse sentido, esta pesquisa investiga a relação de causalidade entre produção científica e reputação acadêmica no campo da sociologia brasileira. Para isto, primeiramente foram identificados os critérios utilizados pelo CNPq e CAPES para avaliar a |
| Row1 | | ? | ? | ? | ? | ? | ? | ? | ? | Os eventos científicos são importantes espaços de articulação dos pesquisadores em suas áreas do conhecimento. Relações sociais de coautoria na produção científica e participação em determinados grupos de trabalho na apresentação oral e divulgação em painéis representam tipos de relação que podemos modelar como redes sociais e analisar seus padrões em busca de entender como uma comunidade científica se articula. O presente trabalho analisa os anais do Encontro Nacional de Pesquisa em Ciência da Informação |
| Row2 | | ? | ? | ? | ? | ? | ? | ? | ? | A pesquisa realizada caracteriza-se como exploratória e descritiva, e analisa as características formais dos periódicos científicos brasileiros da área de Ciências Sociais e de Humanidades indexados na base SciELO. A análise ancorou-se principalmente nos critérios de qualidade extrínsecos de 73 títulos de periódicos, referentes a: entidades editoriais, periodicidade e tempo de existência, fontes de indexação, instruções aos autores, e critérios de avaliação dos artigos. Os resultados revelam que, com relação às características |
| Row3 | | ? | ? | ? | ? | ? | ? | ? | ? | Esta pesquisa objetiva analisar a contribuição científica brasileira no tema "estudos métricos" para a ciência mainstream, por meio dos periódicos indexados na base Scopus, a fim de visualizar a inserção e o impacto internacional na área. Mais especificamente, propõe-se estudar diacronicamente as pesquisas, identificar os autores mais produtivos e a rede de colaboração científica gerada entre eles e identificar também os periódicos nos quais a produção tem sido disseminada. Fundamenta-se nos |
| Row4 | | ? | ? | ? | ? | ? | ? | ? | ? | A proposição desta pesquisa é analisar os dados relativos aos registros de patentes da UNESP desde seu primeiro registro, em 1980, até dezembro de 2010, de forma a fornecer o panorama e a visibilidade das mesmas ao longo dos anos, analisar os autores e áreas mais produtivos, bem como a rede de coautoria de pesquisadores e a rede de coautoria institucional, e ainda calcular os indicadores de rede de densidade e centralidade de grau. Fundamenta-se nos elementos teóricos-metodológicos da Bibliometria, especialmente nos indicadores de |

Fonte: Captura de tela da pesquisa (2022).

Como o módulo *Tika Parser URL Input* gera uma tabela grande com várias colunas, como título, autor, data de criação do arquivo, entre outros, serão apresentadas aqui apenas algumas colunas, pois não haveria espaço de impressão suficiente. É nesse módulo que os apontamentos dos *links* para os arquivos são utilizados, visto que o Knime utiliza a localização do arquivo por meio do *link* para fazer as leituras necessárias e abastecer o *software* com as informações pertinentes. Para esta pesquisa, o que interessa é a coluna *Content*, mostrado na Figura 13, que tem armazenada os conteúdos dos arquivos. Percebe-se ainda que algumas colunas armazenam um ponto de interrogação (?), o qual indica que o *software* não localizou informações correspondentes, por isso, são células sem utilidade para este estudo.

O próximo filtro (módulo *Column Filter*) é utilizado para diminuir essa quantidade excessiva de colunas vazias, sem dados importantes, e deixar armazenado no módulo apenas as colunas que interessam ao pesquisador. A Figura 14 apresenta o módulo *Column Filter*, que está à frente do módulo *Tika Parser URL Input* (pode ser conferido na Figura 9) e faz a filtragem conforme definida nas configurações. A figura, demonstra a ação de filtro já executada, deixando cada linha com a localização do arquivo e o conteúdo armazenado dentro dele.

Figura 14 – Saída de dados do módulo *Column Filter*



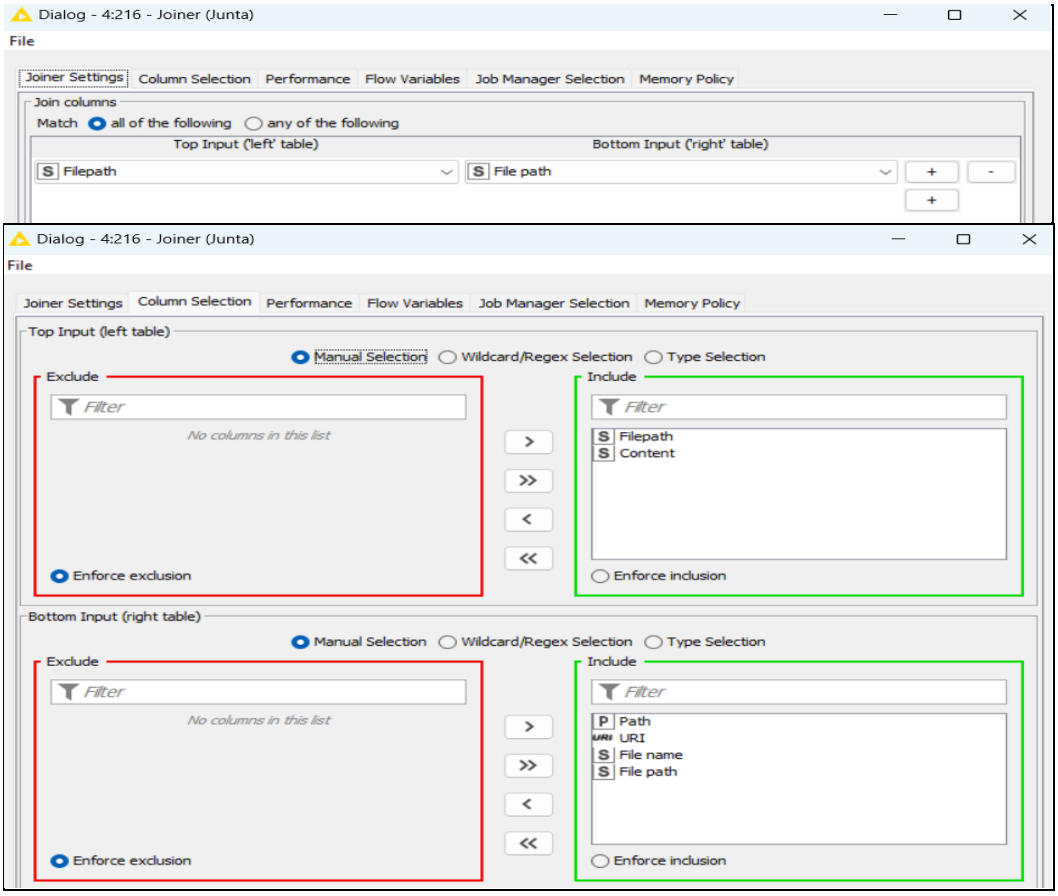

Fonte: Captura de tela da pesquisa (2022).

O módulo *Joiner*, apresentado na Figura 15, faz a junção de colunas de acordo com a configuração do usuário. Nesta pesquisa, tem-se que, para a apresentação dos

dados, seria mais interessante deixar somente o nome do arquivo e seu conteúdo, e não o endereço de localização inteiro, como apresentado na Figura 14 na coluna *Filepath*. Algumas ações são necessárias para se conseguir esse resultado, assim, utiliza-se a saída de dados do módulo *URL to Filepath* (Figura 12) e do último *Column Filter* apresentado (Figura 14), para se conseguir a extração de tais informações (nome e conteúdo do arquivo) e realizar a união que melhor represente o interesse do pesquisador.

Relembre-se que, o primeiro procedimento em um módulo é realizar a configuração dele por meio da opção *Configure*, para que o resultado de saída dos dados seja o desejado. A Figura 15 apresenta parte das opções de configuração do módulo *Joiner*, respectivamente as abas *Joiner Settings* e *Column Selection*. Na primeira aba, faz-se a configuração de quais colunas das duas tabelas devem ser equivalentes para comparação das igualdades e se extrair os dados pertinentes. Na segunda aba, configura-se quais colunas de cada tabela inicial (*Include)* deverão compor a nova tabela (criada após a junção das duas anteriores).

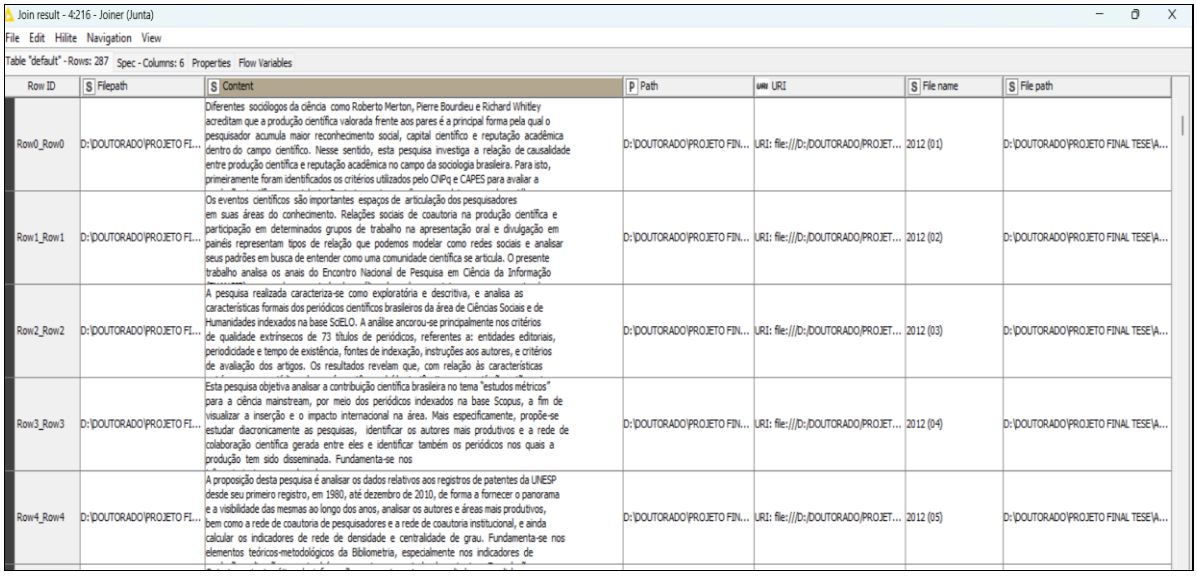
Figura 15 – Abas de configuração do módulo *Joiner*




Fonte: Captura de tela da pesquisa (2022).

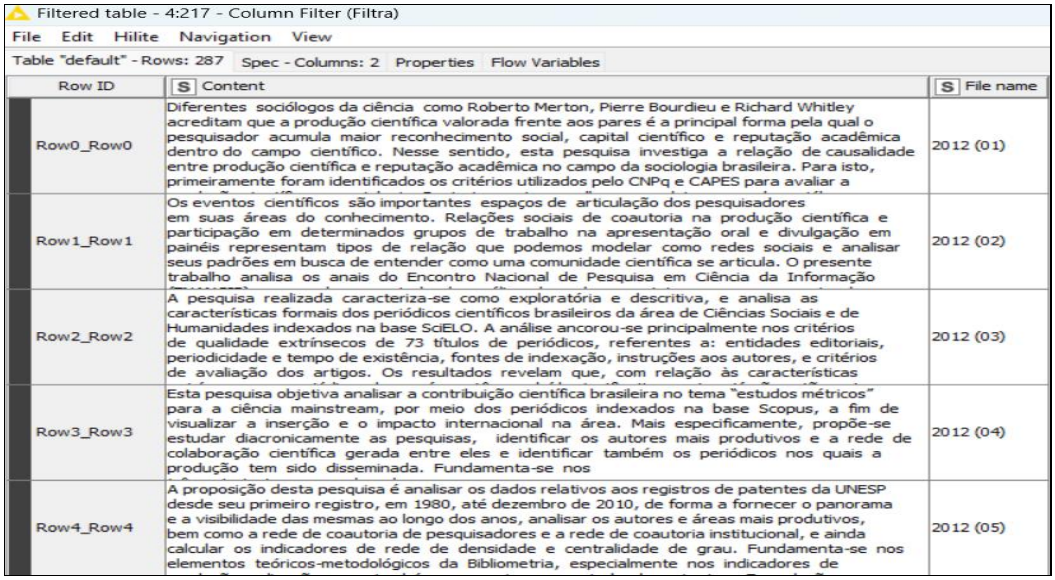
É possível perceber, na Figura 15, quais colunas das tabelas serão carregadas, de modo que essa nova tabela será composta por 6 colunas (destacadas pelos quadrado na cor verde), com as seguintes informações: *Filepath* (1), *Content* (2), *Path* (3), *URL* (4), *File name* (5) e *File path* (6), como demonstrado na Figura 16. Essas informações ainda são necessárias para o correto carregamento e identificação dos conteúdos dos resumos e nomes de arquivos.

Figura 16 – Saída de dados do módulo *Joiner*




Fonte: Captura de tela da pesquisa (2022).

Figura 17 – Saída de dados do módulo *Column Filter*



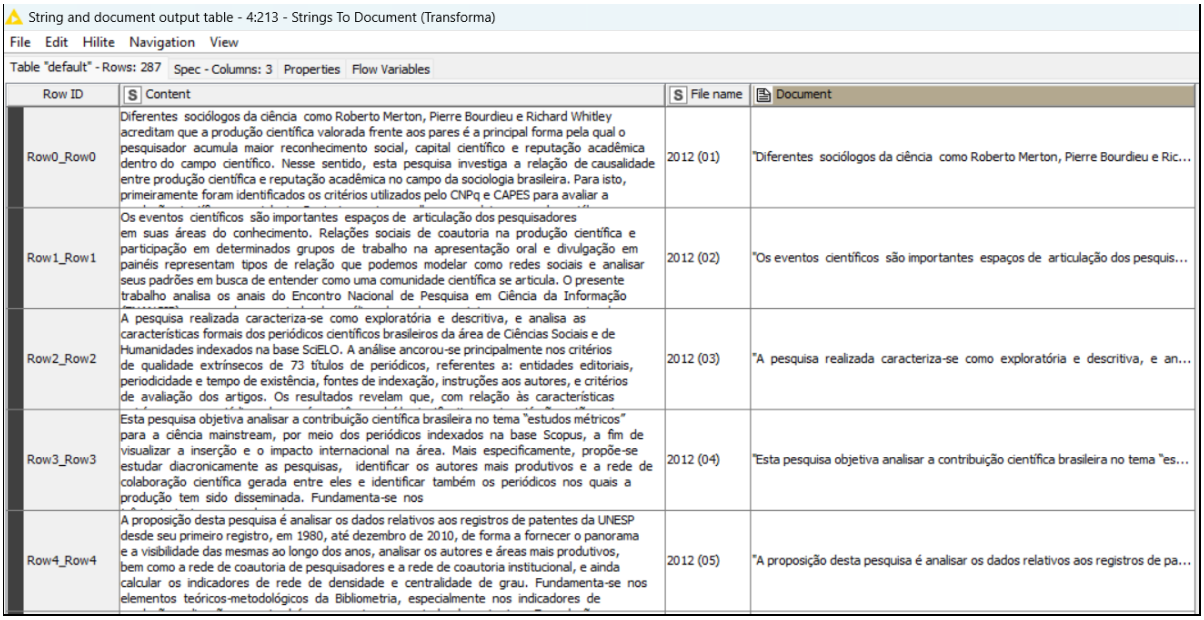

Fonte: Captura de tela da pesquisa (2022).

Com a preparação dos dados quase pronta, faz-se a filtragem das colunas de interesse da pesquisa, de forma que mostrem os nomes que identificam os arquivos

e o conteúdo armazenado em cada um deles, e novamente utiliza-se um módulo *Column Filter* para essa ação, conforme demonstrado na Figura 17.

Do conjunto de módulos que foi considerado nesta pesquisa como Novo Módulo 1, o último a ser utilizado é o *String to Document* (Figura 18), que transforma uma coluna com dados do tipo *string* para o formato *document*, pois, alguns algoritmos de mineração de texto utilizam esse formado de dados. A Figura 18 demonstra o fim da etapa de preparação dos dados e o início das fases de pré-processamento e transformação (Etapas 2 e 3). A figura exibe uma amostra dos 287 documentos manipulados pelo considerado Novo Módulo 1.

Figura 18 – Saída de dados do módulo *String to Document*


String and document output table - 4:213 - Strings To Document (Transforma)

File Edit Hilite Navigation View

Table "default" - Rows: 287 Spec - Columns: 3 Properties Flow Variables

| Row ID | Content | File name | Document |
|---|---|---|---|
| Row0_Row0 | Diferentes sociólogos da ciência como Roberto Merton, Pierre Bourdieu e Richard Whitley acreditam que a produção científica valorada frente aos pares é a principal forma pela qual o pesquisador acumula maior reconhecimento social, capital científico e reputação acadêmica dentro do campo científico. Nesse sentido, esta pesquisa investiga a relação de causalidade entre produção científica e reputação acadêmica no campo da sociologia brasileira. Para isto, primeiramente foram identificados os critérios utilizados pelo CNPq e CAPES para avaliar a | 2012 (01) | "Diferentes sociólogos da ciência como Roberto Merton, Pierre Bourdieu e Ric... |
| Row1_Row1 | Os eventos científicos são importantes espaços de articulação dos pesquisadores em suas áreas do conhecimento. Relações sociais de coautoria na produção científica e participação em determinados grupos de trabalho na apresentação oral e divulgação em painéis representam tipos de relação que podemos modelar como redes sociais e analisar seus padrões em busca de entender como uma comunidade científica se articula. O presente trabalho analisa os anais do Encontro Nacional de Pesquisa em Ciência da Informação | 2012 (02) | "Os eventos científicos são importantes espaços de articulação dos pesquis... |
| Row2_Row2 | A pesquisa realizada caracteriza-se como exploratória e descritiva, e analisa as características formais dos periódicos científicos brasileiros da área de Ciências Sociais e de Humanidades indexados na base SciELO. A análise ancorou-se principalmente nos critérios de qualidade extrínsecos de 73 títulos de periódicos, referentes a: entidades editoriais, periodicidade e tempo de existência, fontes de indexação, instruções aos autores, e critérios de avaliação dos artigos. Os resultados revelam que, com relação às características | 2012 (03) | "A pesquisa realizada caracteriza-se como exploratória e descritiva, e an... |
| Row3_Row3 | Esta pesquisa objetiva analisar a contribuição científica brasileira no tema "estudos métricos" para a ciência mainstream, por meio dos periódicos indexados na base Scopus, a fim de visualizar a inserção e o impacto internacional na área. Mais especificamente, propõe-se estudar diacronicamente as pesquisas, identificar os autores mais produtivos e a rede de colaboração científica gerada entre eles e identificar também os periódicos nos quais a produção tem sido disseminada. Fundamenta-se nos | 2012 (04) | "Esta pesquisa objetiva analisar a contribuição científica brasileira no tema "es... |
| Row4_Row4 | A proposição desta pesquisa é analisar os dados relativos aos registros de patentes da UNESP desde seu primeiro registro, em 1980, até dezembro de 2010, de forma a fornecer o panorama e a visibilidade das mesmas ao longo dos anos, analisar os autores e áreas mais produtivos, bem como a rede de coautoria de pesquisadores e a rede de coautoria institucional, e ainda calcular os indicadores de rede de densidade e centralidade de grau. Fundamenta-se nos elementos teóricos-metodológicos da Bibliometria, especialmente nos indicadores de | 2012 (05) | "A proposição desta pesquisa é analisar os dados relativos aos registros de pa... |

Fonte: Captura de tela da pesquisa (2022).

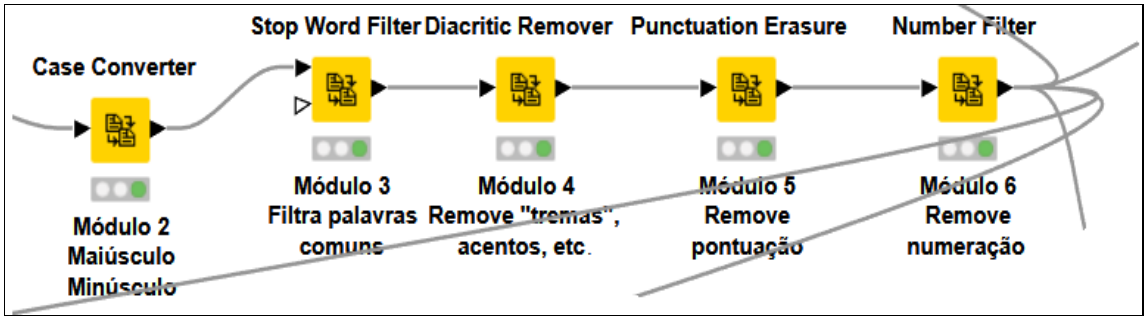
Ainda na Figura 18, observa-se que os documentos estão todos separados, ficando armazenados 1 (um) arquivo em cada linha da tabela, com seus respectivos conteúdos e identificadores definidos nesta pesquisa (nome em que o arquivo foi salvo no computador). Com essas ações executadas, os resumos dos trabalhos apresentados no GT7 nos ENANCIBs de 2012 a 2018, estão todos dentro do *software* de mineração de texto. A execução das ações nos módulos seguintes tem por objetivo fazer com que os textos passem por procedimentos de manipulação dos dados, onde eles serão processados, transformados e preparados para a aplicação dos algoritmos de mineração.

## 5.3 Etapas 2 e 3: manipulação e transformação dos dados

Entende-se que, com as explanações realizadas até esta parte do estudo, os processos de operação e configuração do Knime já estejam mais claros. Assim, desse ponto em diante serão exibidas figuras somente dos módulos mais relevantes, ou a saída de dados já trabalhada por vários módulos, em que os dados manipulados sejam significativos para demonstração. A apresentação dos procedimentos de configuração e saída de dados de todos os módulos tornaria o texto excessivamente extenso.

Dessa maneira, após a conclusão do Módulo 1, que fez o carregamento dos arquivos, relacionando os conteúdos aos seus identificadores (nomes dos arquivos), iniciam-se as Etapas 2 e 3 do processo de descoberta de conhecimento, que são as fases de pré-processamento e transformação dos dados. Interligado aos módulos que formam o Módulo 1 está o nó *Case Converter*, nesta pesquisa designado como Módulo 2 (M2), o qual tem a função de converter as palavras para minúsculo ou maiúsculo (de acordo com a preferência do usuário). Conforme explanado, esse é um exemplo de módulo, cuja função não tem necessidade de apresentação de uma figura apenas para ele. Embora o módulo ajude a padronizar os dados para manipulação futura, compreende-se que seu funcionamento é simples e de fácil entendimento.

Figura 19 – Módulos de transformação de dados




Fonte: Captura de tela da pesquisa (2022).

A Figura 19 apresenta detalhes (parte da Figura 4 – Fluxograma) do encadeamento do conjunto de módulos de 2 a 6. Essa fração da imagem foi extraída do fluxograma principal, demonstrando o nome de cada nó, bem como uma breve descrição da ação que cada um desempenha. Esse grupo de nós, que se inicia no módulo *Case Converter* (M2 - Módulo 2), estão interligados por meio do M2 ao módulo *String To Document*, que é o último nó do Módulo 1, e representa o final da fase de carregamento e leitura.

Interligado ao módulo *Case Converter* (M2) encontra-se o nó *Stop Word Filter* (M3 - Módulo 3), que executa a remoção de palavras consideradas comuns no vocabulário português. Na sequência do processo, o M3 cria uma coluna na tabela para armazenar o conteúdo dos arquivos processados com as alterações realizadas. Os resumos passam então a ficar armazenados sem as palavras auxiliares ou conectivas como interjeições, preposições, artigos e pronomes.

Depois de se efetivar a conversão de todas as palavras para o formato minúsculo (nesta tese) com o *Módulo Case Converter* (M2) e descartar as palavras auxiliares (e, de, para, com, da, a etc.) com o *Módulo Stop Word Filter* (M3), o *Diacritic Remover* (M4 - Módulo 4) faz a exclusão dos diacríticos dos textos (acentos, tremas, cedilhas, til etc.). Em seguida, o *Módulo Punctuation Erasure* (M5 - Módulo 5) remove as pontuações como: ponto-e-vírgula, vírgulas, ponto-final, pontos de interrogação e exclamação etc. O *Módulo Number Filter* (M6 - Módulo 6) retira os numerais dos arquivos, quando necessário, conforme preferência do usuário.

Figura 20 – Transformação dos dados entre os módulos 2 e 6

Table "default" - Rows: 287 Spec - Columns: 2 Properties Flow Variables

| Row ID | File name | Preprocessed Document |
|---|---|---|
| Row0_Row0 | 2012 (01) | "diferentes sociologos ciencia roberto merton pierre bourdieu richard whitley acreditam producao cientifica valorada frente pares principal pesquisador a |
| Row1_Row1 | 2012 (02) | "eventos cientificos importantes espacos articulacao pesquisadores areas conhecimento relacoes sociais coautoria producao cientifica participacao |
| Row2_Row2 | 2012 (03) | "pesquisa realizada caracteriza-se exploratoria descritiva analisa caracteristicas formais periodicos cientificos brasileiros ciencias sociais humanidades i |
| Row3_Row3 | 2012 (04) | "pesquisa objetiva analisar contribuicao cientifica brasileira tema "estudos metricos" ciencia mainstream periodicos indexados base scopus visualizar in |
| Row4_Row4 | 2012 (05) | "proposicao pesquisa analisar dados relativos registros patentes unesp registro dezembro fornecer panorama visibilidade mesmas longo analisar autores |
| Row5_Row5 | 2012 (06) | "tratamento tematico informacao apresenta natureza mediadora dialogar producao usoapropriacao informacao nele verifica-se existencia correntes t |
| Row6_Row6 | 2012 (07) | "conhecimento tradicional assunto interesse cientistas diversas areas conhecimento muitas publicacoes cientificas relacionadas tema desenvolvime |
| Row7_Row7 | 2012 (08) | "presente pesquisa objetivo identificar testes estatisticos quais caracteristicas usuarios reais potenciais interferem quais interferem utilizacao periodic |
| Row8_Row8 | 2012 (09) | "monitoramento sites web diretamente relacionado principal conceito acesso popularidade ambientes digitais reputacao medida so qualidade conteudo |
| Row9_Row9 | 2012 (10) | "estudo objetivo apresentar indicadores gerados base brapci base referencial essencialmente brasileira periodicos publicaram edicoes compara-los indic |
| Row10_Row10 | 2012 (11) | "crescimento desenvolvimento passa atualmente brasil reflete igualmente campo produtividade cientifica pais encontra nacoes representativas visibilidade |
| Row11_Row11 | 2012 (12) | "estudo carater exploratorio inserido contexto pesquisa cujo objetivo tracar panorama pesquisas patentes ciencia informacao analisar aspectos relati |
| Row12_Row12 | 2012 (13) | "apresenta resultados estudo caracteristicas atividades busca acesso uso informacao habitos comunicacao cientifica pesquisadores institutos pesquisa |
| Row13_Row13 | 2012 (14) | "apresenta panorama relacoes autores citados tematicas recorrentes instituicoes produtivas periodo objetivo analisar dinamica institucionalizacao co |
| Row14_Row14 | 2012 (15) | "analise cursos doutorado respectivas teses programas pos-graduacao instituicoes ensino superior brasil publicas privadas arquitetura urbanismo |
| Row15_Row15 | 2012 (16) | "presente artigo intencao verificar estudos realizados co-autoria co-citacao buscando identificar contribuicao brasileira nessas tematicas realizad |
| Row16_Row16 | 2012 (17) | "estudo descritivo utiliza metodo bibliometrico cientometrico verificar caracteristicas tendencias autoria coautoria revistas ciencia informacao brasil |
| Row17_Row17 | 2012 (18) | "analisa insercao tematica 'livro didatico' teses dissertacoes defendidas programas pos-graduacao universidades publicas regioes sul sudeste brasil p |
| Row18_Row18 | 2012 (19) | "estudo exploratorio bibliometrico dedicado identificacao periodicos cientificos expressam base pesquisa tecnico-cientifica producao novas tecnolo |
| Row19_Row19 | 2012 (20) | "artigo trata articulacao matrizes teoricas campo cientifico pierre bourdieu sistemas reputacionais proprios avaliacao producao cientifica richard whitle |
| Row20_Row20 | 2012 (21) | "rede scielo livros lancada brasil scielo livros consorcio edito-ras visa publicacao online colecoes nacionais tematicas livros academicos objeti-vo contribuir |

Fonte: Captura de tela da pesquisa (2022).

