UNIVERSIDADE ESTADUAL PAULISTA “JÚLIO DE MESQUITA FILHO”
FACULDADE DE FILOSOFIA E CIÊNCIAS - CAMPUS DE MARÍLIA
DEPARTAMENTO DE PÓS-GRADUAÇÃO EM CIÊNCIA DA INFORMAÇÃO

FABIO PIOLA NAVARRO

USO DA INTELIGÊNCIA ARTIFICIAL PARA RECUPERAÇÃO DA INFORMAÇÃO COM ABORDAGEM SEMÂNTICA: MODELO DE APLICAÇÃO PARA DOCUMENTOS TEXTUAIS EM AMBIENTES DIGITAIS

MARÍLIA
2021

FABIO PIOLA NAVARRO

USO DA INTELIGÊNCIA ARTIFICIAL PARA RECUPERAÇÃO DA INFORMAÇÃO COM ABORDAGEM SEMÂNTICA: MODELO DE APLICAÇÃO PARA DOCUMENTOS TEXTUAIS EM AMBIENTES DIGITAIS

Tese apresentada ao Programa de Pós-Graduação em Ciência da Informação, da Universidade Estadual Paulista – Campus de Marília, como requisito para a obtenção do título de doutor em Ciência da Informação.

Área de Concentração: Informação, Tecnologia e Conhecimento.

Linha de Pesquisa: Informação e Tecnologia

Orientador: Dr. José Eduardo Santarém Segundo

MARÍLIA
2021

N322u

Navarro, Fabio Piola

Uso da inteligência artificial para recuperação da informação com abordagem semântica : modelo de aplicação para documentos textuais em ambientes digitais / Fabio Piola Navarro. -- Marília, 2021

111 p. : il., tabs.

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

Orientador: Dr. José Eduardo Santarém Segundo

1. Recuperação da informação. 2. Inteligência artificial. 3. Incorporação de palavras. 4. Latent Dirichlet Allocation. 5. Modelos Neurais. I. Título.

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

Essa ficha não pode ser modificada.

FABIO PIOLA NAVARRO

USO DA INTELIGÊNCIA ARTIFICIAL PARA RECUPERAÇÃO DA INFORMAÇÃO COM ABORDAGEM SEMÂNTICA: MODELO DE APLICAÇÃO PARA DOCUMENTOS TEXTUAIS EM AMBIENTES DIGITAIS

Local: Universidade Estadual Paulista
Faculdade de Filosofia e Ciências
UNESP - Campus de Marília

BANCA EXAMINADORA:

_____

Presidente e Orientador: Prof. Dr. José Eduardo Santarém Segundo
Docente no Programa de Pós-Graduação em Ciência da Informação da Universidade Estadual Paulista Júlio de Mesquita Filho, UNESP Campus de Marília.
Docente da Universidade de São Paulo, USP. Campus de Ribeirão Preto.

_____

Membro Titular: Profa. Dra. Silvana Aparecida Borsetti Gregorio Vidotti
Docente no Programa de Pós-Graduação em Ciência da Informação da Universidade Estadual Paulista Júlio de Mesquita Filho, UNESP Campus de Marília.

_____

Membro Titular: Prof. Dr. Leonardo Castro Botega
Docente no Programa de Pós-Graduação em Ciência da Informação da Universidade Estadual Paulista Júlio de Mesquita Filho, UNESP Campus de Marília.

_____

Membro Titular Externo: Prof. Dr. Elvis Fusco
Presidente executivo ASSERTI
(Associação de Empresas de Serviços de Tecnologia da Informação)

_____

Membro Titular Externo: Prof. Dr. Douglas Dyllon Jeronimo de Macedo
Docente no Programa de Pós-Graduação em Ciência da Informação (PPGCin) e do Departamento de Ciência da Informação (CIN) da Universidade Federal de Santa Catarina (UFSC)

MARÍLIA
2021

Dedico à toda minha família.

AGRADECIMENTOS

À Carolina, minha esposa e parceira para as aventuras da vida, por seu amor e companheirismo diário, por seus recados com palavras de carinho, por seus chás que traziam muito mais que o sabor, por ser uma pessoa forte, guerreira e realizadora, a qual tenho o prazer de compartilhar a vida e me espelhar.

Aos meus pais, Maria Cristina Piola Navarro e Nivaldo Cincotto Navarro, por seu amor incondicional e apoio em tudo. Pai e Mãe, vocês são minha base, me ensinaram e ensinam as coisas mais importantes da vida e agradeço sempre por tudo que fizeram e fazem por mim e por me permitirem estudar.

Aos meus irmãos Fernando e Vivian, que indiretamente me permitiram estar longe de casa por muitos anos e cuidaram de tudo por mim. Fer e Vi muito obrigado sempre!!!

Aos meus familiares que sempre me ajudaram, minha sogra Zenaide, meu sogro Álvaro, minhas cunhadas e cunhados, sobrinho e sobrinhas, tios e tias, primos e primas e avós.

Ao professor José Eduardo Santarém Segundo, que acreditou em mim e me orientou durante o doutorado, me passando seu conhecimento de forma muito natural, como de costume pela sua excelente didática, minha admiração por toda a sua carreira.

Ao Prof. Elvis Fusco, que me deu oportunidades tanto na área acadêmica e quanto fora dela, meu agradecimento e estima. Agradeço imensamente suas contribuições.

À Profa. Silvana Aparecida Borsetti Gregorio Vidotti, que me trouxe luz às áreas da Ciência da Informação com seu conhecimento e facilidade de compartilhá-lo. Agradeço imensamente suas contribuições.

Ao Prof. Leonardo Castro Botega, por ter aceito o convite de ser parte da minha banca e por ser meu colega de trabalho que me escutou e deu alguns conselhos e me ajudou em relação a esta tese e ao doutorado como um todo.

Ao Prof. Douglas Dyllon Jeronimo de Macedo, por ter aceito o convite de ser parte da minha banca e contribuir com os seus conhecimentos.

Aos meus amigos e colegas, em especial, Caio Monteiro, Caio Coneglian, Fábio Dacêncio Pereira, Rodolfo Chiaramonte e Leonardo Botega, que participaram e ajudaram na minha carreira profissional.

Aos demais professores da Unesp, em especial, Marta Valentim e Ricardo Sant'Ana, pelo aprendizado.

À Deus, por tudo.

“Não farás uma máquina à semelhança de uma mente humana.”
Dune

"Reconhecerás uma palavra pela companhia que ela mantém."
(FIRTH, 1957)

RESUMO

A quantidade de documentos digitais tem aumentado significativamente ultimamente e a necessidade de armazenamento e recuperação destes levou ao uso de sistemas computacionais como solução para acesso e disseminação da informação. A inteligência artificial, em suas variadas subáreas vem sendo pesquisada como aplicação para a evolução de diversos processos da ciência da informação e em especial para recuperação da informação. Uma das subáreas da inteligência artificial, o processamento de linguagem natural, tem tido repercussão nos ambientes de pesquisa como meio para melhoria dos processos de recuperação da informação nestes sistemas computacionais. Há um conjunto de discussões para melhorias das abordagens semânticas utilizando-se da inteligência artificial e suas tecnologias e técnicas para recuperação da informação e melhor aproveitamento para a sociedade dos conteúdos produzidos nestes sistemas computacionais. Este cenário contribuiu para o interesse em averiguar se modelos de aprendizado de máquina, em especial modelos neurais aplicados ao processamento de linguagem natural podem contribuir com a proposta de um modelo para recuperação da informação com abordagem semântica nestes sistemas computacionais. O objetivo geral deste trabalho está em ampliar a capacidade de Recuperação da Informação com abordagem semântica em ambientes digitais de documentos textuais, para que o usuário, em suas necessidades informacionais, possa expandir o acesso à informação contextual nestes ambientes. Para isso, foi criado um modelo para recuperação da informação com abordagem semântica utilizando técnicas de inteligência artificial em conjunto com processos e algoritmos do processamento de linguagem natural. Além disso, foi realizada pesquisa bibliográfica visando identificar quais as técnicas ligadas ao processo de recuperação da informação em sistemas computacionais utilizando inteligência artificial. A partir dos resultados desta pesquisa, foi realizada uma análise do processamento de linguagem natural usando inteligência artificial para recuperação da informação e foram usadas técnicas como incorporação de palavras, *topic modeling* e algoritmos como Word2Vec e *Latent Dirichlet Allocation* – LDA, que foram aplicados à proposta do modelo de forma integrada. Como resultado, esta tese apresenta a viabilidade do modelo proposto que ao utilizar das técnicas mencionadas de forma integrada e do processo de expansão contextual com o algoritmo Word2Vec possibilitará aos usuários um processo de recuperação da informação com abordagem semântica.

Palavras-chave: Recuperação da Informação. Inteligência Artificial. Incorporação de Palavras. Latent Dirichlet Allocation. Modelos Neurais. Word2Vec.

ABSTRACT

The amount of digital documents has increased significantly lately and the need to store and retrieve them has led to the use of computer systems as a solution for accessing and disseminating information. Artificial intelligence, in its various sub-areas, has been researched as an application for the evolution of several information science processes and in particular for information retrieval. One of the subareas of artificial intelligence, natural language processing, has had repercussions in research environments as a means to improve the information retrieval processes in these computer systems. There is a set of discussions to improve semantic approaches using artificial intelligence and its technologies and techniques for information retrieval and better use for society of the content produced in these computer systems. This scenario contributed to the interest in investigating whether machine learning models, especially neural models applied to natural language processing can contribute to the proposal of a model for information retrieval with a semantic approach in these computer systems. The general objective of this work is to expand the capacity of Information Retrieval with a semantic approach in digital environments of textual documents, so that the user, in his informational needs, can expand access to contextual information in these environments. For this, a model for information retrieval with a semantic approach was created using artificial intelligence techniques in conjunction with natural language processing processes and algorithms. In addition, bibliographic research was carried out to identify which techniques are linked to the information retrieval process in computer systems using artificial intelligence. From the results of this research, an analysis of natural language processing using artificial intelligence for information retrieval was performed and techniques such as word incorporation, topic modeling and algorithms such as word2vec and Latent Dirichlet Allocation - LDA were used, which were applied to the proposal. of the model in an integrated way. As a result, this thesis presents the feasibility of the proposed model that by using the techniques mentioned in an integrated manner and the contextual expansion process with the word2vec algorithm will allow users to process information retrieval with a semantic approach.

Keywords: Information Retrieval. Artificial intelligence. Word Incorporation. Latent Dirichlet Allocation. Neural Models. Word2Vec.

LISTA DE FIGURAS

LISTA DE QUADROS

SUMÁRIO

1 INTRODUÇÃO

No final do século XX e início do século XXI a capacidade de geração de dados e tratamento dos mesmos por parte de recursos tecnológicos cresceu de forma nunca vista. A troca de informação constante e ininterrupta fez necessário meios de armazenamento desta grande quantidade de informação. Um dos grandes desafios a serem tratados pela ciência de modo geral neste período pós-moderno é como recuperar, analisar e estruturar a grande quantidade, variedade e velocidade de dados.

A Ciência da Informação tem um tratamento cuidadoso para dado e informação, segundo Santos e Sant'Ana (2015) há uma pluralidade de conceitos sobre dados e para os autores a definição de dado é colocada levando-se em conta uma preocupação com a interface entre ciência da informação, comunicação e ciência da computação. Neste contexto, esta tese seguirá o conceito definido no artigo supra citado de que dado é o elemento básico dos fluxos informacionais que não contém intrinsicamente um componente semântico.

Neste sentido, tratar informação e dado de uma mesma forma é um equívoco, e apontar a recuperação da informação como sendo o mesmo processo para recuperação de dados também não condiz. Por isso, ao tratar-se de semântica na recuperação da informação, deve-se entender que esta, portanto, está muito mais ligada à informação e não ao dado, porém, processar os dados é matéria-prima para se obter informação.

O termo Web Semântica foi introduzido pelo pesquisador Tim Berners Lee em 2001, no seu artigo intitulado: A Web Semântica, no qual expõe suas ideias sobre tornar a Web estruturada de tal forma que se possa dar significado para os conteúdos e assim permitir que agentes computacionais possam entregar ao usuário maior qualidade nas respostas (BERNERS-LEE; LASSILA; HENDLER, 2001).

Esta tese tenciona que as ideias de Berners-Lee, Lassila e Hendler (2001) podem ser extrapoladas para ambientes mais controlados, como repositórios digitais de documentos, e esses agentes computacionais, ao serem forjados utilizando-se do processamento de linguagem natural e técnicas de aprendizado de máquina, possam entregar aos usuários maior qualidade na recuperação da informação em repositórios digitais.

Segundo Ávila, Silva e Cavalcante (2017), os repositórios digitais surgiram como resposta à industrialização da comunidade científica e tem crescido significativamente nos últimos anos. Resultados experimentais revelam que, respectivamente, 52,0% e 86,9% da graduação e da pós-graduação utilizam os repositórios para algum propósito e 39,3% e 78,8%, respectivamente, os usam como forma de busca por conhecimento acadêmico.

De acordo com Sanchez, Vechiato e Vidotti (2019), vários são os aparatos computacionais para armazenamento de dados, no entanto, os repositórios digitais promovem armazenamento, organização, disseminação, preservação e recuperação de dados de pesquisa. Estes repositórios são formados em sua essência por documentos em formato de texto, consequentemente, dados não estruturados.

Prover uma maior capacidade informacional aos usuários destes ambientes é necessidade premente, permitindo-lhes uma recuperação da informação mais assertiva, isto é, propiciar uma recuperação da informação com abordagem semântica. No entanto, devido à grande quantidade de dados nestes repositórios, tal recuperação incita técnicas mais complexas que possam refletir a verdadeira percepção dos fatos ora registrados.

O processamento e análise de dados por agentes computacionais tem papel fundamental na busca de informação, mas recuperar informação de forma semântica, ou seja, segundo um contexto, com presença de termos de um determinado domínio não é tarefa trivial, uma vez que estes agentes deverão conhecer o *corpus* ao que estão sendo submetidos à análise.

Tecnologias mais recentes no campo da inteligência artificial ou mais especificamente em um dos seus ramos de estudo voltados aos dados textuais não estruturados, como o processamento de linguagem natural (PLN), tornam-se, naturalmente, uma escolha como agente computacional para conhecimento de um documento ou até mesmo de um *corpus*.

O Programa de Pós-Graduação em Ciência da Informação (PPGCI) da Universidade Estadual Paulista “Júlio de Mesquita Filho” (UNESP), Campus de Marília, na linha de pesquisa “Informação e Tecnologia” realiza pesquisas e estudos teóricos, epistemológicos e práticos relacionados à recuperação, à transferência, à visualização, ao design, à arquitetura, à utilização, à gestão e à preservação de dados, informação e

de documentos em ambientes digitais, armazenados em espaços ou sistemas informacionais tecnológicos.

Esta tese é resultado de estudos no PPGCI e apresenta-se posicionada no contexto da Recuperação da Informação, tendo como objeto os repositórios digitais. Ao mesmo tempo, utiliza da interdisciplinaridade com outras ciências, como a ciência da computação e linguística no contexto da pesquisa para melhoria da Recuperação de Informações em repositórios digitais.

A Ciência da Informação tem seu prenúncio de criação como ciência em 1948, quando a obra de Norbert Wiener, *Cybernetics or control and communication in the animal and machine*, e, contextualmente, o livro *The mathematical theory of communication*, de Claude Shannon e Warren Weaver (PINHEIRO; LOUREIRO, 1995).

Fatos da década de 60 como a conferência realizada no *Georgia Institute of Technology* em 1962, o Relatório Weinberg em 1963, o trabalho Informática de Mikhailov, em 1966, o estudo de Rees e Saracevic (1967) e, especialmente, o artigo de Borko (1968) *Information Science: what is it?* marcam fortemente o período e delimitam historicamente a CI, mostrando que a interferência positiva da computação é benéfica para automação de processos.

A Ciência da Informação, como ciência, deve ser o arcabouço intelectual para entender, relacionar, racionalizar e prover meios para resolver problemas que envolvem interdisciplinaridade dos campos da Computação e da Ciência da Informação. No entanto, tal desafio não pode ser alcançado sem ter uma visão holística e complexa dos problemas relacionados à quantidade de dados, sua recuperação e utilização para proveito do Homem.

Conceitos e aportes de outras ciências coirmãs da Ciência da Informação, como a Ciência da Computação, Matemática e Linguística formam a base epistemológica para o desenvolvimento deste trabalho, tendo como cerne algoritmos e técnicas da Computação para solução de problemas pertinentes à Recuperação da Informação dentro da Ciência da Informação.

Esta interdisciplinaridade da Ciência da Informação (CI) com a Ciência da Computação (CC), no sentido de utilizar recursos tecnológicos da CC aliados aos conceitos da CI, representa uma maior capacidade e uma ampliação de como a

recuperação de informação (RI) pode solucionar desafios, como o conhecimento da grande massa de dados a ser analisada por agentes computacionais inteligentes e dessa forma melhorar a assertividade na RI.

Segundo Souza (2006), a Recuperação da Informação traz dificuldades intrínsecas ao conceito de informação, seja com a real necessidade do usuário ou com os documentos que fazem parte do acervo digital. A utilização dos agentes computacionais fornece maior capacidade de análise dos Sistemas de Recuperação da Informação por se utilizar de recursos tecnológicos.

A Ciência da Informação, como definido por Saracevic (1995), faz aportes de outras ciências, como a ciência da computação, pois a recuperação de dados realizado pelos agentes computacionais é passo essencial para a recuperação da informação. Nesse sentido, Sant'ana (2013) pontua que a recuperação de dados é uma das fases do ciclo de vida dos dados (as quais são: coleta, recuperação, armazenamento, descarte) e possui diferença da recuperação da informação. Na recuperação da informação não há uma preocupação com a unicidade dos dados retornados, mas sim com uma miríade de dados que já sofreram alguma interferência e que não estão mais em sua essência.

Assim, Sant'ana (2013) proporciona uma base de análise para corroborar a ideia que, para haver uma recuperação de informação, deve haver antes uma fase de recuperação de dados, ou seja, de alguma forma deve-se haver uma coleta e recuperação de dados que, no caso deste trabalho, serão textos não estruturados, e deve-se processá-los para que posteriormente se possa fazer a Recuperação da Informação (RI).

Desta forma, esta tese propicia um alinhamento interdisciplinar entre as ciências da Informação, da Computação e da Linguística junto à área da Recuperação da Informação em sistemas informacionais tecnológicos e documentos digitais, ao passo que utiliza de elementos das outras ciências, como as acima descritas de forma complementar, como arcabouço tecnológico e conceitual no âmbito da melhoria do processo de Recuperação da Informação.

1.1 PROBLEMA DE PESQUISA

De acordo com Santarém Segundo (2010), a informação em repositórios digitais tem por característica ter sido previamente armazenada e registrada de forma adequada em arquivos, seguindo padrões de catalogação e uso de metadados sobre estes arquivos para que a recuperação de informação seja realizada de forma mais assertiva.

Apesar de ter sido armazenada seguindo os padrões de catalogação e uso de metadados, os processos para recuperação desta informação em repositórios digitais se baseiam em técnicas sintáticas que permitem aos usuários realizar de fato a recuperação da informação, o que delimita e limita informação de valor aos usuários.

> Apesar de a estrutura de armazenamento sugerir um tipo de recuperação mais apropriado ao usuário, ela continua sendo feita de forma sintática, buscando, dentro do conjunto de informações armazenadas, palavras que tenham mesma grafia, e utilizando a técnica baseada no modelo booleano e na teoria de conjuntos, possibilitando apenas o cruzamento de elementos da estrutura na busca de informação (SANTARÉM SEGUNDO, 2010, p. 163).