Na Figura 20, é possível visualizar o resultado das transformações executadas por esse conjunto de módulos (do M2 ao M6). Mesmo compreendendo que a figura expressa as modificações executadas pelos módulos de forma clara, tais alterações podem ficar ainda mais evidentes ao comparar a coluna *Preprocessed Document*, da Figura 20, com a coluna *Content*, da Figura 18, ficando bastante perceptível as

manipulações realizadas. Estão retratados na Figura 20 os resultados de todas as ações realizados pelos módulos que compõem as fases de seleção, pré-processamento e transformação dos dados, correspondentes as etapas, 1, 2 e 3 do processo de descoberta de conhecimento em texto (KDT). A figura então retrata os procedimentos realizados e efetivados desde o Novo Módulo 1 até o módulo 6. O próximo passo é dar início à Etapa 4, que corresponde à aplicação dos algoritmos de mineração de texto.

## 5.4 Etapa 4: mineração de texto

Após a execução dos procedimentos referentes a etapas iniciais do processo de descoberta de conhecimento, notadamente as fases de seleção da coleção documental, pré-processamento e transformação de dados, alguns pontos podem ser abordados antes da próxima etapa. Uma das primeiras questões a serem relembradas foi a percepção da dinâmica operacional do módulo *Flat File Document Parser,* que fez com que se desconsiderasse seu uso no fluxograma da tese.

Embora faça o carregamento dos arquivos/documentos para dentro do *software* de mineração de texto, cada vez que se executa essa ação, pode ocorrer dos arquivos serem dispostos em sequências diferentes. Como já explanado, em alguns casos, esse evento não traria prejuízos, mas em procedimentos nos quais fosse preciso, por exemplo, encontrar padrões e similaridades entre os textos, tais ocorrências poderiam comprometer a apresentação dos resultados. Em uma eventual necessidade de repetição dos procedimentos de descoberta de conhecimento em texto, além de não alimentar o *software* com os arquivos sempre na mesma sequência, seria impossível apontar o nome do arquivo para identificar o conteúdo nele contido.

Em decorrência dessa circunstância, o módulo *Flat File Document Parser* foi substituído pelos seguintes nós: *List Files/Folders*, *Path to URL*, *URL to file Path*, *Column Filter*, *URL Input*, *Column Filter*, *Joiner*, *Column Filter* e *Strings To Document*. O conjunto de nós apresentados foram posicionados respectivamente nessa sequência, conforme demonstrado no fluxograma (apresentado na Figura 4 e detalhado na Figura 9) e, consequentemente, passaram a compor o Novo Módulo 1 (nome definido neste estudo), portanto, doravante, Novo Módulo 1 e Módulo 1 serão utilizados como sinônimos.

Nesse agrupamento de nós substitutos, os módulos *Column Filter* servem para filtrar as colunas e reduzir a extensão das tabelas, para serem manipuladas pelos próximos nós, deixando armazenadas na nova tabela apenas as colunas relevantes à realização de tarefas posteriores. Por sua vez, o módulo *Strings To Document* tem a função de transformar os textos puros (txt) em um formato de dados do tipo "*document*", que é utilizado dentro do Knime. Esse tipo de dados *document* é um formato que permite a manipulação e processamento em módulos que utilizam técnicas de mineração dos textos.

O incremento de vários módulos em substituição a apenas um foi necessário para se fazer a listagem dos arquivos e transformar o endereço de armazenamento dos textos em um *link.* Por meio dos apontamentos (*links),* foi possível carregar o conteúdo do respectivo arquivo para dentro do Knime. Dessa forma, os nomes dos arquivos que identificam os resumos ficam relacionados corretamente com o seu conteúdo, e essas informações poderão ser utilizadas quando necessário. Essa nova série de módulos permite também que, toda vez que for preciso inserir os arquivos no *software*, eles serão carregados sempre na mesma ordem. Isso permite ao pesquisador ajustar os módulos no decorrer de uma pesquisa e executar as ações quantas vezes for preciso, sem alteração dos resultados.

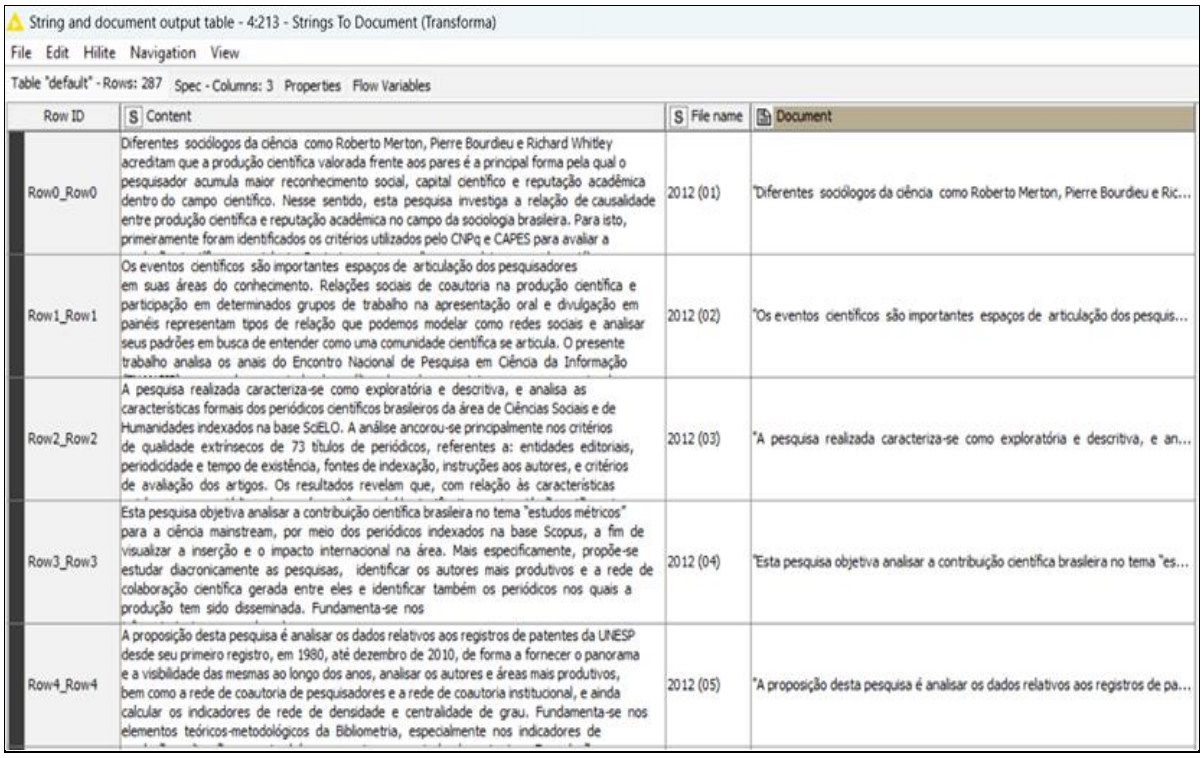
Outro ponto a ser abordado diz respeito aos instrumentos utilizados para a mineração de texto, com intento de viabilizar leituras de uma coleção de documentos. Ficou manifesto tal pressuposto ao se realizar a transformação dos arquivos de PDF para txt, e executar o carregamento dos 287 arquivos, todos de uma só vez e de forma instantânea para dentro da ferramenta. Conforme apresentado na Figura 21, a efetivação do Módulo 1 possibilitou ao Knime encontrar os arquivos, fazer o carregamento, separar e armazenar os nomes dos arquivos e o texto correspondente em células distintas, permitindo aos pesquisadores e ao próprio *software* identificar os arquivos e relacioná-los aos seus conteúdos.

Ao reportar-se à Figura 21, é possível observar, na primeira coluna, os conteúdos dos resumos e, na segunda coluna, o nome de cada arquivo, gravados dessa forma (ano e número sequencial) para identificação dos textos. Saliente-se que, no processo de conversão no *software AntFileConverter* de PDF para txt, o mesmo nome dado ao arquivo no primeiro formato será concebido aos arquivos convertidos, pois o arquivo com nome *2012 (05).pdf* receberá o mesmo nome *2012 (05).txt*,

mudando o respectivo formato e a extensão. Tal informação pode auxiliar o pesquisador para organização e nomeação dos arquivos quando coletados.

A correta identificação dos arquivos possibilita aos algoritmos encontrar as correlações entre eles, ainda que as ligações possam ser demonstradas graficamente. O procedimento de suprir os dados diretamente no *software* de mineração, reduz as chances de ocorrerem equívocos e propicia economia de tempo, pois diminui a intervenção humana na manipulação dos dados. Muitas vezes essas intervenções para manipulação correspondem às ações de copiar e colar o conteúdo dos arquivos em programas, como por exemplo, planilhas eletrônicas (Excel, Calc etc.).

Figura 21 – Saída de dados da Etapa 1


String and document output table - 4:213 - Strings To Document (Transforma)

File Edit Hilite Navigation View

Table "default" - Rows: 287 Spec - Columns: 3 Properties Flow Variables

| Row ID | Content | File name | Document |
|---|---|---|---|
| Row0_Row0 | Diferentes sociólogos da ciência como Roberto Merton, Pierre Bourdieu e Richard Whitley acreditam que a produção científica valorada frente aos pares é a principal forma pela qual o pesquisador acumula maior reconhecimento social, capital científico e reputação acadêmica dentro do campo científico. Nesse sentido, esta pesquisa investiga a relação de causalidade entre produção científica e reputação acadêmica no campo da sociologia brasileira. Para isto, primeiramente foram identificados os critérios utilizados pelo CNPq e CAPES para avaliar a | 2012 (01) | "Diferentes sociólogos da ciência como Roberto Merton, Pierre Bourdieu e Ric... |
| Row1_Row1 | Os eventos científicos são importantes espaços de articulação dos pesquisadores em suas áreas do conhecimento. Relações sociais de coautoria na produção científica e participação em determinados grupos de trabalho na apresentação oral e divulgação em painéis representam tipos de relação que podemos modelar como redes sociais e analisar seus padrões em busca de entender como uma comunidade científica se articula. O presente trabalho analisa os anais do Encontro Nacional de Pesquisa em Ciência da Informação | 2012 (02) | "Os eventos científicos são importantes espaços de articulação dos pesquis... |
| Row2_Row2 | A pesquisa realizada caracteriza-se como exploratória e descritiva, e analisa as características formais dos periódicos científicos brasileiros da área de Ciências Sociais e de Humanidades indexados na base SciELO. A análise ancorou-se principalmente nos critérios de qualidade extrínsecos de 73 títulos de periódicos, referentes a: entidades editoriais, periodicidade e tempo de existência, fontes de indexação, instruções aos autores, e critérios de avaliação dos artigos. Os resultados revelam que, com relação às características | 2012 (03) | "A pesquisa realizada caracteriza-se como exploratória e descritiva, e an... |
| Row3_Row3 | Esta pesquisa objetiva analisar a contribuição científica brasileira no tema "estudos métricos" para a ciência mainstream, por meio dos periódicos indexados na base Scopus, a fim de visualizar a inserção e o impacto internacional na área. Mais especificamente, propõe-se estudar diacronicamente as pesquisas, identificar os autores mais produtivos e a rede de colaboração científica gerada entre eles e identificar também os periódicos nos quais a produção tem sido disseminada. Fundamenta-se nos | 2012 (04) | "Esta pesquisa objetiva analisar a contribuição científica brasileira no tema "es... |
| Row4_Row4 | A proposição desta pesquisa é analisar os dados relativos aos registros de patentes da UNESP desde seu primeiro registro, em 1980, até dezembro de 2010, de forma a fornecer o panorama e a visibilidade das mesmas ao longo dos anos, analisar os autores e áreas mais produtivos, bem como a rede de coautoria de pesquisadores e a rede de coautoria institucional, e ainda calcular os indicadores de rede de densidade e centralidade de grau. Fundamenta-se nos elementos teóricos-metodológicos da Bibliometria, especialmente nos indicadores de | 2012 (05) | "A proposição desta pesquisa é analisar os dados relativos aos registros de pa... |

Fonte: Captura de tela da pesquisa (2022).

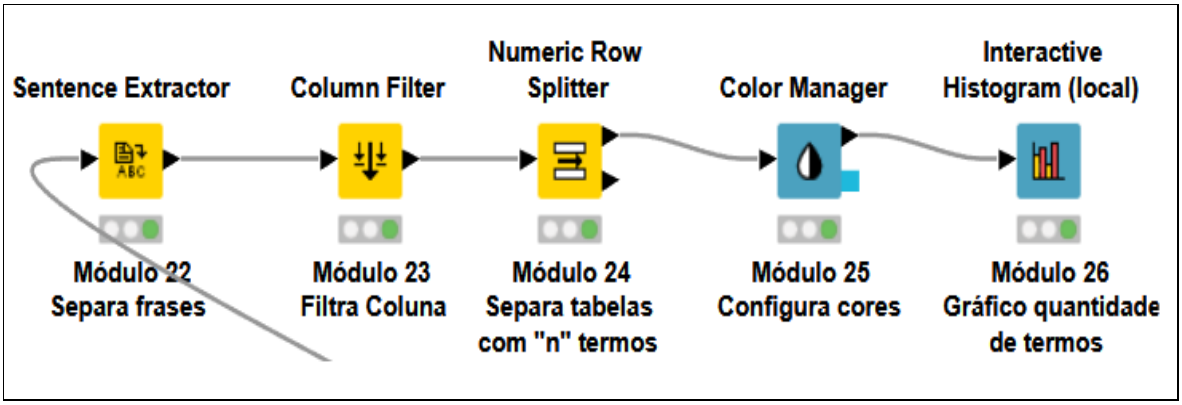
Resolvida essa questão de carregamento correto dos arquivos (Etapa 1 – seleção), passou-se então para as fases 2 e 3, de pré-processamento e transformação; e para isso foram utilizados os módulos: (M2) *Case Converter*; (M3) *Stop Word Filter*; (M4) *Diacritic Remover*; (M5) *Punctuation Erasure* e (M6) *Number Filter*. No intervalo de módulos apresentados (Módulo 2 ao Módulo 6), realizou-se a conversão das palavras para minúsculas, retiraram-se as palavras consideradas comuns no vocabulário português (Brasil) e os diacríticos, além de removerem-se as pontuações e numerações desnecessárias, deixando os resumos limpos para serem trabalhados pelos algoritmos de mineração (conforme demonstrado na Figura 20).

Os próximos tópicos abordam a execução da Etapa 4, em que serão apresentadas as técnicas de mineração de texto. Esses tópicos reproduzem uma parte importante da metodologia proposta nesta tese para o levantamento terminológico, pois demonstram a utilização de algoritmos de mineração textual, que conseguem executar de maneira efetiva e automatizada as técnicas de dedução de frequência, coocorrência de termos e a análise de classificação temática.

### 5.4.1 Padrões das sentenças no *corpus*

O próximo resultado do conjunto de módulos demonstra os padrões de escrita dos autores, referentes à quantidade de termos utilizados em cada sentença. A Figura 22 (fragmento da Figura 4 – Fluxograma) apresenta o encadeamento desses módulos que se inicia no nó M22 e termina no nó M26. O nó *Sentence Extractor* (M22 – Módulo 22) está conectado ao módulo *Number Filter* (M6), pois, esse módulo contém os dados preparados nas etapas anteriores (1, 2 e 3) para serem processados na fase de mineração de texto.

Figura 22 – Módulos de extração de informações das sentenças



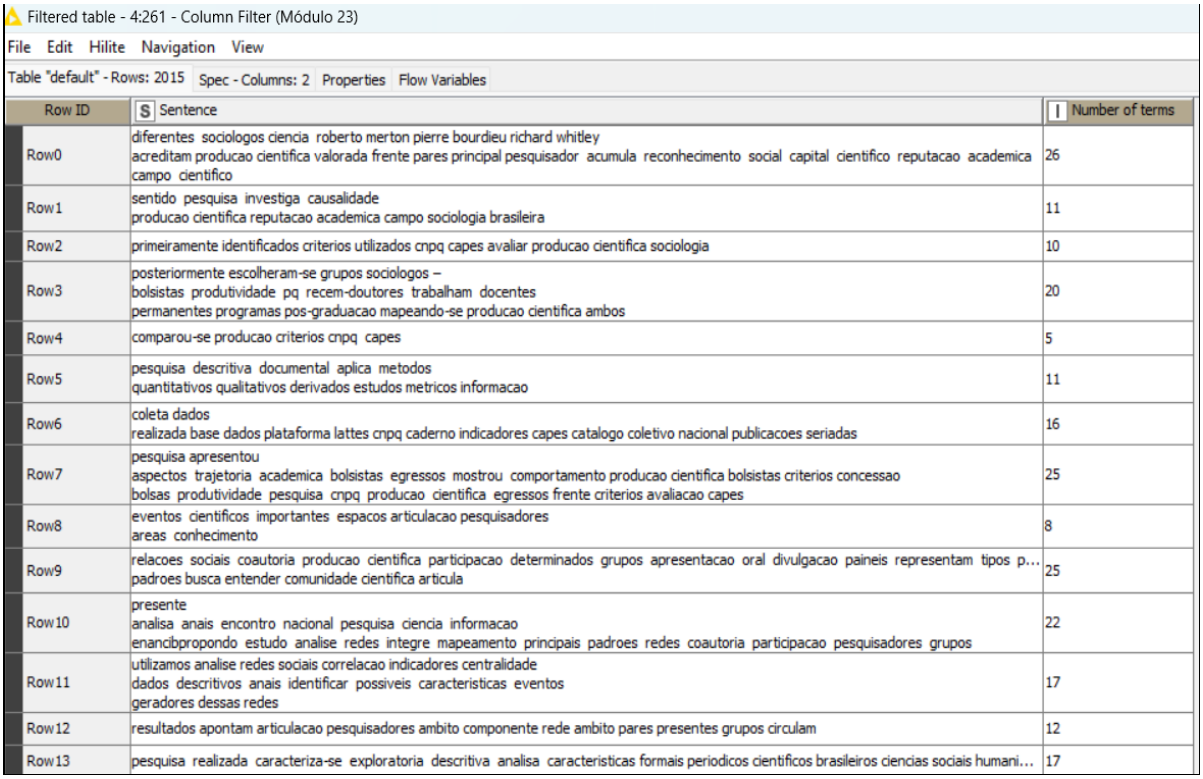

Fonte: Captura de tela da pesquisa (2022).

O módulo *Sentence Extractor* (M22) separa o texto do resumo em frases e faz a contagem da quantidade de termos que compõem as sentenças, armazenando esses dados em linhas distintas, como está demonstrado na Figura 23. Saliente-se que, nesse módulo, já estão descartadas a pontuação e as palavras consideradas comuns, desprezadas em módulos de transformação, anteriores a esta etapa. A figura também ilustra a quantidade total de 2015 sentenças encontradas e separadas, expressando que, os 287 resumos estão agora divididos nessa quantidade de frases e ainda, que o tipo de dados armazenado passa a ser *Sentence.*

A Figura 23 retrata também a utilização de um *Módulo Column Filter* (M23 – Módulo 23) no agrupamento de nós, que novamente serviu para filtrar e apresentar somente as colunas *Sentence* e *Number of Terms* do módulo anterior. Na sequência do encadeamento, o nó identificado como *Numeric Row Splitter* (M24 - Módulo 24), permite dividir e classificar a tabela por quantidade de termos, valor que que deve ser determinado pelo usuário.

Assim, no módulo (M24), é possível realizar uma configuração em que, conforme as definições (quantidade de termos), algumas frases da tabela serão aceitas e outras descartadas, ocorrendo a separação da tabela completa em outras duas. A primeira tabela armazenará as frases que obedecem ao critério definido na quantidade de termos (aceitas) e a outra estará com frases que estão fora do parâmetro (descartadas). Desse modo, há duas possibilidades de saídas de dados do módulo *Numeric Row Splitter* (M24), uma tabela com as frases que obedecem ao critério definido (*Data accepted*) e outra com frases que não obedecem aos critérios (*Data discarded*).

Figura 23 – Filtro da saída de dados do módulo *Sentence Extractor*


Filtered table - 4:261 - Column Filter (Módulo 23)

File Edit Hilite Navigation View

Table "default" - Rows: 2015 Spec - Columns: 2 Properties Flow Variables

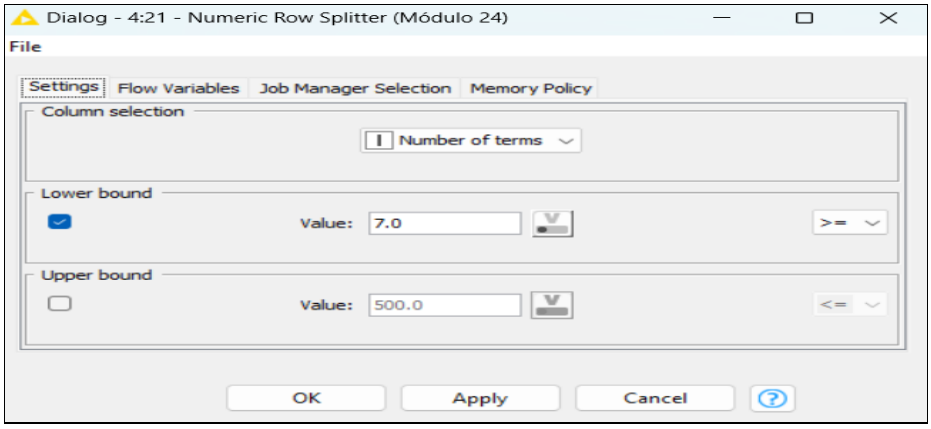
| Row ID | S Sentence | I Number of terms |
|---|---|---|
| Row0 | diferentes sociologos ciencia roberto merton pierre bourdieu richard whitley<br>acreditam producao cientifica valorada frente pares principal pesquisador acumula reconhecimento social capital cientifico reputacao academica<br>campo cientifico | 26 |
| Row1 | sentido pesquisa investiga causalidade<br>producao cientifica reputacao academica campo sociologia brasileira | 11 |
| Row2 | primeiramente identificados criterios utilizados cnpq capes avaliar producao cientifica sociologia | 10 |
| Row3 | posteriormente escolheram-se grupos sociologos –<br>bolsistas produtividade pq recem-doutores trabalham docentes<br>permanentes programas pos-graduacao mapeando-se producao cientifica ambos | 20 |
| Row4 | comparou-se producao criterios cnpq capes | 5 |
| Row5 | pesquisa descritiva documental aplica metodos<br>quantitativos qualitativos derivados estudos metricos informacao | 11 |
| Row6 | coleta dados<br>realizada base dados plataforma lattes cnpq caderno indicadores capes catalogo coletivo nacional publicacoes seriadas | 16 |
| Row7 | pesquisa apresentou<br>aspectos trajetoria academica bolsistas egressos mostrou comportamento producao cientifica bolsistas criterios concessao<br>bolsas produtividade pesquisa cnpq producao cientifica egressos frente criterios avaliacao capes | 25 |
| Row8 | eventos cientificos importantes espacos articulacao pesquisadores<br>areas conhecimento | 8 |
| Row9 | relacoes sociais coautoria producao cientifica participacao determinados grupos apresentacao oral divulgacao paineis representam tipos p...<br>padroes busca entender comunidade cientifica articula | 25 |
| Row10 | presente<br>analisa anais encontro nacional pesquisa ciencia informacao<br>enancibpropondo estudo analise redes integre mapeamento principais padroes redes coautoria participacao pesquisadores grupos | 22 |
| Row11 | utilizamos analise redes sociais correlacao indicadores centralidade<br>dados descritivos anais identificar possiveis caracteristicas eventos<br>geradores dessas redes | 17 |
| Row12 | resultados apontam articulacao pesquisadores ambito componente rede ambito pares presentes grupos circulam | 12 |
| Row13 | pesquisa realizada caracteriza-se exploratoria descritiva analisa caracteristicas formais periodicos cientificos brasileiros ciencias sociais humani... | 17 |

Fonte: Captura de tela da pesquisa (2022).

A Figura 24 apresenta as opções de configuração do módulo *Numeric Row Splitter* (M24). Neste estudo, optou-se por deixar apenas as sentenças que contenham

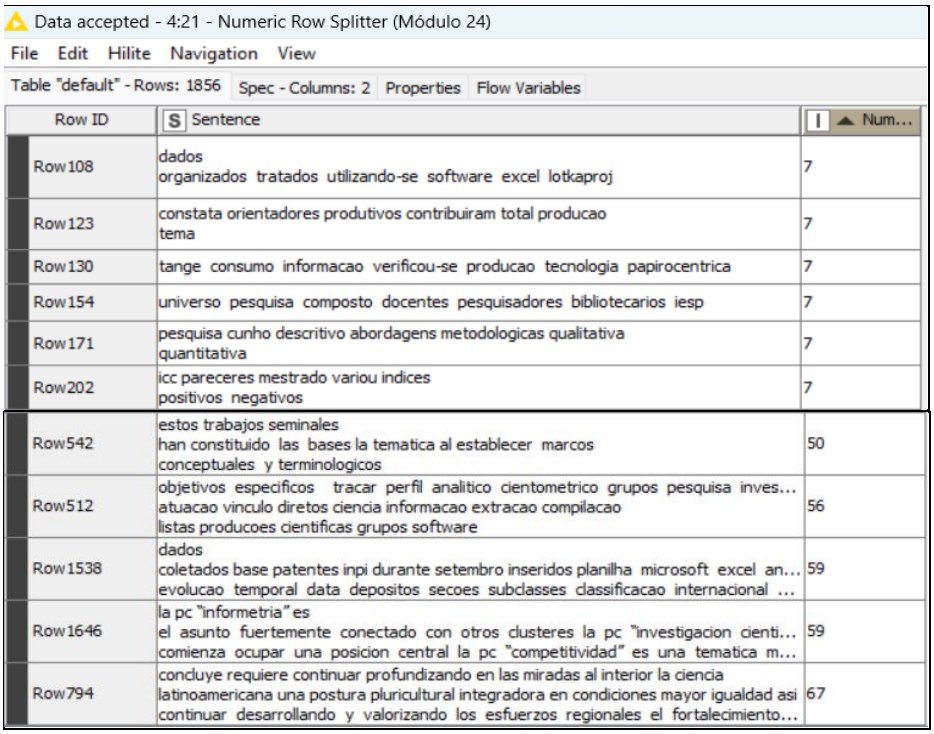
no mínimo 7 palavras, descartando as frases com 6 ou menos, pois, na mineração de texto frases com essas quantidades de termos (1 a 6) são consideradas simples e sem muita relevância. A Figura 25 apresenta frações do início e do final da tabela completa, com as frases que obedecem aos parâmetros (*Data accepeted*) definidos, demonstrando que 1856 sentenças são compostas por 7 ou mais termos.

Figura 24 – Opções de configuração do módulo *Numeric Row Splitter*




Fonte: Captura de tela da pesquisa (2022).