Extrapola-se que não só os repositórios digitais não possuem capacidade para recuperação da informação com um contexto semântico para o usuário, mas também, revistas digitais, bases de artigos, bases de GED (Gerenciamento Eletrônico de Documentos) entre outros vários sistemas computacionais baseados em documentos textuais digitais.

Neste sentido, esta pesquisa busca solução para o seguinte **problema**: a recuperação da informação quando baseada em documentos textuais em sistemas computacionais está limitada a regras básicas sobre os termos digitados, como o uso de operadores booleanos e delimitação da busca pelo usuário, ou seja, a recuperação da informação é feita de forma sintática e não implementam recuperação da informação com abordagem semântica.

1.2 TESE, HIPÓTESE E PROPOSIÇÃO DE PESQUISA

Baseado neste contexto, a seguinte **tese** será defendida: é possível que, baseado em inteligência artificial, seja desenvolvido um modelo de recuperação da informação com abordagem semântica, de forma que os ambientes digitais possam disponibilizar esse tipo de recuperação a seus usuários, enriquecendo e ampliando as possibilidades do modelo sintático já amplamente conhecido e utilizado.

A **hipótese** desta pesquisa é a de que ao se utilizar o aprendizado de máquina para maior reconhecimento do grande volume de textos em um *corpus*, e ao mesmo tempo, utilizar incorporação de palavras através de algoritmos de inteligência artificial para propiciar similaridade semântica, será possível aos ambientes digitais, de documentos textuais, disponibilizar recuperação com abordagem semântica a seus usuários.

Portanto, a **proposta** desta pesquisa é estabelecer um modelo que aplique os conceitos do aprendizado de máquina para permitir uma otimização na recuperação da informação em sistemas computacionais baseados em documentos. Este modelo fará uso de algoritmos para que o processamento da busca seja contextual, ou seja, levando-se em conta o conceito de similaridade semântica entre palavras de um mesmo contexto.

A similaridade semântica é um artefato computacional que permite uma comparação das palavras que estão em um mesmo contexto dentro de um documento. Ao mesmo tempo permite analisar, de maneira global um corpus, verificando quais documentos possuem maior congruência em relação ao outro.

1.3 OBJETIVOS

Este trabalho tem como objetivo geral ampliar a capacidade de Recuperação da Informação em sistemas computacionais baseados em documentos textuais. Para isso, foi criado um modelo para recuperação da informação baseado em técnicas de inteligência artificial em conjunto com processos de aprendizado de máquina e algoritmos de processamento de linguagem natural.

Evidencia-se que este trabalho tem foco no processo interno e funcional da Recuperação da Informação, ou seja, no processo de enriquecimento semântico localizado anteriormente à interação do usuário. Este processo tem caráter autônomo e automático a fim de realizar as tarefas de recuperação de informação sem a necessidade de interação do usuário.

Embora o cumprimento deste objetivo permita consequências benéficas ao usuário, faz-se necessário uma avaliação quantitativa e qualitativa deste processo, etapa que será desenvolvida em um próximo passo do trabalho.

Neste contexto, estabelecido o objetivo geral, para atingi-lo apresenta-se os seguintes objetivos específicos:

- Analisar os processos de recuperação da informação de forma geral;
- Avaliar os processos de recuperação no contexto da inteligência artificial;
- Aprofundar sobre o conceito de incorporação de palavras e demais conceitos chave para a melhoria na recuperação da informação como: algoritmos, tecnologias e conceitos matemáticos envolvidos;
- Desenvolver uma arquitetura que abarque *softwares* e bibliotecas pertinentes à área de Processamento de Linguagem Natural para comprovação da utilização na análise dos documentos em repositórios;
- Apurar quais as técnicas de Processamento de Linguagem Natural no sentido de conhecer e avaliar os *frameworks* utilizados para processamento de linguagem natural;
- Detalhar de forma prática o algoritmo *Latent Dirichlet Allocation* (LDA) como analisador dos documentos para redução de dimensionalidade, conceito importante para capacidade de processamento;
- Utilizar o modelo construído como prova de conceito da melhoria da Recuperação da Informação em sistemas computacionais de documentos digitais textuais através de um *script* (uma execução de uma programação) a fim de avaliar a recuperação contextual da informação.

1.4 METODOLOGIA

Segundo Valentim (2005), a pesquisa na área das ciências sociais aplicadas são plurais e aceitam diferentes tipos de pesquisa. Ainda segundo a autora, existem vários tipos e possibilidades de pesquisa científica e o pesquisador opta pelo tipo que melhor responder às questões do problema de pesquisa.

De acordo com Gil (2002), a pesquisa pode ser organizada como pura e aplicada, sendo que estas não são mutuamente exclusivas, mas sim o contrário em que uma abordagem apoia a outra para construção de avanços científicos e técnicos.

O presente trabalho se caracteriza como uma pesquisa **qualitativa** quanto à sua

natureza, pois pretende gerar conhecimentos para aplicação prática para uso da sociedade ao tratar fenômenos complexos e do mundo real.

Esta pesquisa também se caracteriza como sendo **exploratória**, pois tem a finalidade de apresentar informações sobre o assunto e entendê-lo melhor. Como esta pesquisa gera aplicação prática dirigida à solução de problemas também se caracteriza como sendo **aplicada**.

A metodologia está explicitada no capítulo 2 com detalhes de bases, indexadores, período de abrangência, idiomas, termos de busca, resultados encontrados e critérios de seleção dos materiais bibliográficos utilizados.

1.5 JUSTIFICATIVA

Atualmente grande parte da informação está armazenada em espaços ou sistemas informacionais tecnológicos que mantém dados sobre estudos e pesquisas na forma de documentos digitais que representam um determinado domínio do conhecimento.

Evoluir os meios para recuperação de informação nos ambientes informacionais é essencial para permitir à sociedade esta faculdade de pesquisa, análise, reflexão e tomada de decisão sobre determinados domínios do conhecimento. A discussão e reflexão sobre outros meios no processo de busca nos aparatos tecnológicos de armazenamento de informação incita novas possibilidades de melhoria na recuperação da informação.

Neste sentido, de um prisma acadêmico, este trabalho busca trazer relevância ao consolidar conceitos da Web Semântica, como agentes computacionais aplicados ao processo de Recuperação da Informação em repositórios digitais de documentos, e ao mesmo tempo traz o aporte de conceitos e técnicas da área da Ciência da Computação no âmbito do processamento de linguagem natural ou, de forma mais abrangente, a utilização da inteligência artificial e seus desdobramentos técnicos.

Este trabalho imprime relevância científica e tecnológica ao trazer a proposta de um modelo para melhoria da Recuperação da Informação em repositórios digitais utilizando a técnica de *Topic Modeling* baseada no algoritmo LDA. Ao se utilizar de Inteligência Artificial e Incorporação de Palavras esta conexão traz inovação para melhor

assertividade na Recuperação de Informação, uma vez que possibilita a recuperação da informação com abordagem semântica e desta forma contribui com aspectos culturais e sociais em relação ao conhecimento armazenado em repositórios digitais.

Em resumo, esta tese de doutorado justifica-se ao aventar o uso de estruturas, algoritmos e conceitos da inteligência artificial no âmbito da Ciência da Informação e mais especificamente da recuperação da informação em sistemas computacionais baseados em documentos textuais. Através do processamento de linguagem natural, permitir amplo conhecimento do *corpus* analisado e desta forma habilitar recuperação da informação semântica e contextual com o objetivo de propiciar a implementação deste tipo de tecnologia, favorecendo, posteriormente, a compreensão do usuário acerca de um determinado corpus a partir de suas buscas nestes ambientes

Dessa forma, a importância e a motivação em propiciar uma melhor recuperação da informação em sistemas computacionais de documentos textuais está diretamente ligada a retornar conteúdos mais pertinentes para os usuários e, desta maneira, um melhor aproveitamento das informações para a sociedade.

1.6 ESTRUTURA DO TRABALHO

Além seção corrente, a tese tem sua estrutura em sete seções como descrito a seguir:

Na seção 2, traz-se a metodologia e faz-se uma revisão bibliográfica, trazendo estudos relacionados a esta pesquisa e principalmente como estes trabalhos estão utilizando os conceitos relevantes para essa pesquisa, como Topic Modeling, LDA e Word2Vec.

Na seção 3, faz-se uma revisão sobre Recuperação da Informação levando-se em conta seus tipos clássicos, mas também abrindo a possibilidade de inovações na área de recuperação da informação através da expansão contextual.

A seção 4, contém conceitos sobre Inteligência Artificial e como esta atua no contexto da Recuperação da Informação, além de conceitos sobre processamento de linguagem natural, hipótese distributiva, semântica vetorial, incorporação de palavras e o

algoritmo word2vec e LDA para embasar conceitos e técnicas da área de inteligência artificial aplicadas a textos ou ainda dados não estruturados.

A seção 5 apresenta o modelo proposto para melhoria da recuperação da informação em repositórios digitais usando similaridade por incorporação de palavras, este apresenta como o método LDA compõe a fase de preparação para busca no repositório e como o Word2Vec possibilita a análise de termos similares semanticamente.

Na seção 6 serão apresentadas as considerações finais e trabalhos futuros e as referências a seguir.

2 METODOLOGIA E TRABALHOS CORRELATOS

A pesquisa bibliográfica foi feita como revisão de literatura nas duas grandes ciências que formatam este trabalho, ciência da computação e ciência da informação, em bases de dados nacionais e internacionais como: BRAPCI, SCOPUS e LISA com os termos de pesquisa definidos no quadro 1 e com um recorte temporal de 2000 a 2019, nos idiomas português e inglês.