Figura 25 – *Data accepted* do módulo *Numeric Row Splitter*



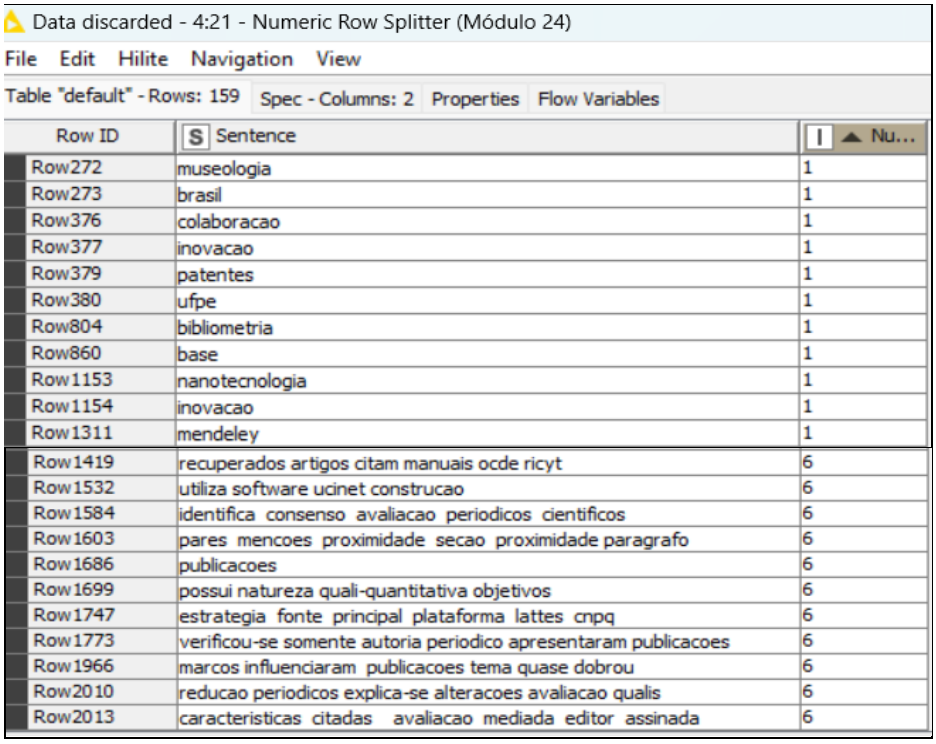

| Row ID | Sentence | Num... |
|---|---|---|
| Row108 | dados<br>organizados tratados utilizando-se software excel lotkaproj | 7 |
| Row123 | constata orientadores produtivos contribuiram total producao<br>tema | 7 |
| Row130 | tange consumo informacao verificou-se producao tecnologia papirocentrica | 7 |
| Row154 | universo pesquisa composto docentes pesquisadores bibliotecarios iesp | 7 |
| Row171 | pesquisa cunho descritivo abordagens metodologicas qualitativa<br>quantitativa | 7 |
| Row202 | icc pareceres mestrado variou indices<br>positivos negativos | 7 |
| Row542 | estos trabajos seminales<br>han constituido las bases la tematica al establecer marcos<br>conceptuales y terminologicos | 50 |
| Row512 | objetivos especificos tracar perfil analitico cientometrico grupos pesquisa inves...<br>atuacao vinculo diretos ciencia informacao extracao compilacao<br>listas producoes cientificas grupos software | 56 |
| Row1538 | dados<br>coletados base patentes inpi durante setembro inseridos planilha microsoft excel an...<br>evolucao temporal data depositos secoes subclasses classificacao internacional ... | 59 |
| Row1646 | la pc "informetria" es<br>el asunto fuertemente conectado con otros clusteres la pc "investigacion cienti...<br>comienza ocupar una posicion central la pc "competitividad" es una tematica m... | 59 |
| Row794 | concluye requiere continuar profundizando en las miradas al interior la ciencia<br>latinoamericana una postura pluricultural integradora en condiciones mayor igualdad asi<br>continuar desarrollando y valorizando los esfuerzos regionales el fortalecimiento... | 67 |

Fonte: Captura de tela da pesquisa (2022).

Na sequência, conforme demonstrado na Figura 26, estão também as frações inicial e final da segunda tabela, as quais contém 6 ou menos termos, evidenciando que foram encontradas 159 sentenças fora do parâmetro (*Data discarded*). Destaque-

se que, as configurações de quantidade para descartar ou não as frases ficam a critério do pesquisador, pois, procurou-se aqui demonstrar que há possibilidades de definição dos valores quando necessário (visto na Figura 24).

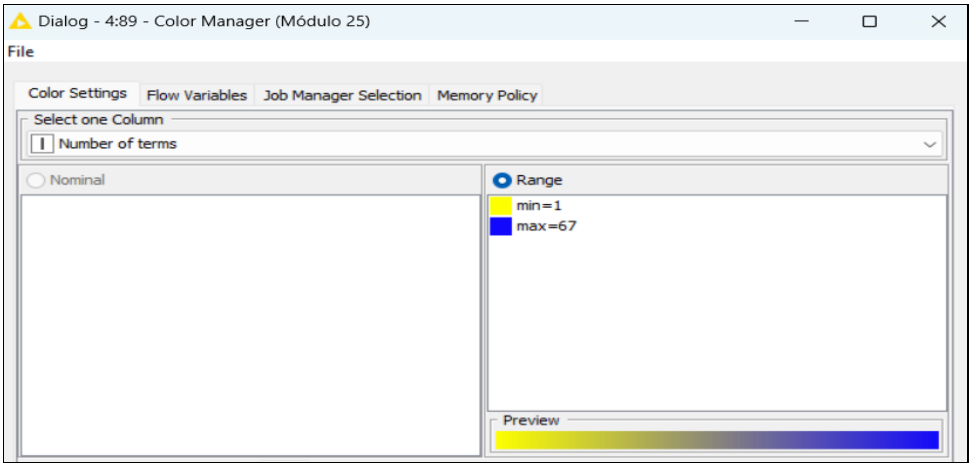
Figura 26 – *Data discarded* do módulo *Numeric Row Splitter*




| Row ID | Sentence | Nu... |
|---|---|---|
| Row272 | museologia | 1 |
| Row273 | brasil | 1 |
| Row376 | colaboracao | 1 |
| Row377 | inovacao | 1 |
| Row379 | patentes | 1 |
| Row380 | ufpe | 1 |
| Row804 | bibliometria | 1 |
| Row860 | base | 1 |
| Row1153 | nanotecnologia | 1 |
| Row1154 | inovacao | 1 |
| Row1311 | mendeley | 1 |
| Row1419 | recuperados artigos citam manuais ocde ricyt | 6 |
| Row1532 | utiliza software ucinet construcao | 6 |
| Row1584 | identifica consenso avaliacao periodicos cientificos | 6 |
| Row1603 | pares mencoes proximidade secao proximidade paragrafo | 6 |
| Row1686 | publicacoes | 6 |
| Row1699 | possui natureza quali-quantitativa objetivos | 6 |
| Row1747 | estrategia fonte principal plataforma lattes cnpq | 6 |
| Row1773 | verificou-se somente autoria periodico apresentaram publicacoes | 6 |
| Row1966 | marcos influenciaram publicacoes tema quase dobrou | 6 |
| Row2010 | reducao periodicos explica-se alteracoes avaliacao qualis | 6 |
| Row2013 | caracteristicas citadas avaliacao mediada editor assinada | 6 |

Fonte: Captura de tela da pesquisa (2022).

Figura 27 – Paleta de configuração do módulo *Color Manager*



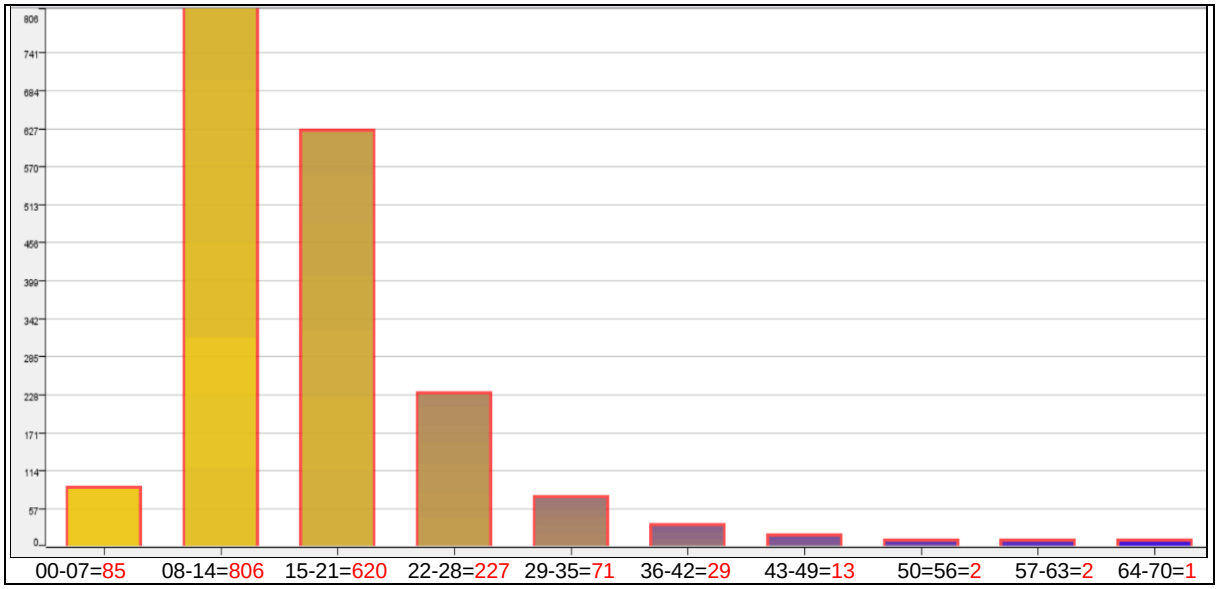

Fonte: Captura de tela da pesquisa (2022).

O módulo *Color Manager* (M25 – Módulo 25) serve apenas para definir cores para um determinado intervalo de valores ou palavras, ou para qualquer outra saída de dados em que seja possível definir cores. É um módulo de auxílio para diferenciação de elementos nas apresentações dos resultados, utilizado para

melhorar a compreensão e visualização gráfica. A Figura 27 mostra a tela com as opções de configuração desse módulo, revelando ainda o intervalo de quantidade de termos encontradas nas sentenças, conforme demonstrado na coluna *Range* da figura (min=1 – max=67).

Nesta pesquisa, a configuração do módulo *Color Manager* será utilizada para auxiliar na apresentação do Gráfico 1, que foi gerado pelo módulo *Interactive Histogram* (local) (M26 - Módulo 26). Além de apresentar o intervalo da quantidade de termos das sentenças, o gráfico também mostra a quantidade de frases escritas e o quantitativo de termos. O módulo *Interactive Histogram* (local) permite definir o número de colunas a serem apresentadas, e neste estudo se utilizou o padrão do *software*, que são 10 colunas.

Gráfico 1 – Quantidade de sentenças e termos




Fonte: Captura de tela da pesquisa (2022).

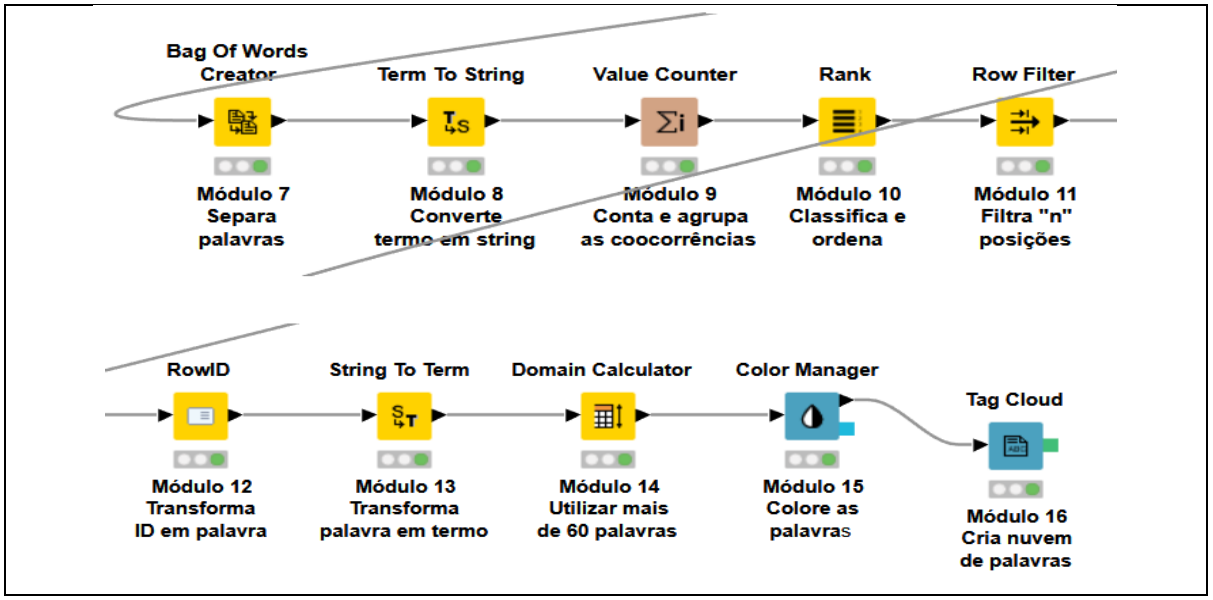
A primeira coluna no gráfico informa que, no intervalo de 00 a 07 termos, foram encontradas 85 sentenças. Ressalta-se que, conforme exposto e demonstrado na Figura 26, as sentenças que possuíam de 01 a 06 termos foram descartadas, as quais totalizavam 159 frases. Por isso, infere-se que nessa primeira coluna estão apenas as sentenças com 07 termos. A segunda coluna agrupa as sentenças do intervalo de 08 a 14 termos, somando 806 frases. E assim, segue-se, consecutivamente, a terceira coluna aponta 620 sentenças que possuem entre 15 e 21 termos, até chegar à décima coluna, a qual mostra que há 1 sentença no intervalo entre 64 e 70 termos. Ao reportar-se a Figura 25, pode-se notar que há uma única sentença com 67 termos, confirmando o que é apresentado no gráfico.

Outra confirmação que o gráfico demonstra acontece ao se somar a quantidade de sentenças apresentadas, (C1-85, C2-806, C3-620, C4-227, C5-71, C6-29, C7-13, C8-2, C9-2 e C10-1), pois, com o cálculo tem-se o total de 1856, novamente confirmando o que se apresentou na Figura 25, a qual mostra que essa é a quantidade de sentenças que satisfazem ao critério de possuírem 7 ou mais termos. Complementando, ao retomar-se a informação de que 159 sentenças não cumpriam a esse requisito e são compostas pelo intervalo de 1 a 6 termos, encontra-se o total de sentenças recuperado pelo módulo *Sentence Extractor*, que é de 2015 (1856 + 159 = 2015), apresentado na Figura 23.

### 5.4.2 Minerando frequência de termos

No agrupamento de módulos seguinte, utiliza-se o método de dedução de frequência dos termos da área da Linguística, com início no módulo 7 e término no módulo 16. Essa sequência de nós foi utilizada para fazer a leitura dos dados e minerar não só a frequências dos termos simples, mas também as coocorrências dos termos compostos no *corpus* estudado. Assim como outros módulos que utilizam os dados já manipulados pelas etapas de seleção, pré-processamento e transformação, o módulo *Bag Of Words Creator* (M7 – Módulo 7) também busca as informações no Módulo 6 (pode ser conferido na Figura 20).

Figura 28 – Módulos de mineração de frequência e coocorrência de termos



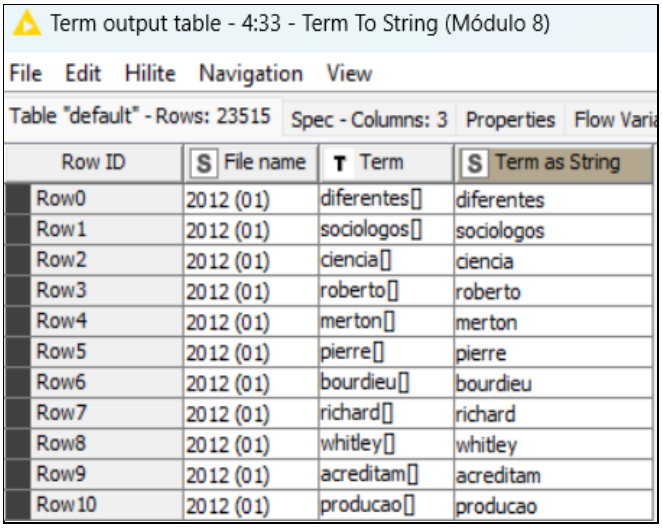

Fonte: Captura de tela da pesquisa (2022).

O encadeamento desses módulos é demonstrado na Figura 28 (fragmento da Figura 4 ampliado para melhor visualização), na qual é possível verificar os nós que

fazem parte do processo de mineração para o ranqueamento dos termos. Os módulos são: *Bag Of Word Creator*, *Term To String*, *Value Counter*, *Rank*, *Row Filter*, *RowID*, *String To Term*, *Domain Calculator*, *Color Manager* e *Tag Cloud*. A figura foi dividida em duas partes e posicionada uma abaixo da outra, apenas para auxiliar na visualização, mas no fluxograma principal (Figura 4) esse conjunto de módulos está em sequência linear, um após o outro.

No módulo *Bag Of Words Creator* (M7 – Módulo 7), o *software* faz a leitura das sentenças e as decompõe em palavras (*token*), gerando uma nova tabela com todas as expressões encontradas, isso significa que cada palavra é disposta separadamente em uma linha específica da tabela. Nesse módulo são armazenados todos os termos encontrados, independentemente da quantidade de repetições eles são arrumados em separado, pois o módulo não realiza o agrupamento dos termos repetidos, e dessa forma o total de 23.515 expressões foram encontradas. O módulo *Term To String* (M8 – Módulo 8) transforma os termos em um formato da classe *string*, fazendo a preparação para um tipo de saída de dados que pode ser lido pelo próximo nó. Uma fração da tabela com os resultados dos módulos 7 e 8 podem ser observados na Figura 29, que também mostra o nome do arquivo a que pertence determinado termo.

Figura 29 – Saída dos dados dos módulos 7 e 8



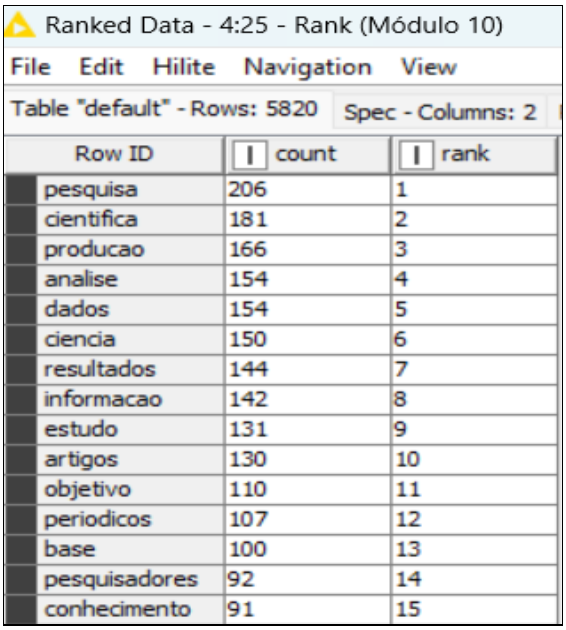

| Row ID | File name | Term | Term as String |
|---|---|---|---|
| Row0 | 2012 (01) | diferentes[] | diferentes |
| Row1 | 2012 (01) | sociologos[] | sociologos |
| Row2 | 2012 (01) | ciencia[] | ciencia |
| Row3 | 2012 (01) | roberto[] | roberto |
| Row4 | 2012 (01) | merton[] | merton |
| Row5 | 2012 (01) | pierre[] | pierre |
| Row6 | 2012 (01) | bourdieu[] | bourdieu |
| Row7 | 2012 (01) | richard[] | richard |
| Row8 | 2012 (01) | whitley[] | whitley |
| Row9 | 2012 (01) | acreditam[] | acreditam |
| Row10 | 2012 (01) | producao[] | producao |

Fonte: Captura de tela da pesquisa (2022).

Conforme se demonstra na Figura 30, o critério de ordenação dos termos neste estudo foi definido como crescente, identificando que o termo “pesquisa” apareceu 206 vezes no *corpus* e a palavra “cientifica” (o acento foi removido nas etapas de limpeza) apareceu 181 vezes, estando as duas expressões na primeira e segunda posição, respectivamente. Pode-se identificar também que, dos 23.515 termos totais

iniciais, quando se considera as repetições e faz-se o agrupamento, passa-se a ter 5.820 termos encontrados no *corpus*. Destaque-se novamente que não estão computados os sinais de pontuações e as palavras consideradas comuns no vocabulário português.

Figura 30 – Saída dos dados dos módulos 9 e 10



Ranked Data - 4:25 - Rank (Módulo 10)

File Edit Hilite Navigation View

Table "default" - Rows: 5820 Spec - Columns: 2

| Row ID | count | rank |
|---|---|---|
| pesquisa | 206 | 1 |
| cientifica | 181 | 2 |
| producao | 166 | 3 |
| analise | 154 | 4 |
| dados | 154 | 5 |
| ciencia | 150 | 6 |
| resultados | 144 | 7 |
| informacao | 142 | 8 |
| estudo | 131 | 9 |
| artigos | 130 | 10 |
| objetivo | 110 | 11 |
| periodicos | 107 | 12 |
| base | 100 | 13 |
| pesquisadores | 92 | 14 |
| conhecimento | 91 | 15 |

Fonte: Captura de tela da pesquisa (2022).

Os módulos 11, 12, 13, 14 e 15 (apresentados na Figura 28) apenas preparam a saída de dados para uma forma legível aos próximos nós. Por exemplo, no módulo *Row Filter* (M11 – Módulo 11) é possível determinar a quantidade ou o intervalo de linhas que se deseja armazenar e apresentar. Neste estudo, configurou-se para serem apresentados os 100 termos mais bem ranqueados, pois, entende-se que a apresentação de uma quantidade maior seria de difícil visualização, entretanto, no *software* ficam armazenados todos os dados. O módulo *RowID* (M12 – Módulo12) armazena em uma coluna o identificador de cada linha, pois o módulo *Value Counter* (M9) transformou os termos encontrados em identificadores de linha. Essa conversão no M12 é necessária manipular os dados novamente, pois os módulos não conseguem realizar algumas ações nos identificadores de linha.

O módulo *String To Term* (M13 – Módulo 13) apenas transforma o dado do tipo *string* do módulo anterior para a classe *term*, o qual pode ser manipulado pelo nó seguinte. O *Domain Calculator* (M14 – Módulo) permite que sejam apresentados mais termos do que a configuração padrão do *software* aceita (60 palavras). Como pretende-se apresentar os 100 primeiros termos, a utilização desse módulo se fez

necessária. O nó *Color Manager* (M15 – Módulo 15), conforme já explanado, serve para diferenciar por meio de cores algumas saídas de dados para apresentação dos resultados em modo gráfico.

No módulo *Tag Cloud* (M16 - Módulo 16), os resultados dos refinamentos feitos nos nós anteriores são apresentados, pois, esse módulo gera a nuvem dos termos mais utilizados pelos autores. A Figura 31 apresenta os 100 principais termos dos 5.820 encontrados e, quanto maior a palavra na imagem e mais posicionada a direita inferior da figura, melhor é sua colocação no ranque. Na figura, é possível perceber que as posições 1 e 2 são ocupadas pelos termos “pesquisa” e “cientifica” e os termos “capes” e “busca” estão na posição 99 e 100, respectivamente.

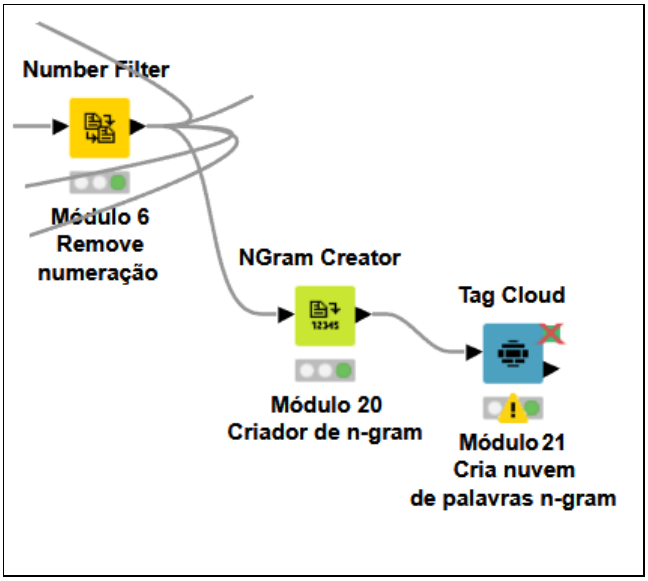
Figura 31 – Principais termos do *corpus*


Fonte: Dados da pesquisa (2022).

### 5.4.3 Minerando coocorrência de n-gramas

Depois de realizar o levantamento dos principais termos utilizados pelos autores, ainda com o método de dedução de frequência e coocorrência de termos, busca-se então encontrar os n-gramas (*cluster* ou combinações de palavras), que serve para demonstrar a estrutura comunicacional dos textos pesquisados. O módulo que executa essa ação é o *Ngram Creator* (M20 – Módulo 20), que também está interligado ao módulo 6 (M6) - (executa a última tarefa das etapas de seleção, pré-processamento e transformação). O módulo *Tag Cloud* (M21 - Módulo 21) dessa sequência de módulos servirá novamente para apresentar os n-gramas extraídos dos resumos. Na Figura 32, é possível observar a interligação desse agrupamento de módulos (ampliação de parte da Figura 4 - fluxograma principal).

Figura 32 – Módulos para extração de n-gramas




Fonte: Captura de tela da pesquisa (2022).

O módulo *NGram Creator* utiliza um algoritmo que possibilita encontrar *clusters* de palavras e a quantidade de ocorrências desses *clusters* no conjunto dos textos. Nesse módulo, ainda é possível realizar a configuração da quantidade de combinação de termos que se pretende encontrar como bigramas, trigramas, tetragramas, e assim sucessivamente. Tais configurações devem ser feitas em conformidade ao escopo e particularidade de cada pesquisa. Nesta pesquisa, apresentaremos apenas os n-gramas compostos por dois e três termos (Figuras 33 e 34), pois o intuito é demonstrar as possibilidades de descoberta de conhecimento com o uso da metodologia, e não a inspeção dos termos em si. As Figuras 33 e 34 foram geradas a partir da seleção do elemento n-grams, presente no menu de opções do módulo *NGram Creator*.

Visualiza-se nessas duas figuras o total de ocorrência dos *clusters* de n-gramas encontrados, os quais são demostrados na coluna *Corpus frequency*. A coluna *Document frequency* representa o número de documentos individuais em que cada n-grama aparece. Os valores da coluna *Sentence frequency* representam a quantidade de frases do *corpus* que contém o n-grama. O *Corpus frequency* é sempre maior ou pelo menos igual às quantias das outras colunas, porque esse campo considera o total de vezes que um n-grama aparece no conjunto completo dos arquivos, em que são computadas todas as repetições em um documento ou nas frases.

Por exemplo, a coluna *Document frequency* é incrementada sempre que um n-grama é encontrado dentro de um documento, mesmo existindo outras ocorrências desse n-grama no texto, ele será contado uma única vez, pois, essa coluna verifica

apenas a existência do referido n-grama no documento, independentemente da quantidade de vezes que ele se repete. Isso também acontece na coluna *Sentence frequency*, pois um mesmo n-grama pode aparecer mais de uma vez em uma frase, contudo, como o cômputo se dá para a presença do n-grama na sentença e não pelo número de repetições, apenas uma ocorrência será considerada. E, é por isso que o valor da coluna *Corpus frequency* sempre será no mínimo igual a contagem dos outros campos, ou maior que eles.

As primeiras linhas da Figura 33 apresentam as combinações dos termos: produção científica; ciência informação; artigos publicados; programas pós-graduação; periódicos científicos e comunicação científica; e tais conjunto de termos são parte do total de 22.238 bigramas encontrados. Pode-se ainda observar na figura que, o bigrama “produção cientifica” aparece ao todo 203 vezes no *corpus* (*Corpus frequency*), e que ele foi encontrado em 120 documentos do total de 287 (*Document frequency*), estando ainda presente em 193 frases (*Sentence frequency*). Já o bigrama “ciência informação” repetiu-se um total de 194 vezes no *corpus*; e 99 documentos continham essa combinação de palavras, com o bigrama estando presente em 170 sentenças. Assim é possível entender e seguir-se com a leitura da figura, que apresente alguns bigramas com suas respectivas frequências no *corpus*, documentos e sentenças.

Figura 33 – Principais bigramas do *corpus*

Ngrams - 3:198 - NGram Creator (Módulo 20)

File Edit Hilite Navigation View

Table "default" - Rows: 22238 Spec - Columns: 4 Properties Flow Variables

| Row ID | Ngram | Corpus frequency | Document frequency | Sentence frequency |
|---|---|---|---|---|
| 10 | producao cientifica | 203 | 120 | 193 |
| 209 | ciencia informacao | 194 | 99 | 170 |
| 950 | artigos publicados | 60 | 48 | 56 |
| 52 | programas pos-graduacao | 58 | 35 | 54 |
| 116 | periodicos cientificos | 55 | 32 | 54 |
| 1106 | comunicacao cientifica | 53 | 37 | 50 |
| 2162 | universidade federal | 49 | 38 | 39 |
| 74 | base dados | 41 | 36 | 40 |
| 195 | redes sociais | 37 | 25 | 35 |
| 1153 | web of | 37 | 32 | 37 |
| 1446 | artigos periodicos | 36 | 26 | 35 |
| 2709 | grupos pesquisa | 36 | 16 | 33 |
| 179 | areas conhecimento | 35 | 31 | 34 |

Fonte: Captura de tela da pesquisa (2022).

Na Figura 34, pode-se inferir que foram encontradas 25.088 combinações com três termos; e os trigramas que aparecem com maior número de coocorrências são: *web of science*; ciência informação brasil; desenvolvimento científico tecnológico;

estudos métricos informação; nacional desenvolvimento científico, análise de redes sociais, além dos outros que podem sem observados na figura.

Novamente, é possível verificar na Figura 34 que, o trigrama "*web of science*" repete-se 35 vezes no *corpus (Corpus frequency)*; e que 30 documentos *(Document frequency)* contém esse trigrama, o qual está presente em 35 frases (*Sentence frequency*). E assim ocorre sucessivamente com os outros trigramas encontrados, conforme é possível visualizar nos dados apresentados na figura. É possível ainda observar de forma mais clara a questão do *Corpus frequency* ser sempre maior ou igual aos outros valores de coocorrência[43], pois, vários trigramas tem valores iguais para os campos *Corpus frequency* e *Setence frequency*.

Figura 34 – Principais trigramas do *corpus.*

Ngrams - 3:198 - NGram Creator (Módulo 20)

File Edit Hilite Navigation View

Table "default" - Rows: 25088 Spec - Columns: 4 Properties Flow Variables

| Row ID | Ngram | Corpus frequency | Document frequency | Sentence frequency |
|---|---|---|---|---|
| 901 | web of science | 35 | 30 | 35 |
| 518 | ciencia informacao brasil | 22 | 18 | 22 |
| 2041 | desenvolvimento cientifico tecnologico | 21 | 16 | 21 |
| 69 | estudos metricos informacao | 18 | 10 | 16 |
| 2040 | nacional desenvolvimento cientifico | 18 | 13 | 18 |
| 220 | analise redes sociais | 14 | 12 | 14 |
| 1727 | pos - graduacao | 14 | 14 | 14 |
| 204 | pesquisa ciencia informacao | 13 | 9 | 12 |
| 1105 | pos-graduacao ciencia informacao | 13 | 10 | 13 |
| 1251 | bolsistas produtividade pesquisa | 13 | 8 | 13 |
| 2905 | producao cientifica brasileira | 13 | 12 | 13 |
| 5676 | indexados base dados | 13 | 13 | 12 |
| 900 | dados web of | 12 | 12 | 12 |
| 1104 | programas pos-graduacao ciencia | 11 | 9 | 11 |
| 1604 | ciencia tecnologia inovacao | 11 | 10 | 11 |
| 3194 | artigos publicados periodicos | 11 | 11 | 11 |
| 3623 | coordenacao aperfeicoamento pessoal | 11 | 11 | 11 |
| 3624 | aperfeicoamento pessoal superior | 11 | 11 | 11 |
| 4428 | universidade federal rio | 11 | 11 | 11 |
| 5183 | periodicos ciencia informacao | 11 | 9 | 11 |

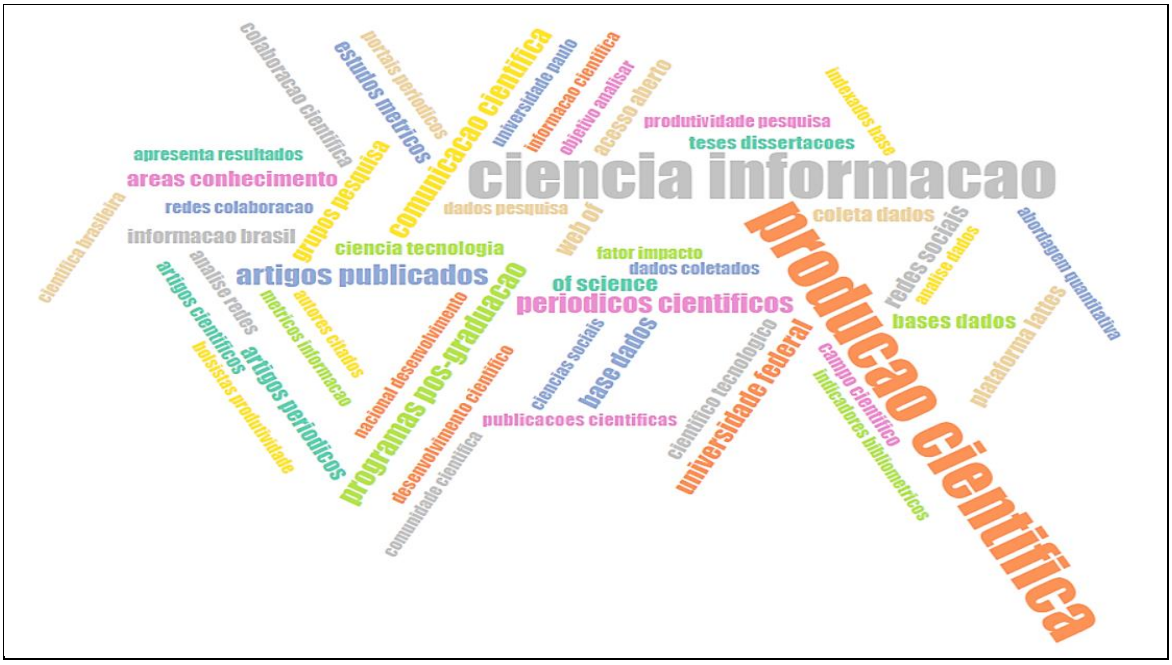
Fonte: Captura de tela da pesquisa (2022).

Salienta-se que, para extração dos bigramas e trigramas, o processo de configuração e ação dos módulos M20 e M21 (ver Figura 32) foram feitos duas vezes. Na primeira vez, configurou-se o módulo *NGram Creator* (M20) para encontrar os bigramas e logo após esse procedimento, executou-se a ação no módulo *Tag Cloud*, para gerar a imagem com a nuvem de palavras dos bigramas (Figura 35). Na sequência, retorna-se ao módulo 20, faz-se a alteração da quantidade de termos para 3, utilizando-se o componente *Configure* no menu de opções (conforme demonstrado

[43] É possível relembrar essa condição retornando a leitura nas páginas 109 e 110.

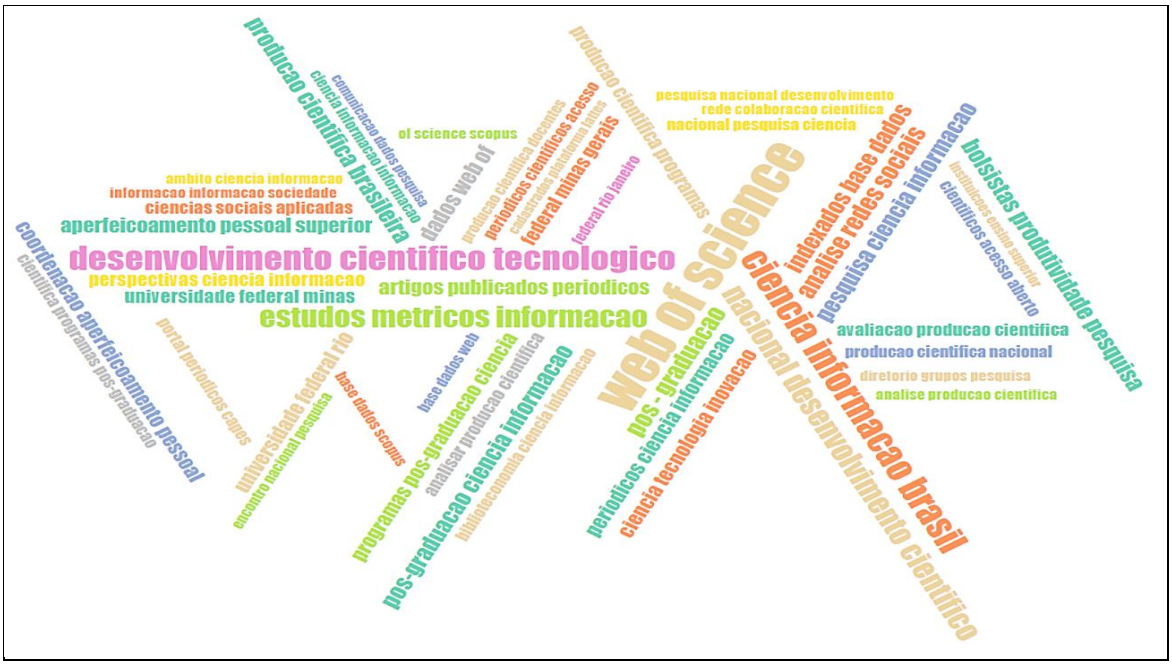
na Figura 6), procede-se o comando de execução da ação (componente *Execute*), e novamente utiliza-se o módulo *Tag Cloud* (M21), o qual gera a nuvem de palavras com os trigramas (Figura 36).

Figura 35 – Nuvem de palavras dos bigramas




Fonte: Dados da pesquisa (2023).

Figura 36 – Nuvem de palavras das trigramas




Fonte: Dados da pesquisa (2023).

Entretanto, diante da dificuldade de visualização completa de forma impressa dos dois resultados minerados (bigramas e trigramas), nas Figuras 35 e 36 são

apresentados somente os 50 n-gramas mais recorrentes no *corpus*. A representação dos dados nessas figuras também é feita pelo módulo *Tag Cloud;* e o posicionamento referente ao ranque é feito pelo tamanho das letras dos n-gramas, de forma que, quanto mais coocorrências dentro do *corpus*, maior os n-grama serão apresentados nas figuras. As Figuras 35 e 36 são obtidas acessando o componente *Interactive View,* presente no menu de opções (ver Figura 6) do módulo *Tag Cloud* (M21).

### 5.4.4 Aplicação da técnica de categorização temática

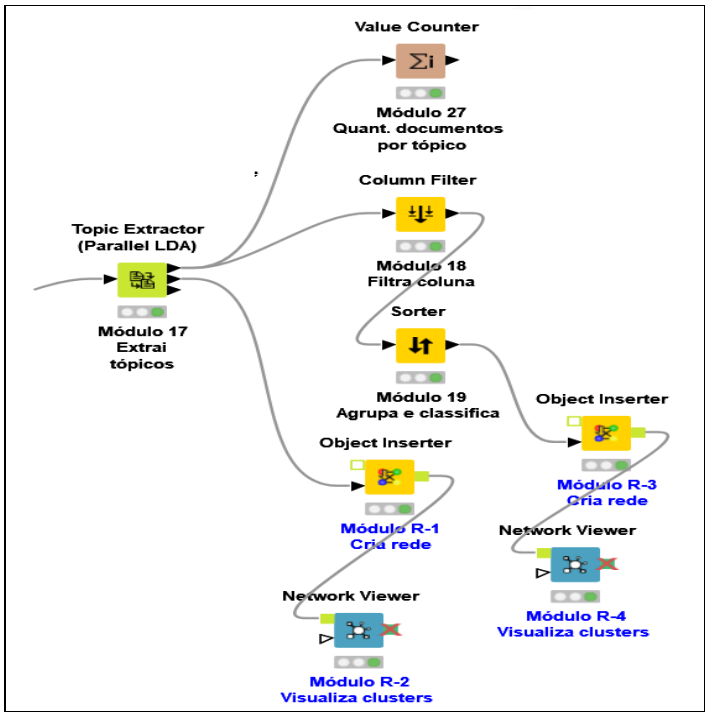
Nessa etapa do processo de mineração, utilizou-se a técnica de análise por categorização temática, que tem por finalidade encontrar os termos que refletem os principais tópicos de um conjunto de documentos. Por meio dessa técnica, tem-se a possibilidade de descobrir os possíveis vínculos entre os resumos, correlacioná-los e agrupá-los por categorias temáticas. Para execução desse processo (minerar os textos; levantamento do conjunto de termos que caracterizarão um tema; classificar e agrupar os resumos em relação aos temas descobertos), fez-se uso de um algoritmo de modelagem de tópicos não supervisionado, denominado de algoritmo de *Alocação Latente de Dirichlet* (*Latent Dirichlet Allocation - LDA*).

Segundo Blei, Ng e Jordan (2003), ao contrário de alguns algoritmos de armazenamento de *cluster* que executam agrupamentos físicos (onde os tópicos são desarticulados), o LDA atribui cada documento a uma mistura de tópicos e os tópicos como uma mistura de termos. (LU; WOLFRAN, 2012). Isso significa que em cada arquivo ou texto (nesta pesquisa, os resumos das comunicações científicas) o algoritmo identifica um ou mais tópicos e atribui um percentual com a probabilidade de pertencimento ou não a determinado tópico, refletindo resultados mais realistas e enriquecendo a pesquisa. O LDA pode buscar os termos de duas maneiras e, nesta tese, utilizou-se o método no qual o usuário define a quantidade de tópicos que se deseja encontrar, por ser considerado um método de fácil entendimento, de configuração simples e que não compromete os resultados.

A Figura 37 demonstra os módulos *Topic Extractor – Parallel LDA* (M17 – Módulo 17), além dos outros nós utilizados para apresentação dos dados extraídos. O módulo *Column Filter* (M18 – Módulo 18) novamente foi utilizado apenas para filtrar as colunas pertinentes; e o módulo *Sorter* (M19 – Módulo 19) serve para classificar as tabelas em ordem crescente ou decrescente nas colunas, conforme designação do

usuário nas configurações. O nó *Value Counter* (M27 – Módulo 27) faz a soma da quantidade de documentos vinculados aos tópicos extraídos.

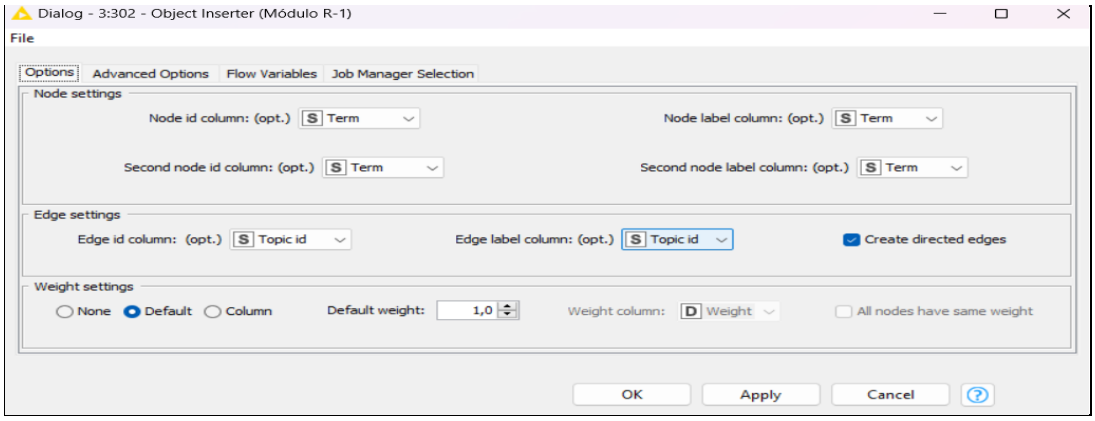
Figura 37 – Algoritmo LDA e módulos de apresentação de resultados




Fonte: Captura de tela da pesquisa (2023).

É possível ainda notar, na Figura 37, que existem dois módulos *Object Inserter* (Módulo R-1 e R-3) e mais dois módulos *Network Viewer* (Módulo R-2 e R-4), que estão com suas descrições na cor azul. Esses módulos são para criação e apresentação de *clusters* ou redes de relacionamentos. Nos módulos R-1 e R-3, são feitas as inserções dos dados na rede, definindo quais são os atores (nós) e por qual atributo eles estão ligados, conforme demonstrado na tela de configuração do nó, apresentada na Figura 38.

Figura 38 – Tela de configuração do módulo *Object Inserter*




Fonte: Captura de tela da pesquisa (2023).

Os módulos R-2 e R-4 servem para apresentar as relações dentro das redes ou os conjuntos de atores que fazem parte de determinado agrupamento. Como não está no escopo dessa pesquisa a Análise de Redes Sociais (ARS), os módulos de visualização de redes serão utilizados apenas para demonstrar os *clusters* de tópicos e conjunto de resumos (atores), que foram classificados conforme as temáticas encontradas (atributo de ligação).

Destaque-se ainda que, o módulo *Topic Extractor* (M17) possui 3 saídas de dados (verificar Figura 37), mas neste estudo, serão utilizas apenas as duas primeiras. A saída de dados 1 apresenta a classificação dos resumos conforme a temática e na saída de dados 2 estão listados os tópicos que definem os temas dos resumos (comunicações científicas). Nesta pesquisa definiu-se que o algoritmo LDA deveria buscar por 5 tópicos e que cada tópico seria representado por um conjunto de 10 termos.

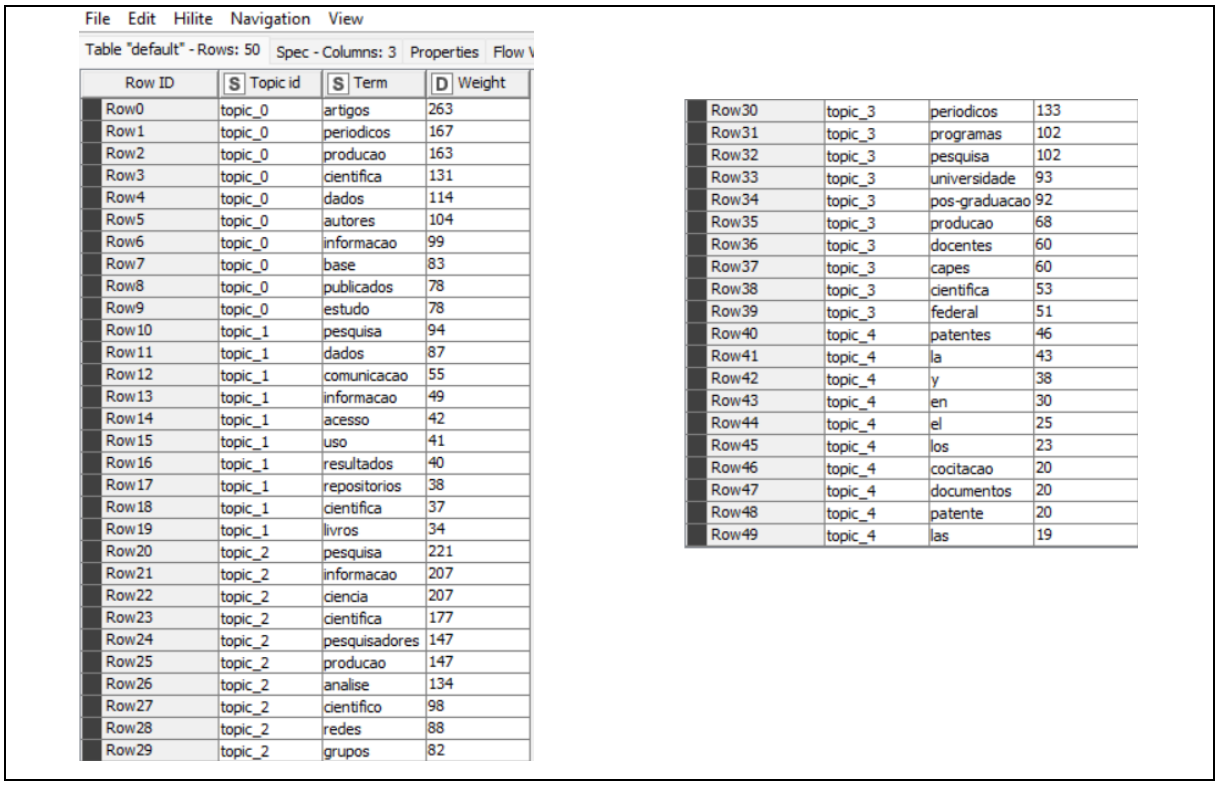
As temáticas obtidas com tópicos extraídos são demonstradas na Figura 39, que corresponde aos resultados apresentados pela saída de dados 2 do *Topic Extractor* (M17). Embora existam metodologias para calcular uma quantidade de tópicos que seja representativa em relação ao número de documentos, por não fazer parte do escopo desta pesquisa, esse tema não será debatido. Contudo, como já exposto, a metodologia aqui proposta permite ao pesquisador definir as quantidades de tópicos e dos termos que compõem cada tópico, de acordo com necessidade do estudo.

Conforme pode-se observar na Figura 39, estão apresentados os 5 tópicos extraídos compostos pelos 10 termos. Cada tópico está organizado na tabela em ordem decrescente, em conformidade aos pesos atribuídos à cada termo dentro da temática; e estão representados da seguinte forma: Tópico 0 (artigos, periódicos, produção, científica, dados, autores, informação, base, publicados e estudo); Tópico 1 (pesquisa, dados comunicação, informação, acesso, uso, resultados, repositórios, científica, livros); Tópico 2 (pesquisa, informação, ciência, científica, pesquisadores, produção, análise, científico, redes e grupos); Tópico 3 (periódicos, programa, pesquisa, universidade, pós-graduação, produção, docentes, capes e científica); Tópico 4 (patentes, la, y, em, el, los, cocitação, documentos, patente e las).

Como já mencionado, a figura demonstra que cada termo que compõe o tópico possui um peso determinado. Por exemplo, no Tópico 0, o termo artigos (263) tem peso bem superior ao termo estudo (98), denotando que o primeiro termo tem maior

importância dentro desse tópico. No Tópico 2, as palavras informação e ciência possuem o mesmo peso (207), estando atrás somente do termo pesquisa (221), demonstrando que elas se equivalem em importância nesse tópico. Outra observação que pode ser feita na Figura 39 é de que a maioria dos termos dos Tópicos 0 e 2 apresentam os maiores pesos, estando mais próximos do maior valor, que é 263, e que o Tópico 4 possui um conjunto de temos com pesos mais baixos.

Figura 39 – Extração de tópicos com o algoritmo LDA.



File Edit Hilite Navigation View

Table "default" - Rows: 50 | Spec - Columns: 3 | Properties | Flow

| Row ID | Topic id | Term | Weight |
|---|---|---|---|
| Row0 | topic_0 | artigos | 263 |
| Row1 | topic_0 | periodicos | 167 |
| Row2 | topic_0 | producao | 163 |
| Row3 | topic_0 | cientifica | 131 |
| Row4 | topic_0 | dados | 114 |
| Row5 | topic_0 | autores | 104 |
| Row6 | topic_0 | informacao | 99 |
| Row7 | topic_0 | base | 83 |
| Row8 | topic_0 | publicados | 78 |
| Row9 | topic_0 | estudo | 78 |
| Row10 | topic_1 | pesquisa | 94 |
| Row11 | topic_1 | dados | 87 |
| Row12 | topic_1 | comunicacao | 55 |
| Row13 | topic_1 | informacao | 49 |
| Row14 | topic_1 | acesso | 42 |
| Row15 | topic_1 | uso | 41 |
| Row16 | topic_1 | resultados | 40 |
| Row17 | topic_1 | repositorios | 38 |
| Row18 | topic_1 | cientifica | 37 |
| Row19 | topic_1 | livros | 34 |
| Row20 | topic_2 | pesquisa | 221 |
| Row21 | topic_2 | informacao | 207 |
| Row22 | topic_2 | ciencia | 207 |
| Row23 | topic_2 | cientifica | 177 |
| Row24 | topic_2 | pesquisadores | 147 |
| Row25 | topic_2 | producao | 147 |
| Row26 | topic_2 | analise | 134 |
| Row27 | topic_2 | cientifico | 98 |
| Row28 | topic_2 | redes | 88 |
| Row29 | topic_2 | grupos | 82 |
| Row30 | topic_3 | periodicos | 133 |
| Row31 | topic_3 | programas | 102 |
| Row32 | topic_3 | pesquisa | 102 |
| Row33 | topic_3 | universidade | 93 |
| Row34 | topic_3 | pos-graduacao | 92 |
| Row35 | topic_3 | producao | 68 |
| Row36 | topic_3 | docentes | 60 |
| Row37 | topic_3 | capes | 60 |
| Row38 | topic_3 | cientifica | 53 |
| Row39 | topic_3 | federal | 51 |
| Row40 | topic_4 | patentes | 46 |
| Row41 | topic_4 | la | 43 |
| Row42 | topic_4 | y | 38 |
| Row43 | topic_4 | en | 30 |
| Row44 | topic_4 | el | 25 |
| Row45 | topic_4 | los | 23 |
| Row46 | topic_4 | cocitacao | 20 |
| Row47 | topic_4 | documentos | 20 |
| Row48 | topic_4 | patente | 20 |
| Row49 | topic_4 | las | 19 |

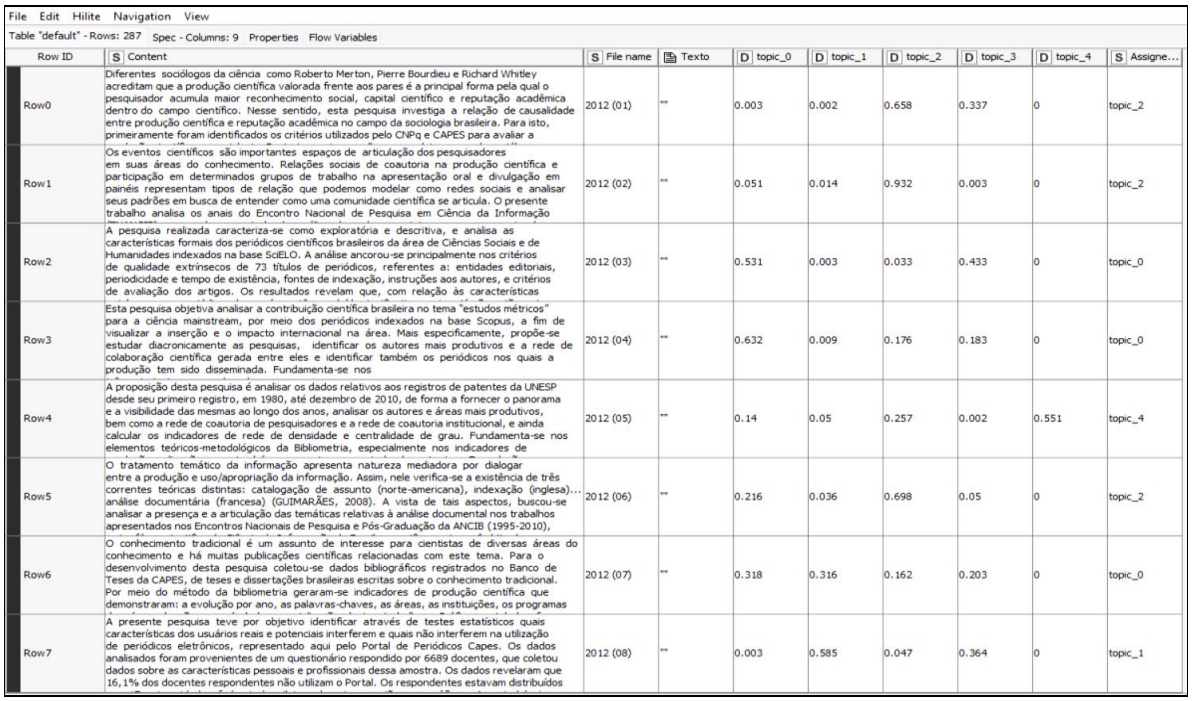
Fonte: Captura de tela da pesquisa (2023).

A explicação pode estar no número de resumos que foram classificados em cada temática, pois, o módulo *Value Counter* (M27), utilizado para verificar a quantidade de documentos existentes em cada tópico, retornou os seguintes quantitativos: Tópico 0 – 83 resumos; Tópico 1 – 57 resumos; Tópico 2 – 80 resumos; Tópico 3 – 53 resumo e Tópico 4 – 14 resumos, totalizando as 287 comunicações científicas utilizadas. Os tópicos 0 e 2 possuem as maiores quantidades de documentos e o tópico 4 forma o menor *cluster*, o que poderia comprovar esse entendimento, pois, se os termos se repetem em mais documentos, maiores pesos (importância) eles terão dentro do *corpus*.

A Figura 40 apresenta o resultado da primeira saída de dados do módulo *Topic Extractor*. São fragmentos dos resultados obtidos com o algoritmo LDA, pois, além de

determinar as temáticas, o algoritmo faz a classificação probabilística dos documentos com base nos tópicos encontrados. É possível ainda perceber na figura que são apresentadas todas as possibilidades de pertencimento aos tópicos e que a ferramenta sempre alocará o documento ao tópico com a maior probabilidade calculada. Para não se estender muito nas comparações, apenas os dois valores probabilísticos mais altos encontrados são delineados em seguida.