Quadro 1 - Termos de busca

Termos em português:
Recuperação da Informação, repositórios digitais, incorporação de palavras, LDA, Latent Dirichlet Allocation, modelagem de tópicos, processamento de linguagem natural
Termos em inglês:
Information Retrieval, digital repositories, word embeddings, LDA, Latent Dirichlet Allocation, topic modeling, natural language processing

Fonte: elaborado pelo autor (2021)

A pesquisa foi realizada em cada base de dados descrita no parágrafo anterior e os quadros 2, 3 e 4 possuem respectivamente a *Query* (apontando para título, palavras-chave e resumo) com os termos de pesquisa e o número de documentos retornados, bem como a data da pesquisa.

Quadro 2 - Base SCOPUS, data da pesquisa 06 de julho de 2020

Query/Termos	**Número de Documentos**
information AND retrieval OR digital AND repositories OR word AND embeddings OR topic AND modeling OR natural AND language AND processing AND SUBJAREA (arts OR busi OR deci OR econ OR psyc OR soci) OR SUBJAREA (comp)	726
TITLE-ABS-KEY (information AND retrieval OR digital AND repositories OR word AND embeddings OR topic AND modeling OR natural AND language AND processing) AND SUBJAREA (arts OR busi OR deci OR econ OR psyc OR soci)	139
TITLE-ABS-KEY (information AND retrieval OR digital AND repositories OR word AND embeddings OR topic AND modeling) AND SUBJAREA (arts OR busi OR deci OR econ OR psyc OR soci) OR SUBJAREA (comp)	660

TITLE-ABS-KEY (information AND retrieval OR digital AND repositories OR word AND embeddings OR topic AND modeling) AND SUBJAREA (arts OR busi OR deci OR econ OR psyc OR soci)	134
TITLE-ABS-KEY (information AND retrieval OR digital AND repositories OR word AND embeddings) AND SUBJAREA (arts OR busi OR deci OR econ OR psyc OR soci) OR SUBJAREA (comp)	1128
TITLE-ABS-KEY (information AND retrieval OR digital AND repositories OR word AND embeddings) AND SUBJAREA (arts OR busi OR deci OR econ OR psyc OR soci)	200
TITLE-ABS-KEY (information AND retrieval AND in AND digital AND repositories) AND SUBJAREA (arts OR busi OR deci OR econ OR psyc OR soci) OR SUBJAREA (comp)	660
TITLE-ABS-KEY (information AND retrieval AND in AND digital AND repositories) AND SUBJAREA (arts OR busi OR deci OR econ OR psyc OR soci)	222

Fonte: elaborado pelo autor (2021)

Na base SCOPUS, quando a pesquisa foi feita adicionando-se a área da Ciência da Computação, o número de documentos retornados foi de 726. No entanto, quando a busca foi restringida para a área de Ciência da Informação, o número de documentos caiu para 139, ou seja, uma diminuição drástica para estes termos de pesquisa dentro da CI.

Quadro 3 - Base BRAPCI, data da pesquisa 07 de julho de 2020

Query/Termos	**Número de Documentos**
Recuperação da Informação	1122
Repositórios digitais	241
Recuperação da Informação, repositórios digitais	177
Recuperação da Informação, repositórios digitais, topic modeling	0
Topic modeling	6
Recuperação da Informação, Topic Modeling	0
Incorporação de palavras	2
Recuperação da Informação, incorporação de palavras	0

LDA, Latent Dirichlet Allocation	2
Recuperação da Informação, LDA, Latent Dirichlet Allocation	0
Modelagem de tópicos	3
Processamento de linguagem natural	29
Repositórios digitais, LDA, Latent Dirichlet Allocation, recuperação da informação, modelagem de tópicos, processamento de linguagem natural	**0**
LDA[1]	7

Fonte: elaborado pelo autor (2021)

As pesquisas denotam que em bases nacionais (BRAPCI) os termos parecem ainda ser novos e não possuem tantos documentos em seus retornos. Termos como processamento de linguagem natural, modelagem de tópicos, *topic modeling* e incorporação de palavras, quando buscados de forma separada tem melhores retornos, mas com uma quantidade de documentos baixa. Quando buscados em conjunto, conforme quadro 4, mostra-se que esta quantidade é pequena quando não, inexistente.

Quadro 4 - Base LISA, data da pesquisa 07 de julho de 2020

Query/Termos	**Número de Documentos**
Recuperação da Informação	292
Repositórios digitais	102
Recuperação da Informação,repositórios digitais	62
Recuperação da Informação,repositórios digitais, topic modeling	0
Topic modeling	6553
Recuperação da Informação, Topic Modeling	1
Word embeddings	100
Word embeddings information retrieval	45

[1] Artigo de autoria própria: NAVARRO, Fabio Piola; CONEGLIAN, Caio Saraiva; SANTARÉM SEGUNDO, José Eduardo. Big Data no contexto de dados acadêmicos: o uso de Machine Learning na construção de sistema de organização do conhecimento. **Informação & Tecnologia**, n. 2, v. 5, p. 181-200, 2018.

LDA, Latent Dirichlet Allocation	166
LDA, Latent Dirichlet Allocation, information retrieval	85
LDA, Latent Dirichlet Allocation, information retrieval, topic modeling	67
Information retrieval, LDA, Latent Dirichlet Allocation, information retrieval, topic modeling, Natural language processing	37
LDA, Latent Dirichlet Allocation, information retrieval, topic modeling, Natural language processing	37
Digital repository, LDA, Latent Dirichlet Allocation, information retrieval, topic modeling, Natural language processing	**10**

Fonte: Elaboração do autor (2021)

A pesquisa bibliográfica na base LISA mostrou-se mais equilibrada entre os termos de busca e por ser uma base voltada para a Ciência da Informação (foram utilizadas as bases *Library & Information Science Abstracts (LISA) e Applied Social Sciences Index & Abstracts (ASSIA)*) mostrou-se promissora para a pesquisa deste trabalho, pois mesmo que os retornos não sejam com uma quantidade tão expressiva (10), estes mostraram que existem trabalhos na área e que esta tese de doutorado pode contribuir.

2.1 DISCUSSÃO DOS TRABALHOS

Esta seção apresenta os trabalhos relacionados ao tema de pesquisa desta tese, são explicitados e discutidos os trabalhos nacionais e internacionais que mais avançaram no âmbito desta pesquisa em relação aos termos como LDA e *Topic Modeling*.

De acordo com a pesquisa bibliográfica na base de dados LISA (ver quadro 4), especificamente, nas bases *Library & Information Science Abstracts (LISA) e Applied Social Sciences Index & Abstracts (ASSIA)*, os artigos mais pertinentes segundo *query* de pesquisa (*Digital repository, LDA, Latent Dirichlet Allocation, information retrieval, topic modeling, Natural language processing*), a qual reúne as palavras-chaves sobre esta tese estão condensados no quadro 5.

Quadro 5 - Trabalhos mais relevantes para comparação com a tese

Trabalhos	Word Embeddings Digital repository	LDA	Information retrieval	Topic modeling	Natural language processing	Word2vec
ABU-SALIH; PORNPIT; CHAN (2018)	Não	Não	Sim	Não	Sim	Não
AHMAD *et al.* (2018)	Não	Parcial	Sim	Parcial	Sim	Não
ARUN; GOVINDAN; MADHU (2020)	Não	Não	Sim	Não	Não	Não
CHEN *et a*l. (2019)	Não	Sim	Sim	Sim	Sim	Não
HUANG (2010)	Parcial	Não	Sim	Não	Sim	Não
MÜLLER *et al.* (2016)	Parcial	Sim	Sim	Sim	Sim	Não
PARK (2013)	Não	Sim	Sim	Parcial	Não	Não
RAMACHANDRAN; RAMASUBRAMANIAN (2018)	Parcial	Sim	Sim	Sim	Sim	Parcial
TOMMASEL; GODOY (2018)	Parcial	Sim	Parcial	Sim	Sim	Não
YAN (2013)	Não	Não	Não	Parcial	Não	Não

Fonte: elaborado pelo autor (2021)

O trabalho de Müller *et al*. (2016) mostra que existe uma grande área a ser pesquisada e um grande potencial para aplicação de técnicas de processamento de linguagem natural em ambientes como os repositórios digitais, em especial o uso de abordagens de aprendizado de máquina em dados estruturados e não estruturados. Adiciona ainda que esta recuperação de informação pode ser ainda mais enriquecida ao ser relacionada com julgamentos humanos.

> [...] a analysis phase, we showed the potential of NLP and machine learning techniques for discovering patterns and relationships in large and unstructured data sets, and how the results of such techniques can be triangulated with human judgements (MULLER *et al*., 2016, p. 10).

Muller *et al*. sustentam uma importante discussão sobre a análise de dados (texto) coletados de repositórios a qual explicita que pesquisas anteriores estavam muito mais focadas em uma análise quantitativa do que qualitativa, nas quais aspectos como tamanho de expressões e análises sintáticas eram mais centrais nas pesquisas. No entanto, pondera que estas análises negligenciam o ponto focal de um documento que é o próprio conteúdo textual, e propõe o uso de modelagem probabilística de tópicos, como o algoritmo **Latent Dirichlet Allocation (LDA)** como auxílio preponderante para reconhecimento do conteúdo de um documento.

> [...] Data analysis While early research on review helpfulness mostly focused on quantitative review characteristics, especially the numerical 'star' rating, more recent studies have started to analyse their textual parts (e.g.,Mudambi & Schuff, 2010; Cao et al, 2011; Ghose & Ipeirotis, 2011; Pan & Zhang, 2011; Korfiatisa et al, 2012). Yet, these studies largely focus on easily quantifiable aspects, including syntactic or stylistic features such as review length, sentiment or readability, and mostly neglect the actual content of the review text (i.e., what topics is a reviewer writing about?). In order to capture a review's content and its impact on review helpfulness, we applied probabilistic topic modelling using the Latent Dirichlet Allocation (LDA) algorithm (MULLER *et al*., 2016, p. 7).

O trabalho de Tommasel e Godoy (2018) introduz uma discussão sobre o uso da expansão contextual para textos curtos (*short text*) que são muito utilizados em redes sociais e pontua que este tipo de texto leva a uma dificuldade de se analisar, compreender relações e inferir expressões pelas técnicas de processamento de linguagem natural.

> [...] However, short-texts coming from the social Web pose new challenges to this well-studied problem as texts' shortness offers a limited context to extract enough statistical evidence about words relations (e.g. correlation), and instances usually arrive in continuous streams (e.g. Twitter timeline), so that the number of features and instances is unknown, among other problems (TOMMASEL; GODOY, 2018, p.1).

Esta reflexão contribui com esta tese de doutorado, pois a análise do texto inserido em uma *query* para busca em um repositório possui as mesmas características de um *short text*, ou seja, limitam análises de inferências e correlações, o que corrobora a motivação desta pesquisa em proporcionar maior capacidade de análise a textos de busca ao utilizar-se como proposta a **expansão contextual** através de incorporação de palavras.

Na base de dados BRAPCI, foram encontrados um total de 1589 trabalhos (quadro 3), no entanto, 1122 são sobre Recuperação da Informação, ou seja, como um termo mais amplo dentro da Ciência da Informação corrobora-se a existência de uma quantidade razoável de artigos.

Na continuação da pesquisa nesta base de dados, foram encontrados 241 artigos sobre Repositórios Digitais, mostrando-se assim ser um termo de pesquisa em crescimento e ainda 177 artigos quando foram somados usando os termos de pesquisa descritos no quadro 3.

Este cenário apresentado no parágrafo acima mostra que os termos de pesquisa de destaque na Ciência da Informação estão em crescimento de pesquisa pela comunidade, porém, quando foram pesquisados termos como *Topic Modeling*, LDA, *Latent Dirichlet Allocati*on, incorporação de palavras, *word embeddings*, a quantidade de documentos retornados diminui muito chegando a se ter nenhum documento encontrado para junção de termos, como recuperação da informação e *Topic Modeling* e para Repositórios digitais, LDA, *Latent Dirichlet Allocation*, recuperação da informação, modelagem de tópicos, processamento de linguagem natural.

Mesmo assim foram encontrados artigos na área, como o trabalho de Silva e Souza (2014) que analisa os fundamentos do processamento de linguagem natural e coloca como proposta a utilização da análise para extração de bigramas. Bi-gramas ou de forma genérica, n-gramas, são possibilidades de se associar mais de uma palavra no processamento de linguagem natural para se chegar à semântica da sentença.

Segundo Sarmento (2005), o texto não é um simples amontoado aleatório de palavras. A ordem da colocação das palavras no texto é que produz o significado. Portanto, o estudo da co-ocorrências das palavras traz consigo uma informação importante. Isso pode indicar que as palavras estão relacionadas, diretamente por composicionalidade ou afinidade, ou indiretamente por semelhança.

Outro conceito importante apontado por Silva e Souza (2014) e corroborado por Zhang *et al*. (2009) é o de que palavras que possuem uma informação contextual, ou seja, possuem posicionamento próximo uma da outra, têm maiores chances de serem similares semanticamente. Os autores denotam haver uma crescente demanda pela utilização de técnicas de processamento de linguagem natural e que esta possui várias

aplicações, especialmente no campo da Recuperação da Informação, tradução automática, sumarização de texto, entre outras.

O trabalho de Fachin (2010) traz um levantamento sobre Recuperação da Informação Inteligente na área da Ciência da Informação. Neste artigo, a autora afirma que o uso de ontologias como representação de uma conceituação sobre um determinado domínio de conhecimento é importante para a recuperação da informação.

A autora destaca ainda que a recuperação da informação é essencial para a evolução das ciências e um desafio para Sistemas de Recuperação da Informação (SRI) que devem recuperar somente documentos com pertinência para o usuário.

> Muitos autores que serão citados ao longo do trabalho apresentam e discutem a recuperação da informação como essencial para a evolução das ciências, reafirmando sua importância na pesquisa e na comunicação científica em todas as áreas do conhecimento. Destaca-se que um dos principais problemas ou desafios dos Sistemas de Recuperação da Informação (SRI) é recuperar somente documentos importantes para o usuário, ou seja, que tenham "relevância" e, infelizmente, esses SRIs possuem relevância parcial (FACHIN, 2010, p. 260).

Fachin (2010) salienta que em sistemas digitais os usuários recuperam a informação através de *browsing* (navegação através de links de documentos ou sites) ou por *searching* (consulta ao um banco de dados) e que em uma ação de *browsing* o usuário nem sempre sabe o que precisa recuperar. Desta forma, analisa que sistemas de recuperação da informação ganham maior inteligência ao utilizar-se de agentes computacionais e ontologias, o que chamou de Recuperação Inteligente da Informação.

Em relação aos artigos da base de dados Scopus são vários os artigos recuperados, no entanto, foi feito um recorte para aumentar o ponto focal dos conceitos e ao mesmo tempo restringir a busca a artigos dentro da Ciência da Informação, e desta maneira seguem os artigos que foram analisados pela sua sinergia com esta tese.

Segundo Esposito *et al*. (2020), uma das formas de se obter uma recuperação da informação é através da pergunta e resposta ou ainda *question answering*, que permite aos usuários obter informação através de linguagem natural recebendo respostas precisas em pedaços de informação e baseadas nas questões ou dicas expressadas pelos usuários.

> [...] Question Answering (QA) is a sophisticated form of Information Retrieval (IR), allowing users to express their information needs in natural language and receive

> answers contain precise pieces of information. QA systems are characterized by a common high-level architecture, according to which, first, a natural language question is processed and a set of clues is extracted from it to understand what it is being asked for [15,42] . Then, these clues are used to query a knowledge base or a collection of documents and retrieve relevant information correlated to the question [3,26]. Finally, the retrieved information is evaluated in order to select, among more possible alternatives, the most confident answer expressed in a succinct form (ESPOSITO *et al.*, 2020, p. 1).

Uma interessante abordagem que Esposito *et al.* (2020) fazem é a chamada *hybrid query expansion* ou expansão híbrida de consulta na qual são utilizados recursos léxicos das técnicas do processamento de linguagem natural em conjunto com modelos do algoritmo Word2Vec.

De forma mais detalhada, o artigo de Esposito *et al.* (2020) mostra que são extraídos em uma primeira fase sinônimos e hiperônimos (são palavras com sentido mais abrangente que pertencem ao mesmo grupo semântico, por exemplo a palavra “animal” é um hiperônimo para “cachorro” de uma base lexical italiana chamada *Multiwordnet*. Em uma segunda fase é utilizado um modelo do **Word2Vec** para ranquear palavras com maior similaridade entre as previamente estabelecidas para propiciar uma busca mais aprimorada.

> [...] In order to augment the effectiveness in retrieving relevant sentences from documents, this paper proposes a hybrid Query Expansion (QE) approach, based on lexical resources and word embeddings, for QA systems. In detail, synonyms and hyper- nyms of relevant terms occurring in the question are first extracted from MultiWordNet and, then, contextualized to the document collection used in the QA system. Finally, the resulting set is ranked and filtered on the basis of wording and sense of the question, by employing a semantic similarity metric built on the top of a Word2Vec model (ESPOSITO *et al.*, 2020, p. 1).

Este artigo traz processos semelhantes a esta tese ao utilizar o algoritmo Word2Vec para análise de similaridade semântica entre termos de busca, e está preocupado com uma melhoria na recuperação da informação trazendo um conceito de expansão contextual híbrida ao adicionar 2 métodos para esta melhoria.

Ainda entre os artigos com maior expressividade da base de dados Scopus, por ter alto grau de relação dos conceitos e processos desta tese, destaca-se o trabalho de Ruas *et al*. (2020) que propõe novos algoritmos que combinam relações semânticas proporcionadas pelo encadeamento lexical com conhecimento prévio de base dados, usando para isso conceitos como **hipótese distributiva** ou *distributional hypothesis*

(apresentada na seção 2) e incorporação de palavras (*word embeddings*) em um único sistema.

> These algorithms combine the semantic relations derived from lexical chains, prior knowledge from lexical databases, and the robustness of the distributional hypothesis in word embed- dings as building blocks forming a single system. In short, our approach has three main contributions: (i) a set of techniques that fully integrate word embeddings and lexical chains; (ii) a more robust semantic representation that considers the latent relation between words in a document; and (iii) lightweight word embeddings models that can be extended to any natural language task. We intend to assess the knowledge of pre- trained models to evaluate their robustness in the document classification task (Ruas *et al*., 2020, p.1).

Suas principais contribuições são propor um conjunto de técnicas como *word embeddings* e *lexical chains* para trabalharem de forma integrada, uma segunda contribuição é mostrar que pode-se obter uma representação semântica mais robusta ao utilizar-se dos conceitos acima descritos (*word embeddings* e *lexical chains*) e uma terceira contribuição de explorar novos modos de análise textual semântica utilizando-se de modelos e treinamentos.

As contribuições de Ruas *et al.* (2020), em grande parte, vão ao encontro da proposta desta pesquisa, pois propor técnicas de processamento de linguagem natural de forma híbrida e propor algoritmos de aprendizado de máquina não supervisionados, como Word2Vec para incorporação de palavras e assim permitir semântica vetorial, é o foco deste trabalho.

2.2 CONTRIBUIÇÃO PERANTE O ESTADO DA ARTE

Esta tese possui contribuições para a área de recuperação da informação dentro da ciência da informação, pois propõe um modelo que relaciona tecnologias para uma melhoria na recuperação da informação especialmente em repositórios digitais. Este modelo proporciona uma abordagem para aumento da semântica na recuperação de documentos, e por conseguinte de informação, usando de conceitos, técnicas e tecnologias tratados pelos trabalhos referenciados.

Neste sentido, esta tese se utiliza dos trabalhos de Tommasel e Godoy (2018) que usam o conceito de expansão contextual e os trabalhos de Souza (2014) e de Zhang *et al*. (2009), pois os mesmos utilizam dos fundamentos do processamento de linguagem

natural para extração de conjuntos de palavras próximas pelo conceito de n-gramas, ou seja, um dos principais conceitos aplicados nesta tese que também se utiliza do processamento de linguagem natural para aprimorar os textos dos documentos relacionados em uma busca, melhoria esta que perpassa por retirar palavras que não são essenciais a busca do usuário por serem repetidas ou até mesmo conjunções que não expressam valor.