Figura 40 – Classificação dos resumos conforme tópicos encontrados



File Edit Hilite Navigation View

Table "default" - Rows: 287 Spec - Columns: 9 Properties Flow Variables

| Row ID | Content | File name | Texto | topic_0 | topic_1 | topic_2 | topic_3 | topic_4 | Assigne... |
|---|---|---|---|---|---|---|---|---|---|
| Row0 | Diferentes sociólogos da ciência como Roberto Merton, Pierre Bourdieu e Richard Whitley acreditam que a produção científica valorada frente aos pares é a principal forma pela qual o pesquisador acumula maior reconhecimento social, capital científico e reputação acadêmica dentro do campo científico. Nesse sentido, esta pesquisa investiga a relação de causalidade entre produção científica e reputação acadêmica no campo da sociologia brasileira. Para isto, primeiramente foram identificados os critérios utilizados pelo CNPq e CAPES para avaliar a | 2012 (01) | "" | 0.003 | 0.002 | 0.658 | 0.337 | 0 | topic_2 |
| Row1 | Os eventos científicos são importantes espaços de articulação dos pesquisadores em suas áreas do conhecimento. Relações sociais de coautoria na produção científica e participação em determinados grupos de trabalho na apresentação oral e divulgação em painéis representam tipos de relação que podemos modelar como redes sociais e analisar seus padrões em busca de entender como uma comunidade científica se articula. O presente trabalho analisa os anais do Encontro Nacional de Pesquisa em Ciência da Informação | 2012 (02) | "" | 0.051 | 0.014 | 0.932 | 0.003 | 0 | topic_2 |
| Row2 | A pesquisa realizada caracteriza-se como exploratória e descritiva, e analisa as características formais dos periódicos científicos brasileiros da área de Ciências Sociais e de Humanidades indexados na base SciELO. A análise ancorou-se principalmente nos critérios de qualidade extrínsecos de 73 títulos de periódicos, referentes a: entidades editoriais, periodicidade e tempo de existência, fontes de indexação, instruções aos autores, e critérios de avaliação dos artigos. Os resultados revelam que, com relação às características | 2012 (03) | "" | 0.531 | 0.003 | 0.033 | 0.433 | 0 | topic_0 |
| Row3 | Esta pesquisa objetiva analisar a contribuição científica brasileira no tema "estudos métricos" para a ciência mainstream, por meio dos periódicos indexados na base Scopus, a fim de visualizar a inserção e o impacto internacional na área. Mais especificamente, propõe-se estudar diacronicamente as pesquisas, identificar os autores mais produtivos e a rede de colaboração científica gerada entre eles e identificar também os periódicos nos quais a produção tem sido disseminada. Fundamenta-se nos | 2012 (04) | "" | 0.632 | 0.009 | 0.176 | 0.183 | 0 | topic_0 |
| Row4 | A proposição desta pesquisa é analisar os dados relativos aos registros de patentes da UNESP desde seu primeiro registro, em 1980, até dezembro de 2010, de forma a fornecer o panorama e a visibilidade das mesmas ao longo dos anos, analisar os autores e áreas mais produtivos, bem como a rede de coautoria de pesquisadores e a rede de coautoria institucional, e ainda calcular os indicadores de rede de densidade e centralidade de grau. Fundamenta-se nos elementos teóricos-metodológicos da Bibliometria, especialmente nos indicadores de | 2012 (05) | "" | 0.14 | 0.05 | 0.257 | 0.002 | 0.551 | topic_4 |
| Row5 | O tratamento temático da informação apresenta natureza mediadora por dialogar entre a produção e uso/apropriação da informação. Assim, nele verifica-se a existência de três correntes teóricas distintas: catalogação de assunto (norte-americana), indexação (inglesa)... análise documentária (francesa) (GUIMARÃES, 2008). A vista de tais aspectos, buscou-se analisar a presença e a articulação das temáticas relativas à análise documental nos trabalhos apresentados nos Encontros Nacionais de Pesquisa e Pós-Graduação da ANCIB (1995-2010), | 2012 (06) | "" | 0.216 | 0.036 | 0.698 | 0.05 | 0 | topic_2 |
| Row6 | O conhecimento tradicional é um assunto de interesse para cientistas de diversas áreas do conhecimento e há muitas publicações científicas relacionadas com este tema. Para o desenvolvimento desta pesquisa coletou-se dados bibliográficos registrados no Banco de Teses da CAPES, de teses e dissertações brasileiras escritas sobre o conhecimento tradicional. Por meio do método da bibliometria geraram-se indicadores de produção científica que demonstraram: a evolução por ano, as palavras-chaves, as áreas, as instituições, os programas | 2012 (07) | "" | 0.318 | 0.316 | 0.162 | 0.203 | 0 | topic_0 |
| Row7 | A presente pesquisa teve por objetivo identificar através de testes estatísticos quais características dos usuários reais e potenciais interferem e quais não interferem na utilização de periódicos eletrônicos, representado aqui pelo Portal de Periódicos Capes. Os dados analisados foram provenientes de um questionário respondido por 6689 docentes, que coletou dados sobre as características pessoais e profissionais dessa amostra. Os dados revelaram que 16,1% dos docentes respondentes não utilizam o Portal. Os respondentes estavam distribuídos | 2012 (08) | "" | 0.003 | 0.585 | 0.047 | 0.364 | 0 | topic_1 |

Fonte: Captura de tela da pesquisa (2023).

Isso posto, ao observar a Figura 40, constata-se na linha 1 (Row 0), que o resumo do trabalho 2012 (01) tem 0.658 de chances de pertencer ao Tópico 2, e 0.337 de chances de pertencer ao tópico 3, por isso ele foi classificado como pertencente ao tópico 2. Na linha 2 da mesma tabela (Row 1), o resumo com identificação 2012 (02) tem 0.932 de probabilidade de pertencer ao tópico 2 e 0.051 de chances de pertencer ao tópico 0, sendo classificado também como pertencente ao tópico 2. A linha 7 (Row 6) armazena os dados do resumo do arquivo 2012 (07) e apresenta 0.318 de chances para o tópico 0 e 0.316 para o tópico 1, o que denota que esse resumo possui quantidades próximas de termos nos dois tópicos. Os cálculos demonstram que o resumo foi classificado como pertencente ao tópico 0, mas significa que seu conteúdo possui bastante relação com a temática presente no tópico 1.

Como não está no mérito desta pesquisa fazer a análise das probabilidades ou classificações, tais explanações serviram para demonstrar o comportamento da

mineração de texto para o levantamento de termos com uso da técnica de categorização temática. Outrossim, servem para mostrar as possibilidades de aplicação da metodologia em outros cenários ou aplicações. A Figura 41 mostra uma parte da saída de dados do módulo *Sorter* (M19), o qual agrupou os resumos, ordenou-os primeiramente do menor tópico para o maior, e no segundo nível organizou os resumos por nome de arquivo (crescente), facilitando a visualização.

A parte a que se refere a Figura 41 corresponde aos 14 resumos que foram classificados como pertencentes ao Tópico 4, e essas linhas são as últimas linhas da tabela completa, pois é o último tópico. Foram feitas algumas configurações, como aumento do tamanho da letra e diminuição da altura das linhas, apenas para tornar a visualização dos dados um pouco mais cômoda. Ao retomar os termos que compõem esse tópico (patentes, la, y, em, el, los, cocitação, documentos, patente e las), algumas impressões podem ficar expressas.

Figura 41 – Resumos classificados como pertencentes ao tópico 4

| | | | |
|---|---|---|---|
| Row4 | A proposição desta pesquisa é analisar os dados relativos aos registros de patentes da U... desde seu primeiro registro, em 1980, até dezembro de 2010, de forma a fornecer o pa... e a visibilidade das mesmas ao longo dos anos, analisar os autores e áreas mais produtiv... | 2012 (05) | topic_4 |
| Row29 | A carência de indicadores e a formação precária de pareceristas são dois dentre ... problemas da revisão por pares. Esta pesquisa aborda ambos, ao medir a confiabili... pareceres dos pares sobre propostas de mestrado em Ciência da Informação. Conduzim... | 2012 (30) | topic_4 |
| Row65 | Analiza el crecimiento de la literatura sobre bibliometría publicada en el Brasil por ... brasileños o extranjeros en la forma de artículos de revistas, capítulos de libros y... presentados en congresos. Analiza el crecimiento de la literatura publicada por trienios, la ... | 2013 (8) | topic_4 |
| Row72 | Analiza el número de publicaciones, el idioma, la tipología documental, la colaboración y el crecimiento de las publicaciones sobre | 2014 (14) | topic_4 |
| Row78 | Dentro del complejo y multidimensional campo de la Educación Superior uno de los temas que ha venido alcanzando destaque es su dimensión | 2014 (2) | topic_4 |
| Row120 | Este artículo asume la perspectiva de los estudios postcoloniales con una crítica al pensamiento hegemónico que sostiene que los saberes producidos por Europa Occident... Unidos) son los únicos que pueden ser considerados universales, valiosos y que constitu... | 2015 (2) | topic_4 |
| Row148 | Esta pesquisa identifica e analisa os documentos de patente correlatos às tecnologias de Gerenciamento de resíduos, depositados na base de patentes do Instituto Nacional da P... Industrial no período de 1/1/2005 a 31/12/2013, pelas 21 universidades públicas brasileir... | 2016 (19) | topic_4 |
| Row166 | Trata-se de um estudo cientométrico que se utiliza de indicadores de patentes com o int... perceber as características da produção tecnológica de biodiesel indexada na Derwent In... entre 2008 e 2009. Através de uma metodologia exploratória e descritiva, esta primeira ... | 2016 (35) | topic_4 |
| Row184 | A revisão por pares é frequentemente afetada por diversas falhas que podem comprom... sua integridade, levando eventualmente à retratação de publicações. Compreender... ajudar a minimizá-las; sendo assim, a questão de pesquisa é: quais os causadore... | 2016 (8) | topic_4 |
| Row191 | Trata-se de um estudo patentométrico que se utiliza de indicadores de patentes com o o... de identificar os aspectos da produção tecnológica brasileira indexada na Derwent... entre 2004 e 2016. Utiliza de metodologia exploratória e descritiva, e observa aspectos r... | 2017 (14) | topic_4 |
| Row223 | Identifica e analisa os documentos de patente relacionados às tecnologias verdes depositados por 21 universidades públicas brasileiras no Instituto Nacional da Propriedad... (INPI), entre 1 de janeiro de 2005 a 31 de dezembro de 2014, totalizando 294 docume... | 2017 (43) | topic_4 |
| Row231 | O presente trabalho tem como tema a Análise de cocitação de autores – ACA. Para tant... objetivos são: 1) determinar a proximidade por seção do artigo e por parágrafo de um ... cocitados de uma ACA tradicional; e 2) identificar indícios de uma subtipologia de cocitac... | 2017 (50) | topic_4 |
| Row237 | Analiza las palabras clave que aparecen en los documentos publicados en Colombia o en otros países sobre bibliometría, cienciometría, informetría y otros términos asociad... palabras que fueron publicados por autores afiliados a instituciones colombianas. S... | 2017 (56) | topic_4 |
| Row242 | Análise patentométrica com o objetivo de compreender o depósito de patentes brasileiras com extensões de depósitos via Tratado de Cooperação de Patentes (PCT). Co... importância do documento de patente tanto em termos informacionais como den... | 2018 (1) | topic_4 |

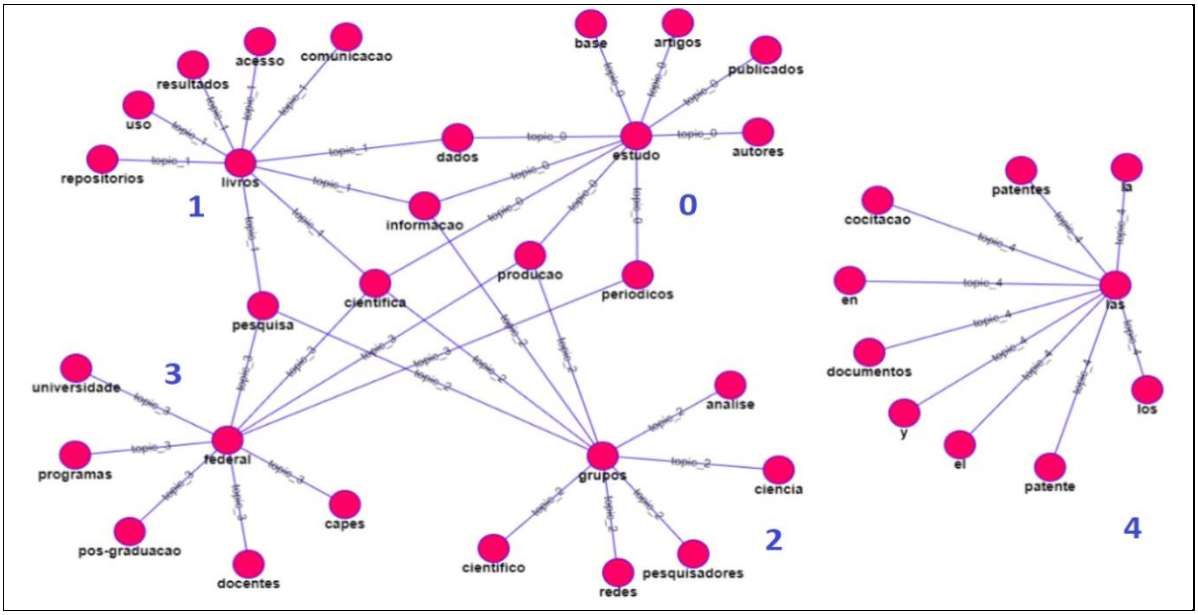
Fonte: Captura de tela da pesquisa (2023).

Ao observarmos os termos do Tópico 4 evidencia-se que o algoritmo LDA encontrou resumos escritos na língua espanhola, o que não causaria estranheza, pois tem-se conhecimento que os ENANCIBs aceitam trabalhos escritos nessa língua. Outra evidência que a Figura 41 aponta é a de que, nesse conjunto de 14 resumos existem estudos que debatem sobre concessões públicas de direito sobre invenções, pois os termos patente e patentes também fazem parte do tópico descoberto, inclusive

a palavra patente é a de maior peso na classificação do agrupamento. Saliente-se que tais exposições foram feitas apenas para melhor entendimento da metodologia proposta nesta tese, de modo que não se tem como propósito a análise dos dados.

Contudo, aqui aparece uma das limitações do trabalho, pois, como as técnicas empregadas provém da área de Linguística e os algoritmos se baseiam na definição do idioma dos documentos, os termos se diferenciam e principalmente são dependentes da língua. Um exemplo seria a questão das *Stop Words* (palavras que podem ser totalmente suprimida, omitida ou ocultada na hora de fazer uma busca, sem que o sentido do que se quer encontrar seja perdido), pois, as palavras consideradas sem importância em uma língua não são as mesmas em outra, e o mesmo ocorre na extração dos termos.

Figura 42 – *Cluster* dos tópicos obtidos com o algoritmo LDA.



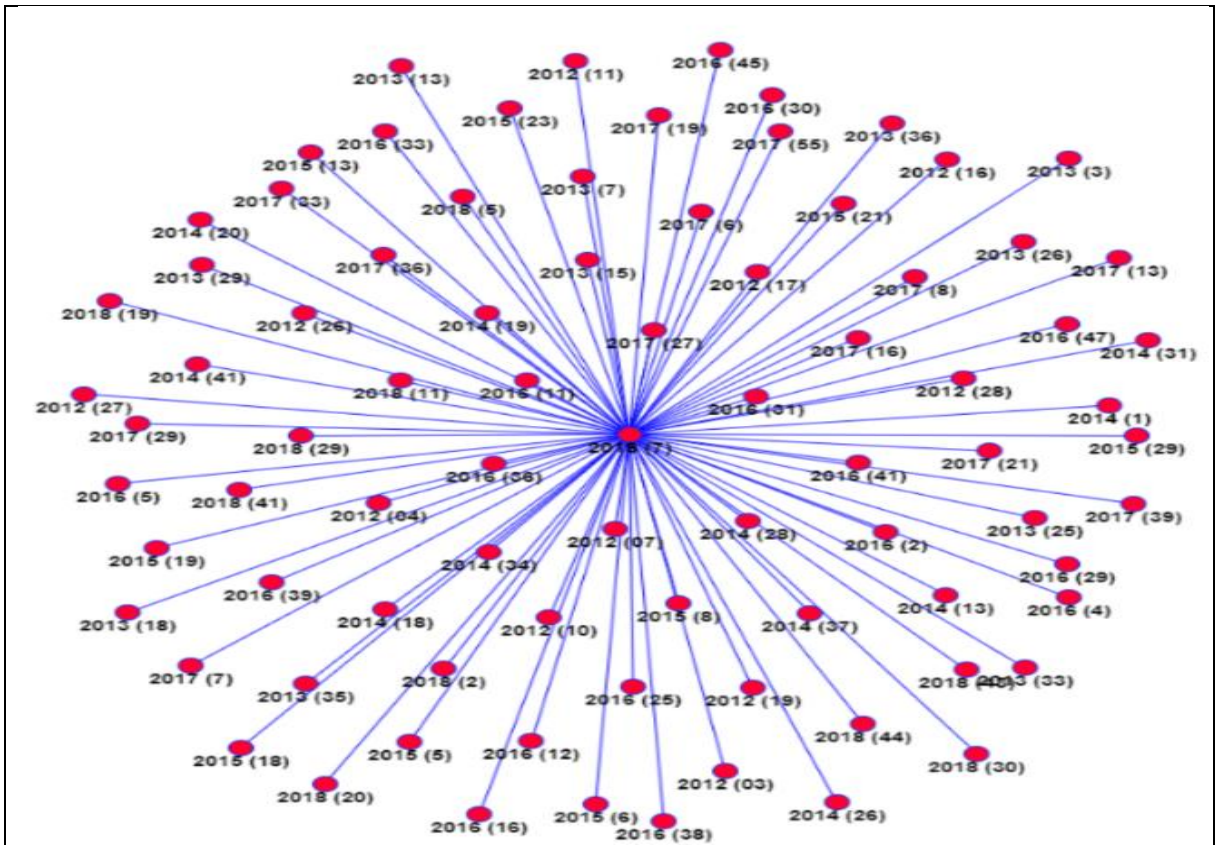

Fonte: Dados da pesquisa (2023).

Como os módulos *Object Inserter* (módulo R-1 e R-3), apresentados na Figura 37, servem apenas para criação de uma rede com os nós e seus atributos, eles não possuem nenhuma saída de dados importante e, por isso, não serão apresentados. Contudo, os nós seguintes, que são de visualização de dados *Network Viewer* (módulo R-2 e R-4), só são passíveis de utilização se a rede for criada, por isso os módulos R-1 e R-3 não devem ser descartados. A Figura 42 apresenta a saída de dados do módulo *Network Viewer* (MR-2), com os *clusters* dos termos que compõem

os tópicos encontrados. Embora demonstrados como rede ego[44], as figuras não são ligações de rede social, apenas mostram os grupos dos termos e seus resumos em que temáticas estão presentes.

Pode-se verificar, na figura, que o conjunto de termos do Tópico 4 não tem ligação com nenhum outo tópico, pois não possui nenhuma palavra em outro grupo de representação temática. Observa se também que o termo “dados” pertence aos Tópicos 0 e 1; que o termo “pesquisa” está contido no Tópico 3, 2 e 1; o termo “informação” se faz presente nos *clusters* dos Tópicos 0, 1 e 2; o termo “produção” pertence aos Tópicos 0, 2 e 3, e o termo “científica” está incluso nos Tópicos 0, 1, 2 e 3. As expressões pertencentes a mais de um tópico formam um grupo composto pelos seguintes termos: científica, produção, informação, pesquisa e dados. As próximas figuras (43, 44, 45, 46 e 47) foram geradas pelo módulo *Network Viewer* (MR-4) e apresentam os *clusters* dos resumos, agrupados segundo as classificações temáticas dos tópicos extraídos.

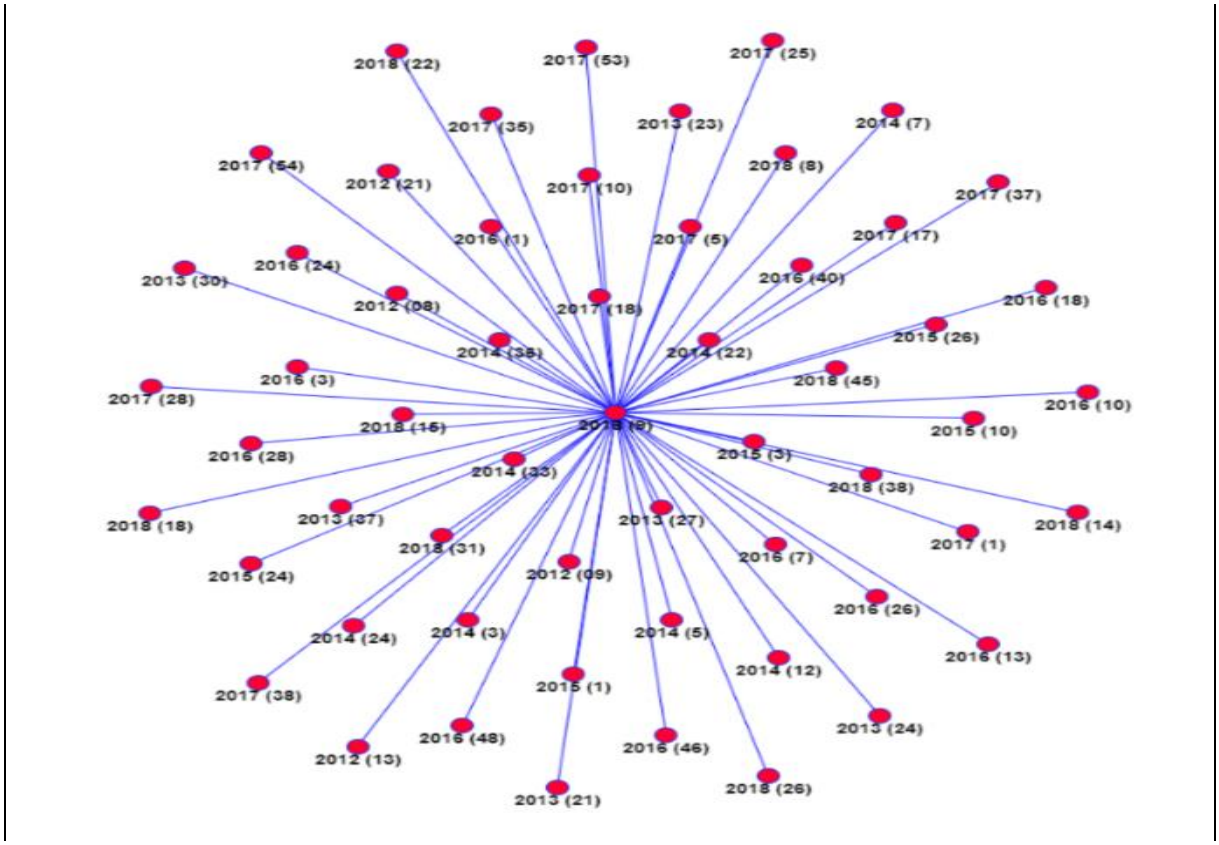
Figura 43 – *Cluster* dos resumos referentes ao tópico 0




Fonte: Dados da pesquisa (2023).

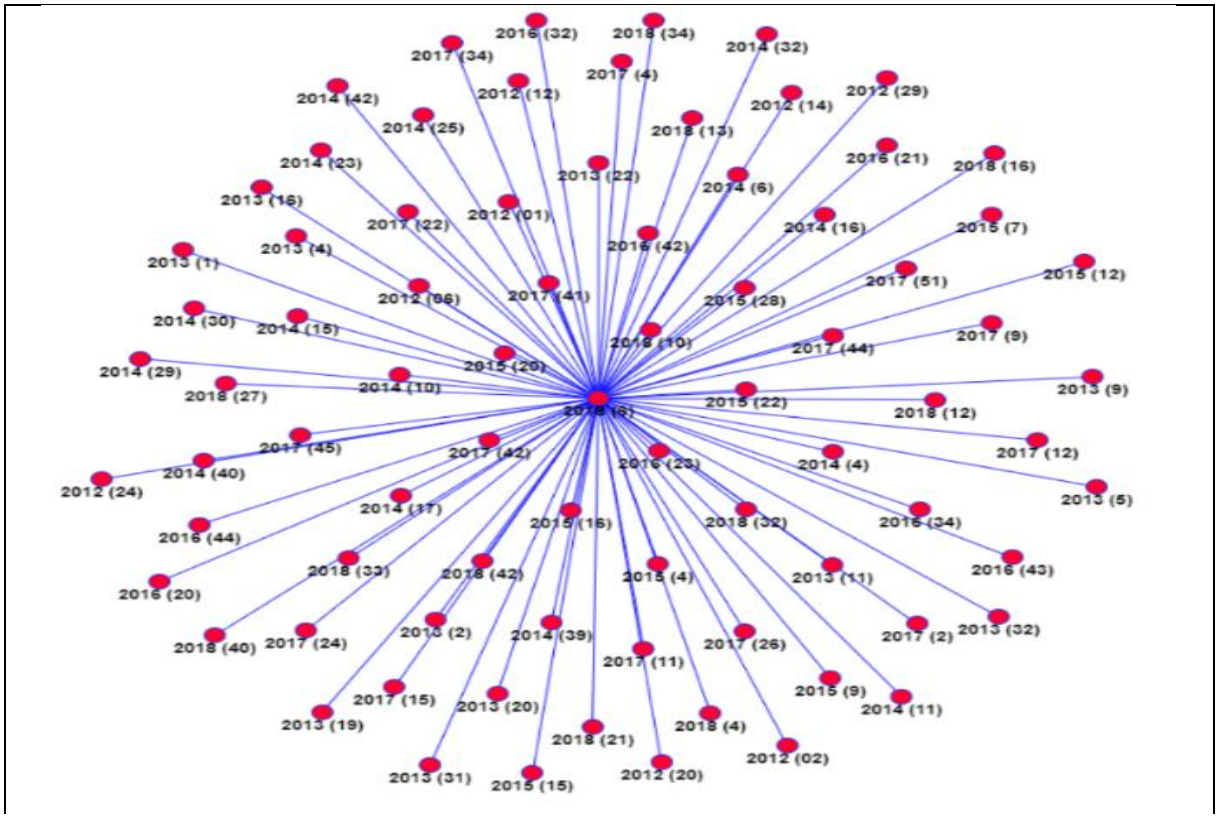
[44] Expressão usada em estudos de Análise de Redes Sociais (ARS).