O trabalho de Ruas et al. (2020) está baseado nos conceitos de hipótese distributiva e incorporação de palavras que também são utilizados no trabalho de Müller et al. (2016). Tais trabalhos fazem uso do conceito de topic modeling baseado no algoritmo LDA, ou seja, uma técnica apoiada por algoritmos que fazem uma análise do texto e permitem descobrir tópicos latentes em um corpus.

Este trabalho é especificamente pertinente e tem relação enfática com a esta tese, pois esta também faz uso do algoritmo e técnica em questão ao posicionar estes no modelo proposto, uma vez que o modelo da tese faz uma comparação das palavras usadas pelo usuário em sua busca (termos de busca) com as palavras arroladas pelo algoritmo/técnica acima expostos.

Este processo permite ao modelo proposto nesta tese uma redução de complexidade, ou seja, uma diminuição da quantidade de palavras a serem analisadas em uma comparação comum, ou até mesmo em uma comparação com somente as palavras-chave definidas no cadastro de cada documento em um repositório.

O artigo de Esposito et al. (2020), utiliza o algoritmo word2vec para encontrar palavras com maior similaridade semântica baseadas no corpus, este tipo de aplicação corrobora a utilização deste tipo de algoritmo para se chegar a uma outra técnica conhecida como expansão contextual. Esta tese também faz uso do mesmo processo na aplicação do modelo proposto no momento ao qual é realizada uma consulta em uma base de vetores pré-definidos chamada NILC.BR.

No entanto, nenhum dos trabalhos utilizam estes conceitos diretamente aplicados em um repositório digital e não possuem um modelo, como a proposta desta tese, que

utilize os conceitos de forma integrada para uma recuperação da informação com abordagem semântica em repositórios digitais.

3 RECUPERAÇÃO DA INFORMAÇÃO E EXPANSÃO CONTEXTUAL

O termo Recuperação da Informação do inglês, *information retrieval*, remonta à década de 1950 com Calvin Mooers, que apresentou a disciplina recuperação da informação, dentro da ciência da informação, e que esta nova disciplina responde e trata de aspectos não somente intelectuais, mas também por sistemas, técnicas e máquinas.

Segundo Mooers (1951), a recuperação da Informação trata dos aspectos intelectuais da descrição da informação e sua especificação para busca, e de qualquer sistema, técnicas ou máquinas que são empregadas para realizar esta operação.

Segundo Saracevic (1995), a interdisciplinaridade entre ciência da informação e ciência da computação é essencial para a área de Recuperação da Informação, e a Ciência da Computação seria o braço tecnológico da ciência da informação. Saracevic enaltece que este relacionamento em torno da Recuperação da Informação vai além e pontua que a interdisciplinaridade concerne a quatro áreas: Biblioteconomia, Ciência da Computação, Comunicação e Inteligência Artificial ou como descrito por Saracevic, ciência cognitiva.

Um dos problemas definidos por Saracevic (1995, p. 2) em relação à Recuperação da Informação diz respeito a uma grande e contínua geração de dados. O autor chamou esse fenômeno de “explosão de informação”. Este fenômeno, de certa forma continua e cresce nos dias de hoje, no entanto, a utilização da tecnologia para lidar com a crescente geração de dados é apontada como capaz de resolver ou abrandar este problema.

> [...] succinctly defined a critical problem that was on the minds of many for a long time, and (2) proposed a solution that was a “technological fix,” in tune with time and strategically attractive. The problem was (and in its basic form still is) “the massive task of making more accessible a bewildering store of knowledge.” This is the problem of “information explosion,” coupled with necessity to provide availability of and accessibility to relevant information, acute to this day. Witness the reasons for evolution of digital libraries. His solution was to use the emerging information technology to combat the problem (SARACEVIC, 1995, p. 2).

Segundo Ferneda (2003), a recuperação da informação está baseada em duas fases. A primeira fase está relacionada à preparação do *corpus* a ser utilizado como base para a recuperação da informação, no qual ainda se contempla a representação da expressão de busca e a função de busca. E a segunda fase, na qual estão envolvidas: a expressão de busca, os resultados de busca e o próprio usuário, que a partir de sua

busca, ou seja, sua necessidade de informação deve interagir com algum sistema que propicie esta interação humano computador e que resulte em documentos que atendam sua necessidade.

Ainda segundo Ferneda (2003), a Recuperação de Informação envolve, além de pessoas, um acervo documental. E em buscas realizadas, mesmo com o apoio de sistemas computacionais, os resultados nem sempre são tão satisfatórios, dado o nível de complexidade em formalizar tais pesquisas.

É neste momento que este trabalho difere e de certa maneira propicia uma nova abordagem na recuperação da informação, pois, normalmente o que se possui é de fato o cenário proposto por Ferneda (2003). No entanto, atualmente, tem-se repositórios de documentos no qual o *corpus* possui muitos *terabytes* de dados, o que pode limitar o resultado de uma busca caso o usuário não tenha conhecimento sobre o *corpus* em si.

De acordo com Santarém Segundo (2010), na Ciência da Informação, a recuperação da informação (RI) tem sido cada vez mais estudada e atualmente sistemas computacionais têm um bom nível de busca sintática. No entanto, quando se trata de buscas semânticas, estes repositórios de dados não estão aptos a retornar informação semântica e contextualizada.

Conforme Saracevic (1996), a dificuldade de encontrar uma informação relevante e de interesse do usuário é um dos grandes problemas observados e discutidos no âmbito da Ciência da Informação, e um sistema de RI tem como principal objetivo “minimizar as dificuldades do usuário em localizar a informação requisitada, ou seja, diminuir o tempo gasto em um processo de busca até que a informação desejada possa ser acessada” (CRISTOVÃO; DUQUE; SERQUEIRA, 2012, p. 5-6).

Soma-se a este cenário que a qualidade e semântica dos documentos retornados precisam, cada vez mais, estarem de acordo com a busca feita pelo usuário. Haja vista que tais documentos, uma vez armazenados em repositórios digitais, necessitam de ferramentas capazes de refletir de fato a busca do usuário e, neste contexto, capacitar este processo nestes ambientes é importante para que não haja:

a) recuperação de documentos **não relevantes** para o usuário baseado em sua consulta.
b) recuperação de uma quantidade de documentos **aquém** da verdadeira capacidade

armazenada dentro do repositório digital.

Portanto, é necessário que melhorias constantes nesta recuperação da informação sejam cada vez mais estudadas e pesquisadas para um retorno e utilização mais fidedigno do repositório, como Tim Berners-Lee, James Hendler e Ora Lassila, em seu artigo publicado em 2001 afirmam que informações marcadas com regras de inferência podem compartilhar um sistema de representação e delegar a máquinas uma forma automática de explorar um conteúdo (BERNERS-LEE; LASSILA; HENDLER, 2001).

Segundo Martins *et al*. (2017), um repositório digital é definido como um meio para armazenar, gerenciar e preservar conteúdos informacionais no formato eletrônico e já possuem formas de se encontrar e recuperar documentos baseados em pesquisa por palavras-chave fornecidas pelo usuário.

Para Santos e Neves (2018), a forma de indexação de um documento em um repositório digital é baseada na mesma forma como o autor se utiliza para apresentar suas ideias no texto, ou seja, linguagem natural.

Neste contexto, como parte da técnica e da tecnologia pontuada por Saracevic (1995), utiliza-se neste trabalho os conceitos do Processamento de Linguagem Natural, a fim de melhorar a contextualização da busca do usuário em repositórios digitais e com isso permitir uma melhoria da Recuperação da Informação nestes ambientes digitais.

Segundo Selton (1968) *apud* Beppler (2008) a Recuperação de Informação é explorada desde a área de pesquisa que se preocupa com a estrutura, análise, organização, armazenamento, recuperação e busca de informação.

> O termo ‘recuperação de informação’ significa, para uns, a operação pela qual se seleciona documentos, a partir do acervo, em função da demanda do usuário. Para outros, ‘recuperação de informação’ consiste no fornecimento, a partir de uma demanda definida pelo usuário, dos elementos de informação documentária correspondentes (SELTON, 1968 *apud* BEPPLER, 2008, p. 15).

Ferneda (2003, p. 14) contextualiza afirmando que: “um modelo de recuperação de informação envolve a especificação formal de três elementos principais: a representação dos documentos, a representação das buscas dos usuários e a maneira como esses dois primeiros elementos serão comparados”.

Ainda segundo Ferneda (2003), a eficiência de um sistema de recuperação de informação está diretamente ligada ao modelo que o mesmo utiliza, sendo eles subdivididos em modelos quantitativos e modelos dinâmicos.

3.1 MODELOS DE RECUPERAÇÃO DA INFORMAÇÃO

Nos Sistemas de Recuperação da Informação tem-se de fato a prática ou a ação do fenômeno da necessidade de informação, seja de forma manual em tempos passados, seja utilizando-se de sistemas informatizados que possuem controle e armazenamento de cada documento, porém em formato digital.

A forma de representação e armazenamento destes conteúdos inseridos nos arquivos digitais são de extrema importância para que possam ser recuperados. Para Santarém Segundo (2010), quanto maior a clareza da representação do conteúdo, mais fácil seria recuperar a informação.

Portanto, a forma como se pode recuperar conteúdo de um Sistema de Recuperação da Informação, ou mais especificamente de um repositório digital, passa pela utilização de diferentes tipos de modelos para esta recuperação da informação.

Nesse contexto, na seção seguinte são abordados os modelos, tanto dinâmicos quanto quantitativos, suas características e como podem ser utilizados em diferentes cenários de aplicação.