Figura 44 – *Cluster* dos resumos referentes ao tópico 1



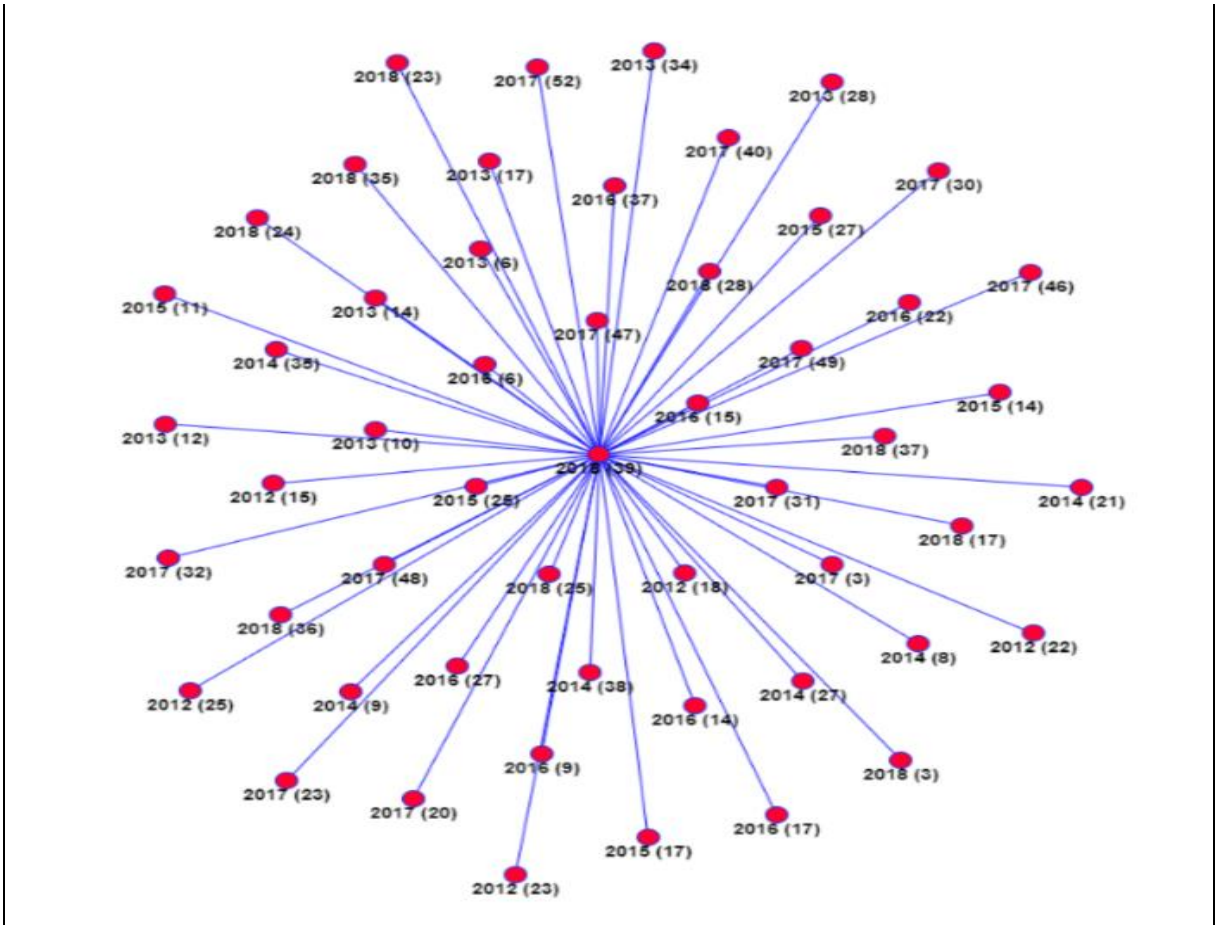

Fonte: Dados da pesquisa (2023).

Figura 45 – *Cluster* dos resumos referentes ao tópico 2



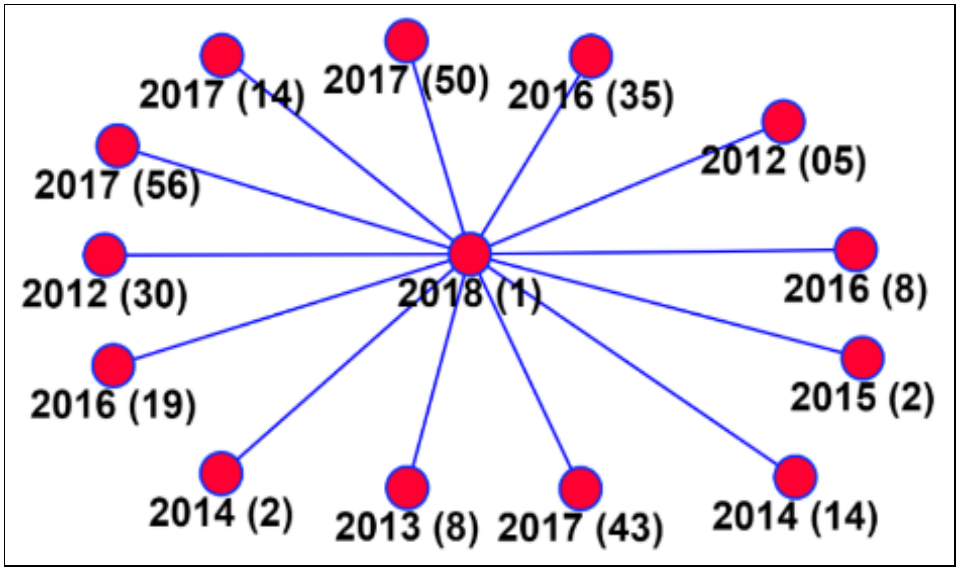

Fonte: Dados da pesquisa (2023).

Figura 46 – *Cluster* dos resumos referentes ao tópico 3




Fonte: Dados da pesquisa (2023).

Figura 47 – *Cluster* dos resumos referentes ao tópico 4




Fonte: Dados da pesquisa (2023).

A modelagem de tópicos é bastante adequada para classificar textos, criar sistemas de recomendação (por exemplo, recomendar livros com base em suas

leituras anteriores), apresentar preferências, sugerir complementos como “próximas palavras” em digitações (sugestão de palavras ao se digitar uma mensagem em aplicativos de celular ou e-mail, por exemplo), direcionar usuários a “produtos” conforme perfil de utilização e até mesmo identificar tendências em futuras publicações *online*.

### 5.4.5 Realização do processo de *stemming*

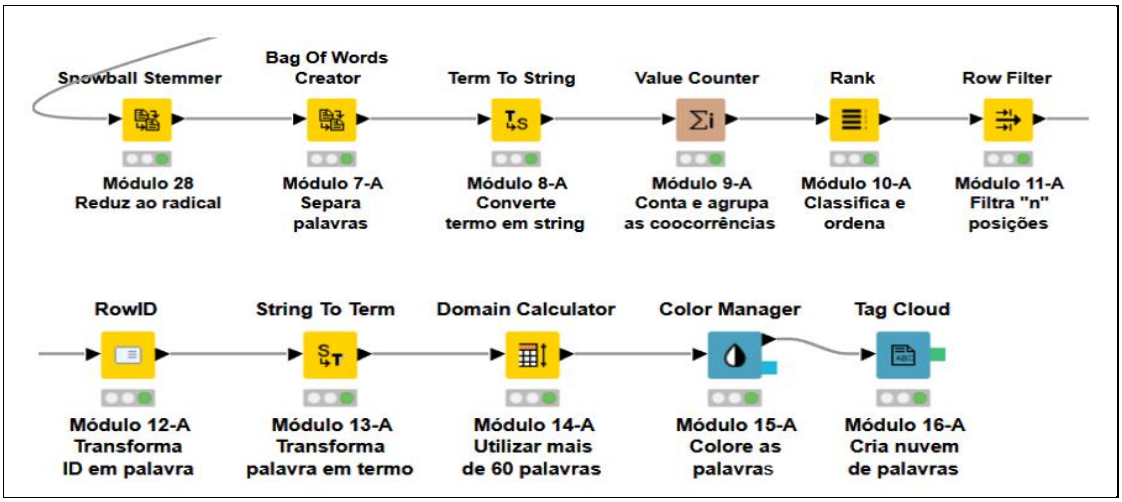
Por fim, a ferramenta de mineração utilizada permite que se realize o procedimento denominado de *stem* (raiz), que reduz os termos encontrados à sua forma primária ou ao seu radical. Segundo Porter (1980), um subconjunto de termos considerados gramaticalmente próximos são abreviados para uma forma canônica inicial, que seja comum a todos os termos considerados semelhantes e que possa representar o conjunto dessas expressões. Os termos são agrupados na sua forma raiz, pois tem-se por pressuposto que as palavras com radical igual são também parecidas no significado, com variações de prefixo e sufixo apenas.

Na Figura 48, são apresentados o conjunto de módulos responsáveis pela realização do processo de *stemming*, manipulação e apresentação dos dados. Para melhor visualização, a figura foi dividida em duas e posicionada uma abaixo da outra, entretanto, no fluxograma principal, o conjunto de módulos aparece em sequência linear. Embora no Knime existam outros algoritmos que executem a tarefa de radicalização dos termos, nesta pesquisa, optou-se por utilizar o módulo *Snowball Stemmer* (M28 – Módulo 28). Assim como outros processos (LDA) têm como base a Linguística, o *stemming* também é dependente do idioma e, pela característica de complexidade da língua portuguesa, poucos algoritmos possuem essa opção. Por isso, a escolha do *Snowball Stemmer*, pois, até o momento, é o único algoritmo que realiza o processo na língua portuguesa (a opção da língua para execução do *stemming* é configurável no módulo).

Saliente-se que a maioria dos algoritmos de *steeming* derivam do algoritmo mais conhecido e mais utilizado na mineração de texto, que é o de Porter (1980). O encadeamento dos módulos de 7 a 16 já foram apresentados na Figura 28, e foram reaproveitados, pois pretende-se demonstrar os dados de forma semelhante, já que são novamente apresentados termos extraídos do *corpus*, mas reduzidos ao seu radical. Conforme demonstrado na Figura 48, os módulos receberam o adendo A,

ficando nomeados como Módulo 7-A, Módulo 8-A, e assim consecutivamente, até o Módulo 16-A, o que na prática não muda a essência de suas tarefas internas.

Figura 48 – Módulos de execução e apresentação do *stemming*




Fonte: Captura de tela da pesquisa (2023).

Os módulos de M7-A ao M16-A são utilizados novamente para aplicação da técnica de dedução de frequência e coocorrência de termos, do mesmo modo que foi executado anteriormente pela sequência de módulos de M7 a M16. Contudo, executou-se apenas a dedução de frequência de termos (unigramas ou termos simples), pois tem-se intenção de apenas demonstrar o resultado do processo de *stemming* e evidenciar que, em circunstâncias específicas particulares à cada pesquisa, o *stemming* pode ser empregue e tornar-se útil. Portanto, não foram realizados procedimentos referentes à coocorrência de termos (termos compostos como bigramas e trigramas).

Depois de se executar o módulo *Snowball Stemmer* (M28 – Módulo 28), o qual realiza a redução das expressões ao seu radical, o módulo *Bag Of Words Creator* (M7-A – Módulo 7-A) destaca cada palavra em separado e as armazena uma por uma em linhas distintas. Sem ainda calcular a frequência das palavras, foram encontrados 21.705 termos reduzidos ao seu modo raiz. O módulo *Term To String* (M8-A – Módulo 8-A) apenas deixa os termos no formato *string* para utilização em nós seguintes.

O módulo *Value Counter* (M9-A – Módulo 9-A) faz a contagem de frequência dos temos no *corpus*, e o módulo *Rank* (M10-A – Módulo 10-A) faz a ordenação dos termos encontrados. O conjunto dos módulos 11-A, 12-A, 13-A, 14-A e 15- tem a função de manipulação dos dados para apresentação. No módulo *Row Filter* (M11-A – Módulo 11-A), define-se a quantidade de linhas (termos) que se pretende

apresentar. Como o M9-A transformou os termos em identificador de coluna, o módulo *RowID* (M12-A – Módulo12-A) transforma-os novamente em um dado manipulável.

O módulo *String To Term* (M13-A – Módulo 13-A) transforma o dado para tipo *term*, e o *Domain Calculator* (M14-A – Módulo 14-A) é utilizado quando se tem intenção de apresentar mais de 60 palavras, já que o Knime limita a quantidade a esse número. Novamente, o *Color Manager* (M15-A – Módulo 15-A) determina cores aleatórias aos termos, para que possam ser diferenciados na demonstração. Por fim, o módulo *Tag Cloud* (M16-A - Módulo 16-A) gera a nuvem de palavras com os dados que foram manipulados pelos nós anteriores.

No primeiro processo de dedução de frequência, foram encontrados 5.820[45] termos. A realização do mesmo processo com o termos manipulados pelo módulo de *stemming* retornou um total de 3.657 expressões, portanto, com a radicalização das expressões, tem-se uma diferença de 2.163 termos a menos. Na Figura 49, são demonstrados os dados após a aplicação do módulo *Snowball Stemmer* (M28), e depois de efetuadas as manipulações entre os módulos M7-A e M14-A (módulos com ações visando a exibição dos resultados).

Figura 49 – Frequência e ranque após o *stemming*

Output Data - 3:316 - Domain Calculator (Módulo 14-A)

File Edit Hilite Navigation View

Table "default" - Rows: 100 Spec - Columns: 4 Properties Flow Variables

| Row ID | count | rank | result | Term |
|---|---|---|---|---|
| pesquis | 234 | 1 | pesquis | pesquis[] |
| analis | 232 | 2 | analis | analis[] |
| cientif | 220 | 3 | cientif | cientif[] |
| result | 173 | 4 | result | result[] |
| estud | 170 | 5 | estud | estud[] |
| cienc | 169 | 6 | cienc | cienc[] |
| period | 168 | 7 | period | period[] |
| produca | 166 | 8 | produca | produca[] |
| objet | 164 | 9 | objet | objet[] |
| dad | 157 | 10 | dad | dad[] |
| artig | 146 | 11 | artig | artig[] |
| apresent | 142 | 12 | apresent | apresent[] |
| informaca | 142 | 13 | informaca | informaca[] |
| public | 123 | 14 | public | public[] |
| bas | 122 | 15 | bas | bas[] |
| identific | 115 | 16 | identific | identific[] |
| indic | 113 | 17 | indic | indic[] |
| utiliz | 113 | 18 | utiliz | utiliz[] |
| conhec | 104 | 19 | conhec | conhec[] |
| brasileir | 95 | 20 | brasileir | brasileir[] |
| univers | 90 | 21 | univers | univers[] |
| autor | 89 | 22 | autor | autor[] |
| brasil | 89 | 23 | brasil | brasil[] |
| desenvolv | 87 | 24 | desenvolv | desenvolv[] |
| bibliometr | 81 | 25 | bibliometr | bibliometr[] |
| are | 80 | 26 | are | are[] |
| tecnolog | 71 | 27 | tecnolog | tecnolog[] |
| camp | 69 | 28 | camp | camp[] |
| colet | 69 | 29 | colet | colet[] |
| realiz | 69 | 30 | realiz | realiz[] |
| red | 67 | 31 | red | red[] |
| metodolog | 66 | 32 | metodolog | metodolog[] |
| temat | 66 | 33 | temat | temat[] |
| comunicaca | 63 | 34 | comunicaca | comunicaca[] |
| publicaco | 62 | 35 | publicaco | publicaco[] |
| pais | 60 | 36 | pais | pais[] |
| instituico | 59 | 37 | instituico | instituico[] |
| present | 59 | 38 | present | present[] |
| princip | 59 | 39 | princip | princip[] |
| soc | 59 | 40 | soc | soc[] |
| especif | 58 | 41 | especif | especif[] |

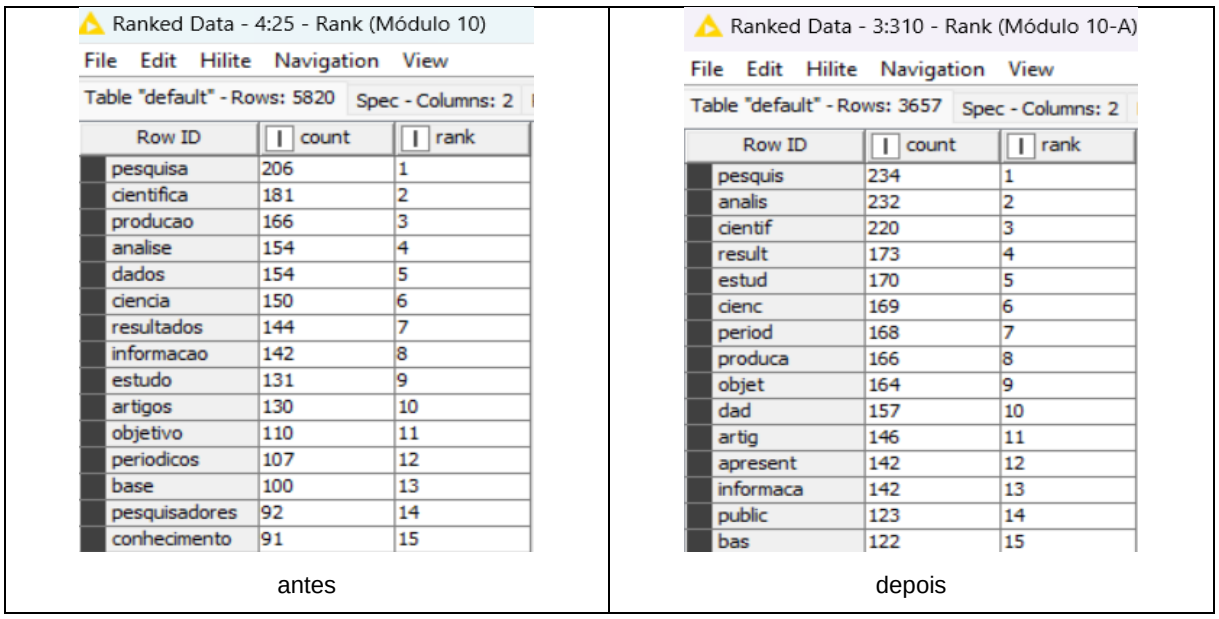
Fonte: Captura de tela da pesquisa (2023).

[45] Ver final da página 106 e início da 107.

Se reportarmos a Figura 30[46], que apresentou os termos, suas frequências e posições no ranque, respectivamente, é possível verificar os seguintes dados: pesquisas – 206 (1); científica – 181 (2); produção – 166 (3); análise – 154 – 4; dados – 154 – 5; ciência – 150 – 6. Ao compararmos a Figura 49 com esses dados da Figura 30, pode-se verificar que a posição 1 do ranque é o radical “pesquis”, que obteve frequência no *corpus* igual a 234. Isso se deu pelo fato de que, além das 206 palavras iniciais, no contexto geral dos resumos, o algoritmo incorporou mais 28 expressões com esse radical (raiz), os quais, após as análises matemáticas, foram entendidas e classificadas como palavras de mesmo significado, totalizando desse modo os 234 radicais (pesquis) encontrados nessa fase.

Ainda comparando os dados constantes nas Figuras 30 e 49, o radical da palavra análise, que teve 154 ocorrências no primeiro processo de ranqueamento, sem a aplicação do *stemming*, passou a ocupar a segunda posição após redução à raiz, contabilizando 232 ocorrências quando agrupadas ao radical “analis”. O radical “cientif” ficou na posição três, totalizando 220 ocorrências, enquanto o radical “result”, tem frequência igual a 173 e está como o quarto mais frequente no *corpus*. Em quinto e sexto lugares, estão os radicais “estud” e “cienc”, com 170 e 169 ocorrências, respectivamente.

Figura 50 – Frequência e ranque antes e depois do *stemming*




Fonte: Captura de tela da pesquisa (2023).

[46] A Figura 30 está na página 107.

Para ilustrar melhor tais comparações, a Figura 50 demonstra uma pequena parte dos resultados obtidos antes e depois da realização do processo de *stemming*, inclusive com o quantitativo de termos já apresentado (pode ser visto na frente da expressão *Rows: xxxx)*. Assim como em outras demonstrações presentes nesta tese, o módulo *Color Manager* (M15-A – Módulo 15-A) e o módulo *Tag Cloud* (M16-A – Módulo 16-A), utilizados nesse conjunto de nós serão usados para geração e apresentação da nuvem de palavras. A Figura 51 corresponde aos radicais dos termos ranqueados após o procedimento de *stemming*.

Conforme demonstrado, são apresentados os cem primeiros radicais armazenados na tabela (essa quantidade pode ser configurada), iniciando-se da parte de baixo e direita da figura, que representa os primeiros radicais mais bem ranqueados, para a parte superior e esquerda, que representam as posições finais. Desse modo, verifica-se o espelhamento da tabela de *stemming* no formato de nuvem de palavra, onde é possível inferir que *pesquis, analis cientif, result* e *estud* representam as cinco primeiras posições no ranque e *citaco, produz, grup, destac* e *caracterist* representam as últimas posições, dentro do agrupamento de cem radicais.

Figura 51 – Nuvem de termos com *stemming*

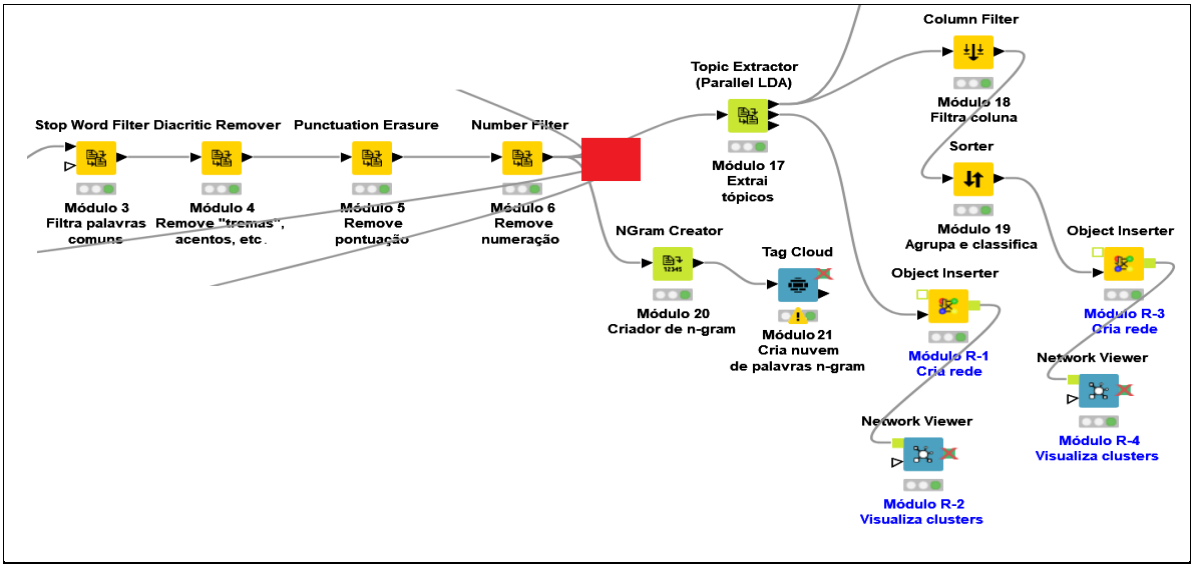

Fonte: Dados da pesquisa (2023).

O processo de *stemming* procura representar um agrupamento de termos em um formato mais simples, que é a raiz ou o radical de cada palavra. Relembrando, as palavras: escrevi, escreveu, escrever, escrevemos e escrevendo poderiam ser reduzidas ao radical “escrev”, tornado uma busca a documentos algo mais abrangente, mas ao mesmo tempo menos específica. Contudo, segundo Porter (1980)

e Berry (2004), é um procedimento que pode contribuir para melhorias em sistemas de recuperação da informação, ao uniformizar as expressões, reduzir as variações e diminuir consideravelmente a quantidade de termos para busca.

Saliente-se dessa forma a flexibilidade do fluxograma proposto, podendo ser utilizado da maneira que melhor servir a pesquisa. A Figura 52 traz um exemplo de alternativa ao posicionamento do módulo que faz o *stemming*, representado pelo quadro na cor vermelha. Se o módulo for colocado naquela posição (entre o Módulo 6 e os seus sucessores) ele realizará as reduções das palavras aos seus radicais e as extrações de tópicos, criação de n-gramas, relacionamento entre os resumos e qualquer outra correlação entre os documentos e termos do *corpus,* serão feitos baseados nas palavras na sua forma básica (formato raiz). Por isso, dependendo do propósito da investigação, existe a possibilidade de posicionar os módulos de forma específica ao projeto.

Figura 52 – Alternativa de posicionamento do módulo de *stemming*




Fonte: Captura de tela da pesquisa (2023).

Outra flexibilidade em relação ao fluxograma diz respeito à saída de dados. Embora se entenda inconveniente migrar dados de um *software* para outro, já que tais ações podem acarretar perda ou imprecisão no momento da transposição dos dados, fazem-se apontamentos sobre a opção do módulo *Excel Writer* identificado também como Módulo Y[47], que aparece no fluxograma posicionado logo no início, mas não tem ligação com outro módulo. O módulo *Excel Writer* faz a exportação da tabela de

[47] Reportar-se a Figura 4 na página 84. O Módulo Y módulo está descrito na cor verde.

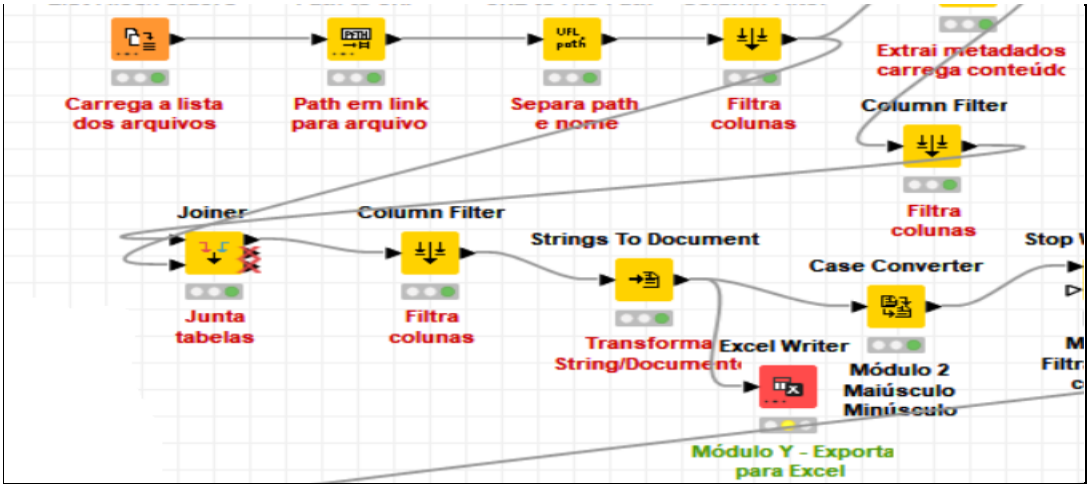
seu módulo antecessor, permitindo que o pesquisador realize o manuseio de dados na referida planilha eletrônica ou em qualquer outro *software* compatível.

Mesmo que a metodologia tenha como finalidade a automatização de processos e a concentração dos procedimentos em apenas uma ferramenta, justamente pela questão de economia de tempo e contenção de falhas na manipulação dos dados, entendem-se como pertinentes as considerações sobre o módulo *Excel Writer*. Na seção de estudos correlatos, percebeu-se que muitos pesquisadores utilizam tal ferramenta em conjunto com outros *software*, demonstrando familiaridade com dados de planilhas eletrônicas.

Para melhor entendimento dos resultados obtidos no Knime, talvez seja conveniente que pelo menos os pesquisadores possam visualizar os dados em *software* que possuem mais familiaridade. Outro fator a destacar é que esse não é um processo de copiar e colar, mas os resultados que estão armazenados na tabela no Knime serão gravados em um arquivo no formato correspondente de planilha eletrônica, diretamente no dispositivo de armazenamento desejado.

A Figura 53 apresenta um exemplo de posicionamento do módulo dentro do fluxograma, logo à frente do módulo *String to Document*. No módulo *Excel Writer*, é possível configurar em que local do computador (unidade, diretório ou pasta) se pretende salvar o arquivo. Por entender que essa configuração de localização e apontamento é comum aos usuários de computador, não se aprofundaram as demonstrações. Pode-se utilizar quantos módulos *Excel Writer* quiser e posicioná-lo para gravação do arquivo sempre que se julgar necessário, desde que o módulo anterior contenha uma tabela.

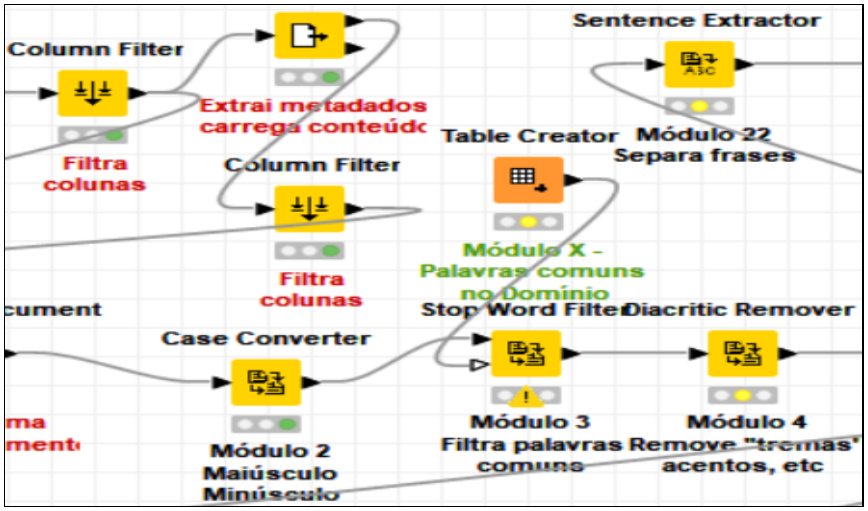
Figura 53 – Módulo para gravar a saída de dados em tabela para o Excel




Fonte: Captura de tela da pesquisa (2023).