3.1.1 Modelo booleano

O modelo booleano utiliza-se dos operadores lógicos (“AND”, “OR” e “NOT”) da álgebra booleana para recuperação de informação ao basear-se na justaposição de termos de busca do usuário com a indexação do documento através de expressões lógicas.

É possível se efetuar uma combinação entre os operadores, aumentando a acurácia do que se propõe a ser buscado, porém, as operações criadas neste método seguem uma ordem de execução que, de acordo com Ferneda (2003), influenciam no resultado.

Algumas limitações são conhecidas do modelo booleano, como:

a) por se basear na comparação binária, expressões diferentes podem retornar um mesmo conjunto de documentos, portanto, o resultado não possui grande valor já que os termos não estão definindo um retorno correto.
b) não há possibilidade de que resultados sejam apresentados de forma parcial, ou seja, somente seguindo a premissa binária de encontrado ou não encontrado.

Neste cenário, o modelo booleano nem sempre expressa confiança no retorno dos documentos, promovendo dessa maneira à utilização de outros modelos para complemento do modelo booleano.

3.1.2 Modelo vetorial

Devido às limitações do modelo booleano, o modelo vetorial ganha notoriedade ao promover a intenção de comparar os documentos por pesos, que são associados pelos documentos e pela própria expressão de recuperação, buscando o grau de similaridade entre ambos.

De acordo com Santarém Segundo (2010, p.33):

> No modelo vetorial, a consulta é realizada em busca dos termos designados, e a classificação apresentada como resultado baseia-se na frequência dos termos no documento em relação ao peso atribuído a cada termo utilizando-se o grau de similaridade calculado.

Ferneda (2003) defende que essa comparação parcial ocorre perante a associação de pesos tanto aos termos de indexação como aos termos da expressão de busca. Esses pesos são utilizados para calcular o grau de similaridade entre a expressão de busca formulada pelo usuário e cada um dos documentos disponíveis. Como resultado, obtém-se um conjunto de documentos ordenados pelo grau de similaridade de cada documento em relação à expressão de busca.

A ideia do modelo vetorial tem a capacidade de garantir respostas mais precisas, em comparação com o modelo booleano, pois trabalha com a possibilidade de uma flexibilização do termo de busca, trazendo documentos que tenham similaridade com parte do termo de busca. Vale salientar que essa parcialidade pode trazer documentos

que não tenham ligação real com o que é procurado pelo usuário, graças a ambiguidade que os termos de busca podem alcançar.

Como exemplo prático, uma busca sobre o termo (Recuperação da Informação) pode trazer resultados referentes à área da RI, assim como (recuperação) e (informação), trazendo documentos relevantes a recuperação de diversas áreas, como recuperação judicial, e qualquer documento que tenha (informação) no seu padrão de indexação.

Baeza-Yates e Ribeiro-Neto (1999, p. 47) citam algumas vantagens a respeito deste modelo, sendo elas:

> a) Melhoria na qualidade da recuperação, devido ao seu esquema de ponderação de termos;
> b) A busca por termos parciais permite a recuperação de documentos mais próximos em relação a consulta;
> c) Seu esquema de *ranking* é ordenado de acordo com o seu grau de similaridade em relação a consulta;
> d) A normalização pelo tamanho do documento está naturalmente ligada ao modelo.

Baeza-Yates e Ribeiro-Neto (1999) afirmam ainda que, em comparação com outros métodos de ranqueamento, o modelo vetorial acaba sendo mais sólido, simples e rápido quando se trata de coleções genéricas.

3.1.3 Modelo probabilístico

Trazido do estudo da probabilidade, onde um mesmo evento pode apresentar resultados diferentes, o modelo probabilístico vem com a ideia de classificar documentos de acordo com a probabilidade em relação aos termos de busca.

Seu sistema de ranqueamento ordena os documentos com maior probabilidade de serem relevantes, trazendo em teoria, melhores resultados. Porém, Baeza-Yates e Ribeiro-Neto (1999) acrescentam que esse ranqueamento não funciona na totalidade, pois a relevância dos documentos pode ser afetada por fatores externos ao sistema.

Baeza-Yates e Ribeiro-Neto (1999, p. 55) elencam algumas desvantagens:

> a) A necessidade em segregar os documentos entre relevantes e não relevantes;
> b) Não leva em consideração o número de ocorrências do termo de indexação;
> c) Os documentos não seguem uma normalização de tamanho.

O modelo probabilístico é o único modelo que utiliza o processo de *Relevance Feedback*, pois deixa a tarefa de reconhecer a relevância do documento com o usuário.

Ferneda (2003, p. 42) ressalta que “o modelo probabilístico pode ser implementado utilizando a estrutura proposta pelo modelo vetorial, permitindo integrar as vantagens desses dois modelos em um sistema de recuperação de informação”. E completa que experimentos realizados com uma quantidade pequena de documentos mostram que o modelo probabilístico não chega a ser muito superior ao modelo booleano, tendo seu certo destaque no contexto da Web. Sua complexidade também é responsável por desestimular desenvolvedores a abandonarem os outros dois modelos.

3.2 MODELOS ESTENDIDOS OU COMPLEMENTARES

Segundo Souza (2006), os modelos de Recuperação da Informação se dividem em clássicos, como os definidos acima (modelo booleano, vetorial e probabilístico) e modelos complementares ou estendidos. Esses últimos são modelos que são complementares aos clássicos, como o *Fuzzy* e booleano estendido para o modelo clássico booleno, o vetorial generalizado e o de Redes Neurais para o Vetorial e para o modelo clássico probabilístico possui-se os modelos Redes de Inferência e Redes de Crença.

Para os modelos estendidos referentes ao probabilístico, o **Redes de Inferência** vinculam variáveis aleatórias ao evento de atendimento de uma *query* e estas variáveis vão sendo alteradas em relação a buscas futuras para melhoria do retorno.

Já o modelo **Redes de Crença** (*belief networks*), mais um modelo estendido do modelo probabilístico, atua como o modelo de redes de inferência, porém *queries* e documentos são adaptados como um subconjunto de um espaço de conceitos e a cada documento e cada *query* se associa a uma probabilidade destes em relação ao espaço de conceitos.

O **modelo Fuzzy** ou também conhecido como de lógica difusa ou ainda nebulosa atua como uma extensão do modelo booleano, seu funcionamento está baseado no grau de pertencimento dos termos da *query* (busca do usuário). Sua extensão está pontualmente em poder se utilizar de algum instrumento de apoio semântico, como um tesauro por exemplo.

Já pela utilização do modelo **booleano estendido** tenta-se exceder o problema de decisão binária do modelo booleano clássico, ao pontuar-se pesos para os termos da busca aproximando-o do modelo vetorial.

O modelo **vetorial generalizado** questiona a independência das palavras dos termos de busca e procura um relacionamento entre as palavras através da análise de co-ocorrências das palavras em cada documento.

Como definido por Souza (2006), dentre os modelos estendidos ou complementares do modelo clássico vetorial, está o modelo **Redes Neurais**, que se utiliza da capacidade das redes neurais para analisar padrões utilizando os termos da busca com os documentos.

> [...] Redes neurais: nesses modelos, utiliza-se o poder das redes neurais para realizar o casamento de padrões entre as queries e os documentos do acervo do sistema. Cada query dispara um sinal que ativa os termos índice, que por sua vez propagam os sinais aos documentos relacionados. Estes, por sua vez, retornam os sinais a novos termos índices, em interações sucessivas. O conjunto resposta é definido através desse processo, e pode conter documentos que não compartilhem nenhum termo-índice com a query, mas que tenham sido ativados durante o processo (SOUZA, 2006, p.1).

Vale uma análise deste último modelo estendido do modelo vetorial, o modelo de redes neurais, pois este modelo atua utilizando-se do conceito de redes neurais em profundidade, ou seja, sinais são disparados entre os documentos e termos da busca e aqueles que foram atingidos pelos sinais são retornados, muitas vezes sem possuir algum termo da busca em si, mas aqueles que foram acionados pela rede neural.

Porém, há de se destacar que este modelo não tem relação com os modelos neurais, como o Word2Vec (utilizado no modelo proposto) que se utilizam dos vetores de palavras (ver 4.6) que foram obtidos segundo aprendizado de máquina sobre o *corpus* de um repositório digital.

3.3 EXPANSÃO DE CONSULTA (QUERY EXPANSION)

Segundo Esposito *et al*. (2020), nos últimos anos a Expansão Contextual (QE) ou ainda expansão de termos de busca tem tido efetividade na Recuperação da Informação ao utilizar-se de diferentes fontes informacionais em conjunto com a automação dos métodos relacionados a este processo.

Os métodos de automação, em sua maioria, estão voltados ao **enriquecimento dos termos por utilização de Processamentos de Linguagem Natural** a fim de permitir que termos semanticamente relacionados sejam enviados, posterior ao enriquecimento, como novos termos de busca para o motor do repositório digital de documentos com o objetivo da melhoria da Recuperação da Informação.

A quantidade de termos a somar-se aos termos originais para envio ao *engine* de busca do repositório é uma análise que deve ser realizada periodicamente e que também depende de cada cenário. Pois, se uma quantidade muito grande de termos for submetida para a busca, pode haver uma superestimação do retorno. Porém, caso sugira uma quantidade de termos pequena, pode haver, por outro lado, uma diminuição da assertividade do retorno.

Este problema continua em aberto, conforme Esposito *et al*. (2020). Porém, o que se propõe como solução no estado da arte neste contexto é a utilização de relacionamentos semânticos com os termos de busca, o que vai ao encontro do proposto nesta tese, assumindo-se assim que o caminho traçado conjumina-se com a pesquisa de vanguarda neste cenário.

Segundo Maron e Kuhns (1960), a expansão de consulta é o processo de reformular uma determinada entrada do usuário em uma consulta (*query*) a fim de melhorar a Recuperação da Informação baseada nas palavras de busca. Dentre as técnicas para esse processo, pode-se enumerar:

a) Encontrar sinônimos e adicioná-los em uma pesquisa;
b) Variações morfológicas das palavras (*stemming*);
c) Correção automática de eventuais erros de digitação.