Outra consideração conveniente a se fazer é do Módulo X[48] (*Table Creator)*, que, assim como o Módulo Y, está posicionado no início do fluxograma principal e não está conectado a nenhum outro módulo. Como não faz parte do escopo desta pesquisa o mérito de conhecimento prévio ou de especialidades, também não se aprofundou ao assunto. Entretanto, tem-se conhecimento da existência de termos que são específicos de determinados domínios, mas são considerados comuns dentro da especialidade, por isso, o módulo *Table Creator* pode ser colocado na posição anterior ao módulo *Stop Word Filter*, como mostra a Figura 54. O *Stop Word Filter* possui duas entradas que permitem essa configuração, pois, conforme demonstrado na figura, há possibilidade de se criar uma tabela *(Table Creator)* com palavras específicas comuns na especialidade pesquisada que serão também suprimidas.

Figura 54 – Módulo que cria tabela com palavras comuns das especialidades



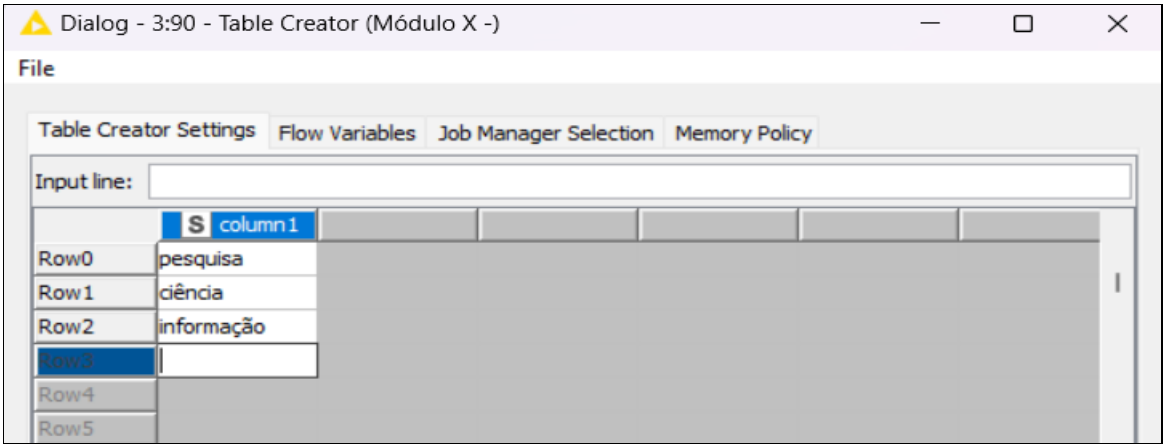

Fonte: Captura de tela da pesquisa (2023).

Por exemplo, ao supor que as palavras “pesquisa”, “ciência” e “informação”, tornam os documentos bastante genéricos na área de CI, ou podem não traduzir de forma aprofundada as temáticas de um *corpus,* pode-se fazer com que elas também sejam desconsideradas na execução da pesquisa. Desse modo, além das palavras comuns, que já são descartadas pelo *Stop Word Filter* de forma automática, as palavras que forem inseridas na tabela do módulo *Table Creator* também serão eliminadas. Em síntese, entende-se que é possível utilizar a metodologia em diversas áreas, possibilitando a realização de estudos mais profícuos na especialidade.

[48] Pode-se verificar a Figura 4 da página 84.

O próprio processo de extração dos termos demonstrado, sem utilização do Módulo X *(Table Creator)*, pode dar indícios de termos que se repetem muito no *corpus* e talvez não sejam significativos para o domínio. Outrossim, um especialista de área pode ter conhecimento prévio de expressões que são convencionais no campo, podendo implementar a tabela segundo sua experiência no domínio. Conforme demonstrado na Figura 55, o pesquisador tem possibilidade de inserir quantas palavras forem necessárias. Esse é um processo simples, mas pode tornar o estudo mais interessante e específico.

Figura 55 – Tela de inserção das palavras na tabela




Fonte: Captura de tela da pesquisa (2023).

Portanto, dada a sobrecarga de informações (DAVIES, 1989; GOLDSCHMIDT; PASSOS, 2005), junto à limitação humana de exploração e registos de grandes quantidades de dados (WITTEN, 2004), percebeu-se que a automatização de processos aqui descritos traz benefícios, tornando os trabalhos menos onerosos (GRÁCIO, 2020). Ao considerar o que sugere a análise de domínio, que o conhecimento não emerge do sujeito individual (HJØRLAND; ALBRECHTSEN, 1995), como uma parte do todo, mas sim o que torna cada parte possível (PÊCHEUX, 1997), acredita-se que a descoberta de interlocuções dos discursos viabilize a caracterização de domínios. Compreende-se então que a operacionalização da abordagem terminológica (HJØRLAND, 2002), pode auxiliar estudos analíticos de domínio.

# 6 CONSIDERAÇÕES FINAIS

A questão que delineou essa tese foi assim proposta: qual a contribuição que as técnicas de mineração de texto podem oferecer para otimizar investigações de análise de domínio na perspectiva da abordagem terminológica? Observando o percurso da metodologia desenvolvida, aplicada como prova de conceito nos resumos do GT7 dos ENANCIBs entre 2012 e 2018 (*corpus*), pode-se admitir que a pesquisa apresenta algumas contribuições significativas. Por exemplo, as técnicas de mineração de texto podem contribuir na instrumentalização para automatizar processos de limpeza, transformação e extração de padrões de escrita a partir de um *corpus* textual qualquer, podendo otimizar investigações que se utilizem de processos de descoberta de conhecimento em texto (KDT).

Assim, presume-se que as hipóteses levantadas foram confirmadas e a metodologia se apresenta como instrumento bastante pertinente para a área de CI, visto que permite a automatização de processos que subsidiam análises de texto e de conteúdo. De modo específico, a operacionalização da abordagem terminológica para análises em domínios (HJØRLAND, 2002), foi atingida com a aplicação dos métodos de dedução de frequência e coocorrência de termos e de análise por categorização temática (oriundos do campo da Linguística), implementados com algoritmos existentes na mineração de texto.

Desse modo, com o uso da metodologia é possível realizar um levantamento terminológico, proporcionando verificar as disposições dos termos nos textos e dos vínculos temáticos entre os documentos de um *corpus*. Os procedimentos utilizados na metodologia, além de amenizar a limitação humana em relação à leitura, exploração e registro de grandes volumes de textos, propicia agilidade e diminui consideravelmente a carga de trabalho. A metodologia viabilizou a descoberta de especificidades no *corpus*, como padrões de linguagem e estruturas de conhecimento presentes nos discursos.

Tais apurações são ratificadas ao reportar-se aos objetivos específicos, pois, verificou-se que os resultados apresentados podem ser considerados promissores em relação a aplicabilidade da abordagem terminológica. Por exemplo, de um total de 1.856 sentenças tidas como válidas, a maior parte delas, 806, eram compostas pelo intervalo entre 08 e 14 termos, e outras 620 possuíam de 15 a 21 termos. Esse resultado demonstra que, mesmo sem trocar informações a respeito do

desenvolvimento dos textos, o grupo de autores possui um laço comum entre eles, que é o padrão de escrita, referente à quantidade de termos presentes nas frases dos discursos.

Assim como na extração da quantidade de termos por sentença, a metodologia propiciou a automatização do levantamento das ocorrências dos termos (simples e compostos) mais utilizados pelos autores, revelando um tipo de interlocução entre eles, que mesmo sem conhecimento prévio, apresentam perspectivas textuais parecidas, formando uma comunidade com discursos correlatos. (HJØRLAND; ALBRECHTSEN, 1995; MACIAS-CHAPULA, 1998). São apontamentos que podem estender-se a estudos que visem a compreensão dos discursos das comunicações científicas, ou de relações teórico-metodológica entre autores ou documentos, ou ainda entre ambos.

Por meio do fluxograma desenvolvido, a ferramenta de mineração de texto realizou a varredura nos documentos, separou os termos individualmente, efetuou a contagem, fez o agrupamento, o ranqueamento e gerou uma lista com todos os vocábulos descobertos. Inicialmente foram encontradas um total de 23.515 palavras, que depois de agrupadas e ordenadas, passaram a ser representadas por 5.820 termos, já descartadas as pontuações e palavras comuns do vocabulário português. Assim, foi possível gerar um gráfico de nuvem de palavras com os 100 primeiros termos mais bem ranqueados.

Da mesma forma, a metodologia permitiu que fossem encontrados *cluster* de palavras, os quais são passíveis de configuração, podendo variar na quantidade de termos que o pesquisador planeja encontrar (2, 3, 4, 5 e assim por diante). Neste estudo, foram apresentados os n-gramas compostos por dois e três termos (bigramas e trigramas), nos quais encontraram-se 22.238 combinações de dois termos e 25.088 com três palavras. Fez-se a demonstração dessa possibilidade em gráficos de nuvem de palavras, em que os bigramas: produção científica, ciência informação, artigos publicados, programas pós-graduação, periódicos científicos e comunicação científica são os mais recorrentes. Os trigramas com maior número de coocorrências são: web of science; ciência informação brasil; desenvolvimento científico tecnológico; estudos métricos informação; nacional desenvolvimento científico e análise de redes sociais[49].

[49] As figuras dos bigramas e trigramas extraídos encontram-se na página 110 e 111 respectivamente.

Tais constatações apontam novamente que as técnicas de dedução de frequência e coocorrência de termos, empregues com a mineração de texto, foram adequados para o cumprimento dos objetivos, pois, possibilitaram apurar as dinâmicas de uso dos termos e n-gramas, revelando especificidades desconhecidas da linguagem dos autores nos seus discursos, como apontado por López-Huertas (2015) sobre as possibilidades da análise de domínio revelar estruturas de conhecimento, dinâmicas na construção dos discursos e padrões de linguagem e comunicação em domínios específicos.

Relembrando os paradigmas da CI (físico, cognitivo e social), ressalta-se que os algoritmos de mineração encontram os indicadores com cálculos que consideram o *corpus* textual como um todo*,* e não por documentos individuais, permitindo uma visão mais abrangente (CAPURRO, 2003) das comunicações científicas. Tais preceitos estão em conformidade aos apontamentos dos estudos terminológicos (HJØRLAND, 2002; 2022) empregados para análise de domínio (HJØRLAND; ALBRECHTSEN, 1995; HJØRLAND, 2017), principalmente o reportar-se a Hjørland e Albrechtsen (1995) quando definem que, para a área de CI, os domínios devem encarados como grupos de pensamento ou comunidades discursivas, que constituem a divisão social do trabalho.

Outros dois objetivos eram, encontrar tópicos que representassem os temas abordados no *corpus* e verificar os conjuntos de resumos pertencentes à determinado tema. Tais objetivos foram atingidos com emprego da técnica de análise por categorização temática, que teve o processo automatizado por meio do algoritmo LDA (mineração de texto). Conforme as configurações do módulo, o algoritmo foi capaz de distinguir os 5 (cinco) tópicos, contendo 10 (dez) palavras cada um, que serviram para identificar as principais temáticas do *corpus*. E ainda, por meio de cálculos probabilísticos, o algoritmo LDA classificou e agrupou os resumos de acordo com a proximidade dos temas[50].

No último objetivo, em que se visou apresentar uma alternativa complementar para instrumentalização da abordagem terminológica, propôs-se a realização do processo de *stemming* (redução ao radical) dos termos. Sem o processo de agrupamento das palavras ao seu radical foram encontrados 5.820 vocábulos, e após

[50] As figuras que apresentam os agrupamentos dos tópicos e os *clusters* dos resumos estão no intervalo de páginas de 119 a 122.

a realização do *stemming* chegou-se ao quantitativo de 3.657, portanto, o processo de agrupamento das palavras ao seu radical conseguiu reduzir cerca de 2.163 termos. Palavras de mesmo radical têm significados semelhantes, variando somente seus afixos – prefixo e sufixo – (PORTER, 1980), e a redução e padronização dos termos podem auxiliar no processo de indexação de documentos e no aperfeiçoamento de sistemas de recuperação da informação (PORTER, 1980; BERRY, 2004).

Diante dos resultados obtidos pode-se deduzir que os objetivos propostos foram alcançados. A proposta de uma metodologia para instrumentalização automatizada para abordagem terminológica (HJØRLAND, 2002) foi elaborada empregando as técnicas de dedução de frequência e coocorrência de termos e análise por categorização temática (PÊCHEUX, 1997), implementadas por meio de um fluxograma para mineração de texto, criado na ferramenta Knime. Entende-se que a metodologia desenvolvida possibilitou encontrar nos discursos (*corpus* – resumos), padrões de linguagem que não eram conhecidos, como: quantidade de sentenças; termos simples e compostos mais utilizados; tópicos e temáticas representativas; e agrupamento de textos por semelhança temática. Tais padrões podem exprimir até relações estruturais de conectividade teórico-metodológica de um domínio.

Entendendo a metodologia como promissora para a área, cabe um adendo das impressões de colaboração para a CI, bem como a pertinência das temáticas abordadas. Inicialmente, no processo de revisão da bibliografia correlata, percebeu-se apreciação de estudos analíticos de domínio em vários textos. (LEE; KIM; KIM, 2010; GUIMARÃES; GONZÁLEZ; ALENCAR, 2012; LU; WOLFRAM, 2012; GANDRA; DUARTE, 2013; SMIRAGLIA, 2013; GRÁCIO; OLIVEIRA, 2014; SUENAGA; CERVANTES, 2014; DANTE *et al.,* 2015; GUIMARÃES *et al.,* 2015; AMORIM; CAFÉ, 2016; GUIMARÃES *et al.,* 2017; ALMEIDA; DIAS, 2018; MANHIQUE; CASARIN, 2018; NAKANO *et al.*, 2018; TOGNOLI; SILVA; SILVA, 2019; DAMUS; ACUÑA, 2019; TRABADELA-ROBLES *et al.*, 2020; REGO-PIVA; CASARIN; GUIMARÃES, 2021; BARROS; LAIPELT, 2021).

Outra temática bastante debatida na bibliografia foi o emprego de análises bibliométricas (medidas de citação e cocitação) para levantamento de autores expoentes e influenciadores epistemológicos em diversas áreas. (SANTAREM, 2011; GUIMARÃES; GONZÁLEZ; ALENCAR, 2012; LU; WOLFRAM, 2012; SMIRAGLIA, 2013; GRÁCIO; OLIVEIRA, 2014; GUIMARÃES *et al.,* 2014; ROSAS; GRÁCIO, 2015; DANTE *et al.,* 2015; GUIMARÃES *et al.*, 2015; TOGNOLI; SILVA; SILVA, 2019;

TRABADELA-ROBLES *et al.,* 2020; REGO-PIVA; CASARIN; GUIMARÃES, 2021; FERREIRA; CORREA, 2021).

Além desses assuntos, outros estudos realizaram investigações temáticas, baseadas nas técnicas de dedução de frequências e coocorrências de termos (copalavras), geralmente implementadas com a mineração de texto. (LEE; KIM; KIM, 2010; CALVO FUENTE; CANTOS MATEO; ZULUETA GARCÍA, 2013; SMIRAGLIA, 2013; DANTE *et al.,* 2015; SÉRGIO; SILVA; GONÇALVES, 2016; GUIMARÃES *et al.,* 2017, TOGNOLI; SILVA; SILVA, 2019; HSU; LI, 2019; JOO; OH, 2019; JOO, 2020; MOKHTARPOUR; KHASSEH, 2021; REGO-PIVA; CASARIN; GUIMARÃES, 2021; MARQUES; MARQUES; MACULAN, 2021).

Desse modo, entende-se que os direcionamentos dados por esta tese estão condizentes com vários estudos já realizados e, aparentemente, não se distanciam de temáticas atuais. Outrossim, além de observar a importância dada a estudos das produções científicas em diversas áreas, percebeu-se também manifesta particularidade a respeito da manipulação dos dados serem feitas de modo manual (SMIRAGLIA, 2013; ROSAS; GRÁCIO, 2015; MANHIQUE; CASARIN, 2018; HSU; LI, 2019; JOO; OH, 2019). Embora alguns estudos se apresentem com maiores e outros com menores quantidades de documentos, o fato a ser destacado é que a fase de pré-processamento dos dados, supostamente, tem sido motivo de muito trabalho. Como se já não bastasse essa etapa apresentar-se como dificultosa, imagina-se que correlacionar os documentos "*per se"* seja outra tarefa extremamente complexa.

Sobre a questão de necessidade de preparação dos arquivos para serem carregados no *software* de mineração utilizado nesta pesquisa, entende-se que esse é um processo trivial no cenário acadêmico. Desse modo, compreende-se que tal circunstância não é uma exclusividade na utilização do *software* Knime. Percebeu-se a ferramenta como bastante apropriada para realização de descoberta de conhecimento em texto (pré-processamento, transformação, manipulação, mineração de texto) em todas suas etapas. Ainda que o custo computacional se altere em decorrência da quantidade de textos do *corpus* a ser estudado; e entendendo que o conjunto documental aqui utilizado é considerado pequeno (287) para a área de mineração de texto, a metodologia elaborada foi executada no *software* em poucos minutos.

Observando que os dados utilizados nesta pesquisa foram coletados e preparados para serem lidos pelo Knime, entende-se que, igualmente, dados obtidos

em bases como LISA, Scopus e WoS também teriam que ser organizados para utilização em diferentes *software*. Desse modo, a metodologia pode ser replicada, independentemente da origem dos dados, bastando apenas estarem no formato próprio de inserção na ferramenta.

Assim, acredita-se que a tese pode corroborar a expansão de estudos da literatura científica para além das bases mais conhecidas (LISA, Scopus, WoS), mesmo em países à margem delas (OLIVEIRA; GRÁCIO, 2009). O próprio corpus utilizado na tese é um exemplo disso (comunicações científicas de eventos), não se prendendo a dados estruturalmente organizados. Essa flexibilidade contribuiria novamente para diminuição de trabalhos manuais; e mais, os procedimentos descritos podem ser aplicados às diversas seções dos textos (palavras-chave, títulos, resumos etc.). Destaque-se novamente a flexibilidade de posicionamento dos módulos[51], que podem se adequar a diferentes estudos em diversas especialidades.

Retomando novamente López-Huertas (2015), a análise de domínio pode descobrir estruturas de conhecimento, dinâmica, padrões de linguagem e comunicação e comportamento de cooperação em domínios específicos. Entretanto, bem pouco se encontrou nos estudos correlatos sobre como seriam operacionalizados tais análises, parecendo existir uma lacuna entre o debate “do que é possível descobrir” e “como é possível descobrir”. Ao se reportar à Aranha e Passos (2006), tem-se que a mineração de texto é vista como ferramenta capaz de encontrar regularidades, padrões e tendências em *corpus* textuais. As conotações dos autores parecem bastante afinadas, portanto, uma metodologia que consiga preencher essa lacuna é perfeitamente oportuna.

Outra constatação de pertinência da tese aparece em Urbizagastegui-Alvarado (2021). Segundo o autor, a mineração de texto vem se destacando no mundo acadêmico como uma ferramenta útil para investigações, já que a comunidade científica tem se interessado por estudos em linguagem natural. Ainda conforme o autor, o crescente interesse tem como base a necessidade de se encontrar aplicações relacionadas a recuperação da informação, extração de dados, resumo de documentos, descoberta de padrões, associações, regras e a realização de análises tanto qualitativas quanto quantitativas em documentos textuais.

---

[51] Ver o final da página 76 e a 128 sobre a questão de posicionamento e inserção de módulo.

O que também parece ratificar a pertinência da pesquisa são os apontamentos de Marian (2015) sobre área de *Linguística de Corpus*, a qual entende que os processos computacionais são extremamente úteis, ao possibilitar estudos de linguagem por ângulos probabilísticos, quantificando padrões com ferramentas estatísticas e compreendendo que as ocorrências de termos nos textos não são casuais, pois, tais ocorrências denotam significados aos discursos.

Ainda, Macias-Chapula (1998, p. 134) assevera que, “se o documento é a expressão de uma pessoa ou de um grupo trabalhando em uma frente de pesquisa, podemos dizer alguma coisa sobre as relações entre as pessoas a partir dos próprios documentos.” Desse modo, percebe-se que a metodologia preserva os apontamentos relativos aos estudos sociocognitivos, referenciados para análise de domínio (HJØRLAND; ALBRECHTSEN, 1995; HJØRLAND, 2002) e a socioterminologia (GAUDIN, 1993, 2014; CABRÉ, 2005; FAULSTICH, 2006), estando ainda em consonância com a subárea linguística de *corpus* (MARIAN, 2015), que preconiza o uso da tecnologia computacional para análises de textos.

É o que se desenvolveu nesta tese, uma Metodologia para Abordagem Terminológica da Análise de Domínio baseada na Mineração de Texto, doravante denominada de MATAD-MT ou simplesmente de MATAD. Dado o caráter de incipiência da metodologia, provavelmente ainda serão encontrados hiatos. Uma das limitações já encontradas refere-se a questão da língua em que os documentos estão escritos[52]. Por exemplo, a execução de procedimentos como a limpeza de palavras comuns (*Stop Words*) de um vocabulário é dependente da língua. Palavras comuns da língua espanhola são diferentes das palavras comuns da língua portuguesa, assim como elas também se diferem de outras línguas.

Como as técnicas empregadas advêm da linguística, e ainda, alguns algoritmos para mineração de texto precisam ser configurados em determinada língua para extraírem resultados de forma correta, convém organizar os documentos e executar a metodologia em textos separados por língua. Como trabalho futuro, poder-se-ia implementar algum módulo que reconhecesse a língua do texto e fizesse a separação dos documentos de forma automática.

---

[52] Assunto exposto no início da página 119.

Outra limitação está na questão de apresentação dos *cluster* dos tópicos e dos resumos[53]. Embora apareçam na tese como figuras de rede ego, essa representação não significa ligações de uma rede social, por exemplo, graus de relacionamentos entre os nós (grau de intermediação, grau de proximidade, ou centralidade) [54]. As figuras apenas demonstram que os resumos possuem interlocuções temáticas, que foram obtidas com a extração dos tópicos. Em algum trabalho futuro poderia se verificar possibilidades de melhorias nesse quesito. Ainda no âmbito de trabalhos futuros, percebeu-se que a metodologia pode ser aplicada nos estudos métricos da informação, não só na bibliometria e cienciometria, mas há aparente conformidade em pesquisas que utilizem a informetria, já que não se prende a dados estruturados. Como não era escopo da pesquisa definir ou comparar tais métodos de estudos, traz-se apenas um breve demonstrativo de suas características.

Figura 56 – Características da bibliometria, cienciometria e informetria

| Tipologia | Bibliometria | Cienciometria | Informetria |
|---|---|---|---|
| Objetos de estudo | Livros, documentos, revistas, artigos, autores, usuários | Disciplinas, assunto, áreas, campos | Palavras, documentos, bases de dados |
| Variáveis | Número de empréstimos (circulação) e de citações, freqüência de extensão de frases etc. | Fatores que diferenciam as subdisciplinas. Revistas, autores, documentos. Como os cientistas se comunicam. | Difere da cienciometria no propósito das variáveis; por exemplo, medir a recuperação, a relevância, a revocação etc. |
| Métodos | Ranking, freqüência, distribuição | Análise de conjunto e de correspondência. | Modelo vetor-espaço modelos booleanos de recuperação, modelos probabilísticos; linguagem de processamento, abordagens baseadas no conhecimento, tesauros. |
| Objetivos | Alocar recursos: tempo, dinheiro etc. | Identificar domínios de interesse. Onde os assuntos estão concentrados. Compreender como e quanto os cientistas se comunicam. | Melhorar a eficiência da recuperação. |

Fonte: Macias-Chapula (1998, p. 135).

Macias-Chapula (1998, p. 135) define a informetria como “o estudo dos aspectos quantitativos da informação em qualquer formato, e não apenas registros

[53] Pode-se vera nas figuras que estão no intervalo de páginas de 119 a 122.
[54] Esses termos são utilizados na metodologia de Análise de Redes Sociais (ARS).

catalográficos ou bibliografias, referente a qualquer grupo social, e não apenas aos cientistas.” Segundo o autor, são estudos que podem incorporar, utilizar e ampliar pesquisas que fazem avaliação da informação fora dos limites da bibliometria e da cienciometria. Os aspectos de cada tipo de estudo estão expressos na Figura 56, que apresenta uma tabela sintética desenvolvida pelo autor.

Pode-se observar ainda que o *corpus* de análise foi formado por um conjunto de textos centrado em um tema ou grupo de trabalho (GT7), ou seja, monotemático. Seria interessante analisar o emprego da metodologia em análise de textos sobre vários temas, a fim de se verificar se o método é eficiente para *corpus* politemático, portanto, considera-se essa questão ainda como limitante da pesquisa, podendo se desenvolver trabalhos futuros nesse âmbito.

Nessa mesma linha, pode-se destacar a junção de estudos métricos da informação já realizados, com novos estudos de abordagem terminológica. Ao replicar a abordagem terminológica em um mesmo período dos estudos bibliométricos, seria possível caracterizar um domínio por meio de duas abordagens, como sugere Hjørland (2002). Outros estudos que podem ser realizados seguem a linha do que foi o proposto por Guimarães (2019), um paralelo entre a terminologia de um domínio e as linguagens de indexação da área. Poder-se-ia também utilizar o mesmo *corpus* deste estudo, empregando o processo de *stemming*, como forma de comparação aos termos aqui encontrados sem a redução ao radical.

Ainda seria possível a execução de estudos das evoluções terminológicas dentro dos domínios, pois, a realização do levantamento terminológico em intervalos diferentes poderia indicar mudanças diacrônicas características de uma especialidade. Sobre a questão de periodicidade, a metodologia também serviria para indicar evoluções temáticas em um domínio, bem como o crescimento ou declínio de um tema dentro da especialidade ao comparar-se períodos diferentes.

Com apresentado na pesquisa, além de conseguir destacar microdomínios dentro de um domínio, existe a possibilidade de realização da abordagem terminológica no microdomínio. Por exemplo, ao encontrarmos os *clusters* com os resumos que formavam os conjuntos referentes aos tópicos 0, 1, 2, 3 e 4, poderia se separar os resumos em pastas distintas e aplicar novamente a metodologia aos resumos pertencentes ao tópico 0, conseguindo encontrar especificidades dos discursos dentro desse microdomínio.

## REFERÊNCIAS

ALMEIDA, D. P. R.; ANTONIO, D. M.; BOCCATO, V. R. C.; GONÇALVES, M. C.; RAMALHO, R. A. S.. Paradigmas contemporâneos da Ciência da Informação: a recuperação da informação como ponto focal. *In:* **Informação & Cognição**, v. 6, n. 1, p. 16-27, 2007. DOI: https://doi.org/10.36311/1807-8281.2007.v6n1.745.

ALMEIDA, J. F. V. R. de; DIAS, G. A.. Estado da arte sobre análise de domínio no campo da Ciência da Informação brasileira. *In:* **Brazilian Journal of Information Science: research trends**, [S. l.], v. 13, n. 3, p. 26-45, 2019. DOI: 10.36311/1981-1640.2019.v13n3.04.p26.

AMORIM, I. S.; BRÄSCHER, M.. Cartografia: debates sobre os métodos da análise de domínio. *In:* Encontro Nacional de Pesquisa em Ciência da Informação, ENANCIB 17, 2016, Salvador, BA. **Anais...** Salvador, BA: UFBA, 2016. Disponível em: http://www.ufpb.br/evento/lti/ocs/index.php/enancib2016/enancib2016/paper/viewFile/3586/2235. Acesso em: 07 jun. 2022.

AMORIM, I. S.; CAFÉ, L. M. A.. Análise de domínio hjørlandiana sob a luz da filosofia de Deleuze. *In:* Encontro Nacional de Pesquisa em Ciência da Informação, ENANCIB 15, p. 1045-1051, 2014, Belo Horizonte, MG. **Anais...** Belo Horizonte, MG: UFMG, 2014. Disponível em: http://enancib2014.eci.ufmg.br/documentos/anais/anais-gt2. Acesso em: 08 jul. 2022.

AMORIM, I. S.; CAFÉ, L. M. A.. Os conceitos de comunidade discursiva, domínio e linguagem na análise de domínio hjørlandiana. *In:* Encontro Nacional de Pesquisa em Ciência da Informação, ENANCIB 17, 2016, Salvador, BA. **Anais...** Salvador, BA: UFBA, 2016. Disponível em: http://hdl.handle.net/20.500.11959/brapci/190585. Acesso em: 10 jun. 2022.

ANCIB. **Site da Ancib**, 2022. Páginas de histórico e sobre a Associação Nacional de Pesquisa e Pós-Graduação em Ciência da Informação (ANCIB). Disponível em: https://ancib.org/. Acesso em: 19 jul. 2022.

ARANHA, C.; PASSOS, E.. A tecnologia de mineração de textos. *In:* **Revista Eletrônica de Sistemas de Informação – RESI**, v. 5, n. 2, 2006. DOI: https://doi.org/10.21529/RESI.2006.0502001.

ARAÚJO, C. A. A.. O que é ciência da informação? *In:* **Informação & Informação**, v. 19, n. 1, p. 1-30, 2014. DOI: 10.5433/1981-8920.2014v19n1p01.

BARROS, T. H. B.; LAIPELT, R. C. F.. Uma análise de domínio da área de Organização e Representação do Conhecimento no contexto do periódico Em

Questão. *In:* **Em Questão**, Porto Alegre, v. 27, n. 4, p. 438–468, 2021. DOI: 10.19132/1808-5245274.438-468.

BRASIL. **Site do Ministério da Educação**. Página da história e missão da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES). Disponível em: https://www.gov.br/capes/pt-br/acesso-a-informacao/institucional/historia-e-missao. Acesso em: 26 jun. 2022.

BERRY, M. W.; LINOFF, G.. **Data Mining techniques for Marketing, Sales, and Customer Support.** New York: John Wiley & Sons. 1997.

BERRY, M. W., org.. **Survey of Text Mining**. New York, NY: Springer New York, 2004. DOI: https://doi.org/10.1007/978-1-4757-4305-0.

BLEI, D. M.; NG, A. Y.; JORDAN, M. I.. Latent Dirichlet Allocation. **Journal of Machine Learning Research**, v. 3, p. 993-1022. 2003. Disponível em: https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf. Acesso em: 04 mar. 2022.

BORKO, H.. Information science: what is it? **American Documentation**, Washington, v. 19, n. 1, p. 3-5, jan. 1968.

CABRÉ, M. T.. **La Terminología:** representación y comunicación, elementos para una teoría de base comunicativa y otros artículos. 2005. Girona: Documenta Universitaria.

CALVO FUENTE, V.; CANTOS MATEOS, G.; ZULUETA GARCÍA, M.. Á. Delimitación temática de la investigación española en fisioterapia a través del análisis de co-palabras. *In:* **Scire:** representación y organización del conocimiento, v. 19, n. 2, p. 98-101, 2013. Disponível em: http://hdl.handle.net/20.500.11959/brapci/167923. Acesso em: 09 jun. 2022.

CAPURRO, R.. Epistemologia e Ciência da Informação. *In:* Encontro Nacional de Pesquisa em Ciência da Informação, ENANCIB 5, 2003, Belo Horizonte, MG. **Anais...** Belo Horizonte, MG: ANCIB, 2003. Disponível em: http://www.capurro.de/enancib_p.htm. Acesso em: 06 abr. 2022.

CAPURRO, R.; HJØRLAND, B.. O conceito de informação. **Perspectivas em Ciência da Informação**, v. 12, n. 1, p. 148–207, jan. 2007. DOI: https://doi.org/10.1590/S1413-99362007000100012.

CASTELL, M.. **A sociedade em rede**. 8 ed.; tradução de Roneide Vanancio Majer com colaboração de Klauss Brandini Gerhardt. São Paulo: Paz e Terra, 2005.

DAMUS, M. A.; ACUÑA, G. N.. Aproximación al Análisis de Dominio (AD) desde la investigación en Bibliotecología y Ciencia de la Información. *In:* **e-Ciencias de la Información**, [S. l.], v. 9, n. 2, 2019. DOI: https://doi.org/10.15517/eci.v9i2.37497.

DANTE, G.; LA ROSA, G.; LOPEZ, P.; BAYONA, A. L.. Domain Analysis of the research in professional competences, technology and engineering cluster. *In:* **Procedia** – Social and Behavioral Sciences, v. 182, p. 163-172, 2015. DOI: https://doi.org/10.1016/j.sbspro.2015.04.752.

DAVIES, R.. The creation of new knowledge by information retrieval and classification. *In:* **Journal of Documentation**, v. 45, n. 4, p. 273-301, 1989. DOI: https://doi.org/10.1108/eb026846.

EVERITT, B. S.; LANDAU, S.; LEESE, M.; STAHL, D.. **Cluster analysis:** Wiley series in probability and statistics. 5 ed., United Kingdom: John Wiley & Sons, 348 p., 2011.

FAULSTICH, E. L. J.. Princípios formais e funcionais de variação em terminologia. *In:* **Seminário de Terminologia Teórica**. 1999. Barcelona-Espanha.

FAULSTICH, E. L. J.. A socioterminologia na comunicação científica e técnica. *In:* **Ciência e Cultura**, v. 58, n. 2, p. 27-31, abr./jun. 2006. Disponível em: http://cienciaecultura.bvs.br/pdf/cic/v58n2/a12v58n2. Acesso em: 11 ago. 2022.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P.; UTHURUSAMY, R.. **Advances in Knowledge Discovery and Data Mining**. California, USA: AAAI Press, 1996.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P.. **From Data Mining to Knowledge Discovery in Databases**. v. 17, n. 3, p. 37-54, California, USA: AAAI Press, 1996.

FELDMAN, R., SANGER, J.. **The Text Mining Handbook**: advanced approaches in analyzing unstructured data. Cambridge: Cambridge University Press, 2007. DOI: 10.1017/CBO9780511546914.

FELDMAN, R.; DAGAN, I.. Knowledge Discovery in Textual Databases (KDT). **Proceedings of the First International Conference on Knowledge Discovery and Data Mining** (KDD'95). AAAI Press, p. 112–117, 1995. Disponível em: https://www.aaai.org/Papers/KDD/1995/KDD95-012.pdf. Acesso em: 25 jul. 2022.

FERREIRA, M. H. W.; CORREA, R. F.. Mineração de textos científicos: análise de artigos de periódicos científicos brasileiros da área de Ciência da Informação. *In:* **Em Questão**, Porto Alegre, v. 27, n. 1, p. 237–262, 2020. DOI: 10.19132/1808-5245271.237-262.

FRAKES, W. B.; BAEZA-YATES, R.. **Information retrievel:** data structures algorihms. New Jersey: Prentice Hall, 1992.

GANDRA, T. K.; DUARTE, A. B. S.. Interlocuções entre a análise de domínio e os estudos de usuários da informação: contribuições para uma abordagem sociocognitiva. *In:* XIV Encontro Nacional de Pesquisa em Ciência da Informação, 2013. Florianópolis, SC. **Anais...**. Florianópolis, SC. ENANCIB 2013. Disponível em: http://hdl.handle.net/20.500.11959/brapci/184477. Acesso em: 12 jun. 2022.

GAUDIN, F.. **Pour une socioterminologie:** des problemes semantiques aux pratiques institutionnelles. Rouen: Publications de l'Universite de Rouen. 1993.

GAUDIN, F.. **Socioterminologie**: une approche sociolinguistique de la terminologie. Bruxelas: De Boeck Supérieur, Duculot (Coleção Champs linguistiques), 2003, 288 p. DOI: 10.3917/dbu.gaudi.2003.01.

GAUDIN, F.. Socioterminologia: um itinerário bem-sucedido. In: ISQUERDO, A. N., DAL CORNO, G. M. (org.). As ciências do léxico: lexicologia, lexicografia, terminologia. v. VII. Campo Grande: Editora UFMS, 2014.

GOLDSCHMIDT, R.; PASSOS. E.. **Data Mining:** um guia prático. Editora Campus, Rio de Janeiro: Elsevier, 2005.

GRÁCIO, M. C. C.; OLIVEIRA, E. F. T de. Estudos de análise de cocitação de autores: uma abordagem teórico-metodológica para a compreensão de um domínio. *In:* **Tendências da Pesquisa Brasileira em Ciência da Informação**, v. 7, n. 1, p. 1-22, 2014. Disponível em: http://hdl.handle.net/11449/114829. Acesso em: 12 jun. 2022.

GRÁCIO, M. C. C.. **Análises relacionais de citação para a identificação de domínios científicos: uma aplicação no campo dos Estudos Métricos da Informação no Brasil**. Marília/SP: Editora Unesp, 252 p. 2020. DOI: https://doi.org/10.36311/2020.978-65-86546-12-5.

GUIMARÃES, J. A. C.. Análise de domínio como perspectiva metodológica em organização da informação. *In:* **Ciência da Informação**, [S. l.], v. 43, n. 1, p. 13-21, 2014. DOI: 10.18225/ci.inf.v43i1.1415.

GUIMARÃES, J. A. C.; GONZÁLEZ, J. A. M.; ALENCAR, M. F.. A análise documental no universo científico dos Enancibs: elementos para uma análise de domínio. *In:* XIII Encontro Nacional de Pesquisa em Ciência da Informação, 2012, Rio de Janeiro. **Anais...**. Rio de Janeiro: ENANCIB, 2012, v. XII, p. 2-17. Disponível em: http://hdl.handle.net/20.500.11959/brapci/182251. Acesso em: 17 ago. 2022.

GUIMARÃES, J. A. C.; SANTOS, A. Y. dos; CÂNDIDO, G. G.; PINHO, F. A.. A dimensão temática da pesquisa em organização do conhecimento: uma análise de

domínio dos congressos nacionais e regionais da ISKO (Brasil, Espanha e América do Norte). *In:* **Scire**: representación y organización del conocimiento**,** v. 20, n. 2, p. 19-25, 2014. Disponível em: http://hdl.handle.net/11449/232326. Acesso em: 08 ago. 2022.

GUIMARÃES, J. A. C.; TOGNOLI, N. B.. Provenance as a domain analysis approach in archival knowledge organization. *In:* **Knowledge Organization**. Wurzburg: Ergon-verlag, v. 42, n. 8, p. 562-569, 2015. Disponível em: http://hdl.handle.net/11449/164830. Acesso em 15 dez. 2022.

GUIMARÃES, J. A. C.; MATOS, D. F. de O.; DOS SANTOS, A. Y.; SALES, R.. La dimensión conceptual de la organización del conocimiento en el universo científico de la ISKO: un análisis de dominio de los congresos de ISKO-Brasil, ISKO-España, ISKO-Norteamerica e ISKO-Francia. *In:* **Scire**: representación y organización del conocimiento, [S. l.], v. 21, n. 2, p. 13–26, 2015. DOI: 10.54886/scire. v21i2.4245.

GUIMARÃES, J. A. C.; MARTÍNEZ-ÁVILA, D.; OLIVEIRA, A. M.; GOMES, P. H. C.. Análise de domínio em ciência da informação: uma análise da produção científica internacional. *In:* **Scire:** representación y organización del conocimiento, v. 23, n. 2, p. 37-43, 2017. Disponível em: http://hdl.handle.net/20.500.11959/brapci/168150. Acesso em: 18 ago. 2022.

HAN, J.; KAMBER, M.. **Data Mining:** concepts and techniques. 2 ed. San Francisco: Morgan Kaufmann Publishers, 2006.

HERRERA VARELA, R.. **Bibliomining:** minería de datos y descubrimiento de conocimiento en bases de datos aplicados al ámbito bibliotecario. 2006. Disponível em https://fdocuments.ec/document/bibliomining-mineria-de-datos-y-descubrimiento-de-crisp-dm-dentro-de.html. Acesso em: 03 fev. 2021.

HJØRLAND, B.; ALBRECHTSEN, H.. Toward a new horizon in Information Science: Domain-Analysis. *In:* **Journal of the American Society for Information Science**, v.6, n. 6, p. 400-425, 1995.

HJØRLAND, B.. Domain analysis in information: eleven approaches-traditional as well as innovative. *In:* **Journal of Documentation**, v. 58, n. 4, p. 422-462, 2002.

HJØRLAND, B.. Domain Analysis. *In:* **Knowledge Organization**, v. 44, n. 6, p. 436-464, 2017, DOI: https://doi.org/10.5771/0943-7444-2017-6-436.

HJØRLAND, B.. Terminology. **ISKO** – Encyclopedia of Knowledge Organization, eds. Birger Hjørland and Claudio Gnoli, 2022. Disponível em: https://www.isko.org/cyclo/terminology. Acesso em: 19 dez. 2022.

HSU, W.-C.; LI, J.-H.. Visualising and mapping the intellectual structure of medical big data. **Journal of Information Science**, v. 45, n. 2, p. 239-258, 2019, DOI: https://doi.org/10.1177/0165551518782824.

JOO, S.. Exploring the domain of information “users”: semantic analysis of wikipedia articles. *In:* **Journal of Library and Information Studies**, v. 18, n. 1. p. 1-23, 2020. DOI: https://doi.org/10.6182/jlis.202006_18(1).001.

JOO, S.; OH, K. E.. Differences in the research domains of knowledge organization between academic researchers and library practitioners: preliminary results. *In:* ACM/IEEE **Joint Conference on Digital Libraries** (JCDL) p. 357-358, 2019, DOI: https://doi.org/10.1109/JCDL.2019.00069.

KUHN, T. S.. **A estrutura das revoluções científicas**. 7. ed. São Paulo: Perspectiva, 2003.

LANGRIDGE, D.. **Classificação:** abordagem para estudantes de biblioteconomia. 1. ed. Rio de Janeiro: Interciência, 1977.

LEE, J. Y.; KIM, H.; KIM, P. J.. Domain Analysis with Text Mining: analysis of digital library research trends using profiling methods. *In:* **Journal of Information Science**, v. 36, p.144-161. 2010. DOI: https://doi.org/10.1177/0165551509353251.

LOBAINA, E. M. R.; SUÁREZ, C. P. R.. Resultados obtenidos en un proceso de minería de datos aplicado a una base de datos que contiene información bibliográfica referida a cuatro segmentos de la ciência. *In:* **JISTEM – Journal of Information Systems and Technology Management (Online)**, v. 15, 2018. DOI: 10.4301/S1807-1775201815003.

LLOYD, C.. **As estruturas da história**. Rio de Janeiro: Zahar, 1995. 400 p.

LOPES, M. C. S.. **Mineração de dados textuais utilizando técnicas de *Clustering* para o idioma português**. 2004, 191 f. Tese (Doutorado em Engenharia Civil), Universidade Federal do Rio de Janeiro-RJ.

LÓPEZ-HUERTAS, M. J.. Domain analysis for interdisciplinary knowledge domains. *In:* **Knowledge Organization**, n. 42, p. 570–580, 2015. Disponível em: https://www.nomos-elibrary.de/10.5771/0943-7444-2015-8-570/domain-analysis-for-interdisciplinary-knowledge-domains-jahrgang-42-2015-heft-8?page=1. Acesso em: 05 ago. 2022.

MACIAS-CHAPULA, C. A.. O papel da informetria e da cienciometria e sua perspectiva nacional e internacional. *In:* **Ciência da Informação**, [S. l.], v. 27, n. 2, 1998. DOI: 10.18225/ci.inf.v27i2.794.

MANNING, C.; RAGHAVAN, P.; SCHUETZE, H.. **Introduction to information retrieval.** [online]. Cambridge: Cambridge University Press, 2008. DOI: doi:10.1017/CBO9780511809071.

MARIAN, J.. Terminologia especializada: um estudo baseado na linguística. *In:* **Caderno PAIC**, v. 16, n. 1, p. 475-487, 2015. ISSN: 2447-8954. Disponível em: https://cadernopaic.fae.edu/cadernopaic/article/view/109/108. Acesso em: 15 set. 2022.

MANHIQUE, I. L. E.; CASARIN, H. de C. S.. Estrutura intelectual dos estudos da competência informacional na perspectiva fenomenográfica: uma análise por meio da citação e cocitação. *In:* **Revista Ibero-Americana de Ciência da Informação**, [S. l.], v. 11, n. 3, p. 751-768, 2018. DOI: https://doi.org/10.26512/rici.v11.n3.2018.10460.

MARQUES, F. B.; MARQUES, Y. B.; MACULAN, B. C. M. dos S.. Coocorrência de palavras-chave em dados abertos da Capes: teses e dissertações em Ciência da Informação. *In:* **Múltiplos Olhares em Ciência da Informação**, [S. l.], n. Especial, 2021. DOI: https://doi.org/10.35699/2237-6658.2021.37157.

MOKHTARPOUR, R; KHASSEH A. A.. Twenty-six years of LIS research focus and hot spots, 1990–2016: a co-word analysis. *In:* **Journal of Information Science**, v. 47, n. 6, p. 794-808, 2021. DOI: https://doi.org/10.1177/0165551520932119.

MOURA, M. F.. **Contribuições para a construção de taxonomias de tópicos em domínios restritos utilizando aprendizado estatístico.** 2009, 137 f. Tese (Doutorado em Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, SP, Biblioteca(s): Embrapa Agricultura Digital.

NAKANO, N.; MARTINEZ-AVILA, D.; JORENTE, M. J. V.; CANTISANI, M.. Information design, information science, and knowledge organization: a domain analysis from the perspective of complexity. *In:* **Scire**, v. 24, n. 1, p. 67-75, 2018. Disponível em: http://hdl.handle.net/11449/171160. Acesso em: 08 ago. 2022.

NICHOLSON, S.. The bibliomining process: data warehousing and data mining for library decision-making. *In:* **Transinformação**. v.16, n.3, p.253-261, set./dez. 2004. DOI: 10.1590/S0103-37862004000300005.

NOBRE, L. N.; FREITAS, R. R.. A evolução da Pós-graduação no Brasil: histórico, políticas e avaliação. *In:* **Brazilian Journal of Production Engineering**, [S. l.], v. 3, n. 2, p. 26–39, 2017. DOI: 10.0001/v3n2_3.

NOGUEIRA, E. T.; REIS, E. V. dos; OLIVEIRA, E. F. T. de. Tendências de pesquisa do GT-7: comparando citações em dois períodos. *In:* XX Encontro Nacional de Pesquisa em Ciência da Informação, 2019, Florianópolis. **Anais...**. Florianópolis: ENANCIB, 2019. v. XX. s.p. Disponível em:

https://conferencias.ufsc.br/index.php/enancib/2019/paper/view/938. Acesso em: 12 jan. 2021.

OLIVEIRA, E. F. T.. **Análise de Domínio em "Estudos Métricos" no Brasil:** produção, impacto e visibilidade em âmbito nacional e internacional. 2013. 193 f. Tese (Livre-Docência em Ciência da Informação). Faculdade de Filosofia e Ciências, Universidade Estadual Paulista, Marília-SP.

OLIVEIRA, E. F. T. de; GRÁCIO, M. C. C.. A produção científica em Organização e Representação do Conhecimento no Brasil: uma análise bibliométrica do GT-2 da ANCIB. *In:* X Encontro Nacional de Pesquisa em Ciência da Informação, 2009, João Pessoa. **Anais...**. João Pessoa: ENANCIB, 2009. v. X. p. 2037-2056.

ONU. Organização das Nações Unidas – Brasil. **Objetivos do desenvolvimento sustentável**. Disponível em: https://brasil.un.org/pt-br/sdgs. Acesso em: 28 fev. 2023.

PÊCHEUX, M.. **Análise automática do discurso**. *In:* GADET, F.; HAK, T. (Orgs.). Por uma análise automática do discurso: uma introdução à obra de Michel Pêcheux p. 61-161, 1997. Campinas: Editora da Unicamp.

PORTER, M.F.. An algorithm for suffix stripping. *In:* **Program: electronic library and information systems**, v. 14, n. 3, p. 130-137. 1980. DOI: https://doi.org/10.1108/eb046814.

PRIETO-DÍAZ, R.. Domain analysis: an introduction. *In:* **ACM SIGSoft Software Engineering Notes**, v. 15, n. 2, p. 47-54, 1990. DOI: https://doi.org/10.1145/382296.382703.

REGO-PIVA, L. M.; CASARIN, H. de C.; GUIMARÃES, J. A. C.. Análise de domínio como perspectiva metodológica para avaliação de periódicos científicos: o caso do BRAJIS. *In:* **Abec Meeting**, [S. l.], 2021. DOI: 10.21452/abecmeeting2021.38.

REIS, E. V.; TOMAÉL, M. I.. Infocomunicação: delineamento e representação. In: Marta Lígia Pomim Valentim; Cecilia Leite de Oliveira; Antonio Miranda. (Orgs.). **Gestão da informação, comunicação e tecnologia**. 1ed. Brasília: Universidade de Brasília, 2016, v. 1, p. 251-268.

ROSAS, F. S.; GRÁCIO, M. C. C.. Colaboração científica como procedimento para a análise de um domínio: uma aplicação na área de zootecnia. *In:* **Encontros Bibli:** revista eletrônica de biblioteconomia e ciência da informação, [S. l.], v. 20, n. 43, p. 115–132, 2015. DOI: 10.5007/1518-2924.2015v20n43p115.

ROSENBAUM, H.. Information use environments and structuration: towards an integration of Taylor and Giddens. *In:* BONZI, S. (Ed.), ASIS 93, **Proceedings of**

**the 56th ASIS annual meeting**, v. 30, p. 235-245, 1993. Medford, NJ: Americal Society for Information Science & Learned Information, Inc.

SAGER, J. C.. A Practical Course in Terminology Processing. Amsterdam, the Netherlands: John Benjamins, 1990.

SALAZAR-LÓPEZ, M. E.; VANIN, A. A.; CAZELLA, S. C.; LEVANDOWSKI, D. C.. Consequências na alimentação de crianças órfãs após a morte materna: uma investigação por meio de *software* de mineração de texto. *In:* **Cadernos de Saúde Pública**. v. 36, n. 3, 2020. DOI: 10.1590/0102-311X00189717.

SANT'ANA, R. C. G.. Tecnologias da informação e comunicação na ciência da informação: identificando dados. *In:* **BIBLOS**, [S. l.], v. 34, n. 2, 2020. Disponível em: https://periodicos.furg.br/biblos/article/view/12199. Acesso em: 12 ago. 2022.

SANTAREM, L. G. S.. Caracterização dos pesquisadores em tratamento temático da informação: um estudo da produção científica por meio da análise de domínio. *In:* XII Encontro Nacional de Pesquisa em Ciência da Informação, 2011, Brasília. **Anais...**. Brasília: ENANCIB, 2011. v. 12. p. 2341-2360. Disponível em: http://hdl.handle.net/20.500.11959/brapci/183376. Acesso em: 05 ago. 2022.

SARACEVIC, T.. Interdisciplinary nature of information science. *In:* **Ciência da Informação**, Brasília, v. 24, n. 1, p. 36-41, 1995. Disponível em: labds.eci.ufmg.br:8080/jspui/handle/123456789/75. Acesso em: 04 abr. 2022.

SÉRGIO, M. C.; SILVA, T. do N. da; GONÇALVES, A. L.. Descoberta de conhecimento a partir de informações não estruturadas por meio de técnicas de correlação e associação. *In:* **Em Questão**, Porto Alegre, v. 22, n. 2, p. 87–113, 2016. DOI: 10.19132/1808-5245222.87-113.

SMIRAGLIA, R. P.. Is FRBR A Domain? Domain Analysis applied to the Literature of The FRBR Family of Conceptual Models. *In:* **Knowledge Organization**, v. 40, n. 4, p. 273-282, 2013. DOI: https://doi.org/10.5771/0943-7444-2013-4-273.

SMIRAGLIA, R. P.. **Domain analysis for knowledge organization:** tools for ontology extraction Oxford: Chandos Publising, 2015.

STEYVERS, M.; GRIFFITHS, T. L.. Probabilistic topic models. *In:* LANDAUER, T. *et. al.* (Eds.), **Handbook of latent semantic analysis**, [*S.l.*], v. 427, n. 7, p. 424-440, 2007. Disponível em: https://cocosci.princeton.edu/tom/papers/SteyversGriffiths.pdf. Acesso em: 23 abr. 2022.

SUENAGA, C. M. K.; CERVANTES, B. M. N.. Discurso documental e representações sociais de domínios: uma perspectiva a partir da análise de domínio. *In:* XV Encontro Nacional de Pesquisa em Ciência da Informação, 2014, Belo Horizonte. **Anais...**. Belo Horizonte: ENANCIB, 2014. v. 15. p. 534-544.

Disponível em: http://hdl.handle.net/20.500.11959/brapci/190093. Acesso em: 10 ago. 2022.

TAN, A. H.. **Text Mining:** the state of the art and the challenges. Nanyang Technological University, p. 65-70, 1999. Singapore: Kent Ridge Digital Labs. Disponível em: https://docer.com.ar/doc/e1enxs1. Acesso em: 10 set. 2022.

TAYLOR, R. S.. Information use environments. *In:* DERVIN, B.; VOIGHT, M. J. (Eds.), **Progress in communication sciences**, v. 10, p. 217-255, 1991. Norwood, NJ: Ablex Publishing Corporation.

TAYLOR, R. N.. **A History of Software Engineering in ICS at UC Irvine.** Institute for *Software* Research. ICS2-221. University of California, Irvine-CA, 2018. Disponível em: https://isr.uci.edu/sites/isr.uci.edu/files/techreports/UCI-ISR-18-5.pdf. Acesso em: 05 jun. 2022.

TOGNOLI, N. B.; SILVA, A. M. S.; SILVA, A. P.. Organização do conhecimento e arquivologia: uma análise de domínio nos periódicos Knowledge Organization e Scire. *In:* **Informação & Informação**, v. 24, n. 3, p. 52-77, 2019. DOI: 10.5433/1981-8920.2019v24n3p52.

TRABADELA-ROBLES, J.; NUÑO-MORAL, M.-V.; GUERRERO-BOTE, V. P.; DE-MOYA-ANEGÓN, F.. Análisis de dominios científicos nacionales en Comunicación (Scopus, 2003-2018). *In:* **Profesional de la información**, [S. l.], v. 29, n. 4, 2020. DOI: 10.3145/epi.2020.jul.18.

URBIZAGASTEGUI-ALVARADO, R.. La bibliometría brasileña: minería de textos. *In:* **Revista ACB**, [S.l.], v. 26, n. 1, p. 1-18, jul. 2021. ISSN 1414-0594. Disponível em: https://revista.acbsc.org.br/racb/article/view/1768. Acesso em: 17 set. 2022.

VIEIRA, A. F. G; VIRGIL, J.. Uma revisão dos algoritmos de radicalização em língua portuguesa. *In:* **Information Research**, ISSN 1368-1613, v. 12, n. 3, 2007. Disponível em: http://informationr.net/ir/12-3/paper315.html. Acesso em: 09 set. 2022.

WEISS, S.; INDURKHYA, N.; ZHANG, T.; DAMERAU, F.. **Text Mining:** predictive methods for analyzing unstructured information. New York: Springer, 2005. 237 p.

WILSON, P.. Communication efficiency in research and development. *In:* **Journal of the American Society for Information Science**, v. 44, p. 376-382, 1993.

WITTEN, I. H.. Text mining. *In:* SINGH, M. P., editor. **Practical handbook of internet computing.** Boca Raton, FL: Chapman and Hall/CRC Press; p. 1–22. 2004. Disponível em: https://www.cs.waikato.ac.nz/~ihw/papers/04-IHW-Textmining.pdf. Acesso em: 10 ago. 2022.

WIVES, L. K.. **Utilizando conceitos como descritores de textos para o processo de identificação de conglomerados *(clustering)* de documentos.** Tese (Doutorado em Ciência da Computação), 2004. 126 f. Programa de Pós-graduação em Computação. Universidade Federal do Rio Grande do Sul, Porto Alegre-RS.

YANG, Y; PEDERSEN, J. O.. A comparative study on feature selection in text categorization. *In:* ICML ’97: **Proceedings of the fourteenth International Conference on Machine Learning**, p. 412–420, 1997. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

ZHONG, S; GHOSH, J.. A unified framework for model-based clustering. *In:* **Journal of Machine Learning Reseach**, v. 4, p. 1001–1037, 2003. Disponível em: https://www.jmlr.org/papers/volume4/zhong03a/zhong03a.pdf. Acesso em: 07 ago. 2022.