No entanto, a técnica de se utilizar vetor de palavras (ver 4.6) para incorporação semântica é a técnica do estado da arte, segundo Esposito *et al*. (2020), e as técnicas enumeradas acima não levam em conta a questão semântica.

Saar, Shtok e Kurland (2016) pontuam a utilização de vetores de incorporação de palavras para resolução de problemas semânticos entre os termos de consulta e os documentos de um *corpus*, o que corrobora a menção de Esposito *et al*. (2020) em relação à utilização de tais métodos para melhoria da Recuperação da Informação.

Um exemplo é o algoritmo Word2Vec, utilizado tanto em Saar, Shtok e Kurland (2016) quanto em Esposito *et al*. (2020). Este algoritmo permite que cada palavra em um texto seja aprendida, ou seja, armazenada em uma estrutura chamada vetor de palavra. O processo de aprendizado das palavras, e do contexto em que elas se encontram, passa pela fixação de um tamanho de janela que é deslizado para cada palavra no *corpus*.

Esta janela pode ser explicada por um exemplo: na sentença "*o gato pulou no mato*", a janela seria a quantidade de palavras que se quer obter em relação ao contexto da palavra central, ou seja, se a janela tem tamanho 1, o algoritmo irá obter a palavra central (*pulou*) e uma palavra posterior à central (*mato*) e uma palavra anterior à central (*gato*).

Este processo, iterado milhares de vezes pelo algoritmo, permite que seja armazenada uma estrutura na qual cada palavra esteja em um contexto de suas palavras sucessoras e predecessoras, e esta estrutura vai permitir uma análise de quais palavras possuem similaridade contextual. Neste sentido, a utilização do algoritmo Word2Vec possibilita análises semânticas baseado em vetores de palavras e é utilizado como parte da técnica do modelo e será explicado em detalhes em seções posteriores.

Em resumo, a expansão contextual surge como conceito base para que os sistemas computacionais possam retornar informação contextual para o usuário e desta maneira possibilitar recuperação da informação com abordagem semântica. Estes conceitos em conjunção aos conceitos e técnicas das seções seguintes que abordam inteligência artificial serão a base para o modelo apresentado no capítulo 5.

4 INTELIGÊNCIA ARTIFICIAL E PROCESSAMENTO DE LINGUAGEM NATURAL

Nesta seção, é apresentada uma discussão sobre a inteligência artificial, como vem sendo utilizada no processo de recuperação da informação, como modelos como o LDA e técnicas com a de *Topic Modeling*, aos quais estão baseados conceitos e técnicas de Aprendizado de Máquina, interagem com o processo de RI. Como o Processamento de Linguagem Natural (PLN) e técnicas de incorporação de palavras (*word embeddings*), hipótese distributiva, além da TF-IDF e semântica vetorial estão vinculados para propiciar recuperação da informação contextual ao usuário.

4.1 CONCEITOS DA INTELIGÊNCIA ARTIFICIAL

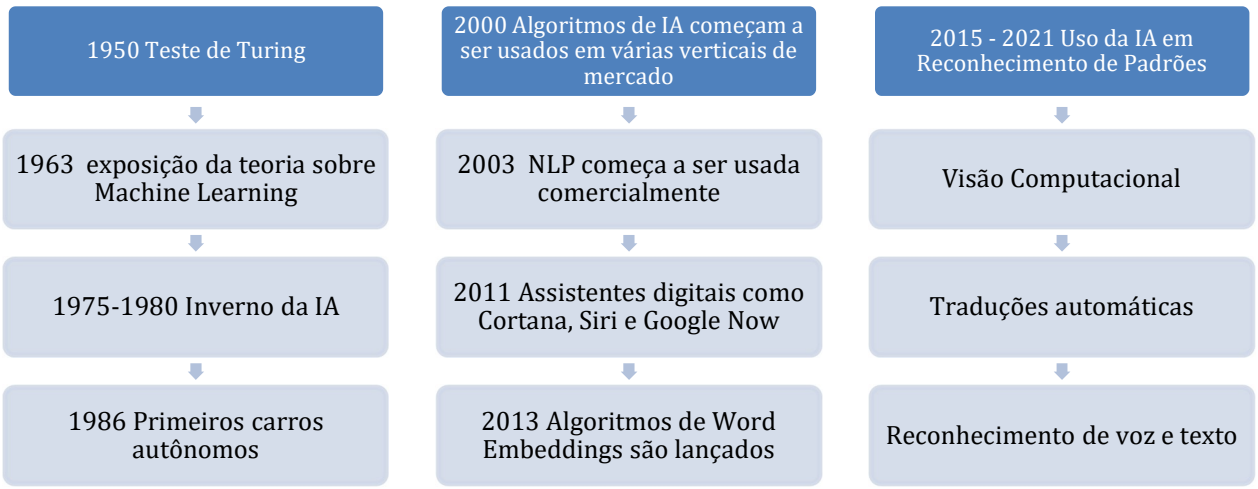
A inteligência artificial (IA), nos últimos anos, vem ganhando forte espaço nos mais variados ambientes digitais como análise de imagens, visão computacional para veículos não tripulados entre outros. No campo da linguística e análise textual, em específico o processamento de linguagem natural, a forma como propicia ao ser humano interagir com artefatos digitais de maneira mais assertiva na recuperação de informação vem crescendo e, também, ganhando grande visibilidade.

Por definição, segundo Russel e Norvig (2016) a Inteligência Artificial é a capacidade das máquinas imitarem ou simularem comportamentos cognitivos humanos, como a capacidade de resolver problemas e aprender. Ainda segundo os autores a IA em sua completa capacidade teria a possibilidade de tomar decisões de forma racional e chegaria perto da capacidade humana.

É comum se estabelecer uma diferença quando usado o termo inteligência artificial, esta diferença está na chamada IA Forte ou do inglês Strong IA e IA Fraca ou Weak IA. A forte é aquela que imitaria perfeitamente um ser humano, desde as faculdades de elaboração de pensamentos e tomadas de decisão até a capacidade de locomoção e conhecimento de si. Já a IA fraca, a que mais avançou nos últimos tempos, está ligada a resolução de problemas mais específicos, ainda que complexos, tais como elaboração de diagnósticos, composições musicais, discernimento analisando imagens, identificação de padrões, análise de documentos entre outras.

Muitas são as capacidades, atualmente, quando a inteligência artificial é aplicada, destas pode-se destacar áreas como: educação, saúde, automações, financeiro, segurança, mídias sociais, entretenimento, jogos, agricultura, e-commerce, robótica entre outros. Por ser uma área muito abrangente, a inteligência artificial abarca várias outras áreas e tem tido um crescimento e variedade de aplicações. Neste sentido, a figura 1 relata uma linha do tempo para melhor compreensão das técnicas, conceitos e tecnologias envolvidas pela aplicação da IA.

Figura 1 – Histórico da Inteligência Artificial



Fonte: elaborado pelo autor (2021)

Neste trabalho a inteligência artificial está focada em dois campos específicos, o processamento de linguagem natural e o uso de algoritmos de aprendizado de máquina, que serão tratados nas seções seguintes. Na próxima seção, são mostrados dois modelos base para o tratamento de dados para aplicação da IA, além de mostrar como estes modelos se relacionam ao modelo proposto no capítulo 5.

4.1.1 Modelos estatísticos

Os modelos estatísticos estão vinculados aos processos da inteligência artificial desde seu início, especificamente, no processamento de linguagem natural, estes

modelos começam a ser aplicados nas décadas de 1980 e 1990, onde há um crescimento do conceito e maior disponibilidade de poder computacional.

Um modelo estatístico faz uma estimativa probabilística sobre as unidades às quais trabalha. Especificamente, no processo do modelo desta tese, estas unidades de trabalhos são textuais, como palavras, ou ainda frases ou sentenças.

Segundo Everitt (2006), a **Distribuição probabilística** é a função matemática que fornece as probabilidades de ocorrência de diferentes resultados possíveis para um experimento. Este conceito forma a base dos modelos estatísticos e analisa que para uma determinada variável (por exemplo uma palavra) existe uma gama de possibilidades, ou seja, um mapeamento das probabilidades de possíveis saídas dada uma entrada.

Em especial para o foco deste trabalho, no processamento de linguagem natural, um modelo estatístico atua de forma quantitativa, ou seja, em um primeiro momento é realizada uma coleta de todas as palavras de um determinado *corpus*, após esta coleta é possível fazer levantamentos como quantidade de palavras, frequência de palavras, categorização de palavras, contagem de palavras que estão a uma distância de uma ou mais palavras e inferências sobre probabilidades de ocorrência de palavras.

Estas funcionalidades dos modelos estatísticos são frequentemente reutilizadas no processamento de linguagem natural, e por serem processos repetitivos foram criados *frameworks* (ver quadro 6), ou seja, arcabouços de funcionalidades que estão disponíveis à comunidade para a aplicação dos processos característicos do PLN.

Quadro 6 - Frameworks PLN

Framework	**URL**
PyTorch	https://pytorch.org/
Stanford CoreNLP	https://stanfordnlp.github.io/CoreNLP/
SciKitLearn (usado neste trabalho)	https://scikit learn.org/stable/
Spacy	https://spacy.io

Fonte: elaborado pelo autor (2021)

Especificamente, neste trabalho foi escolhido o *framework Scikit-Learn* para uso nos processos estatísticos do processamento de linguagem natural. Esta decisão se baseia em duas principais características do framework, estar escrito na linguagem python, linguagem padrão do trabalho em relação à prova de conceito e na habilidade do autor em relação à linguagem de programação necessária. Mesmo assim, no quadro 6,

são apresentados outros possíveis frameworks comuns à comunidade que podem desempenhar papel similar.

4.1.2 Modelos neurais

Segundo Sangiorgi (2019), as redes neurais artificiais, ou como definido pelo autor as NARs (Redes Neuro-Artificiais), podem relacionar informações variadas e de diferentes locais, o que possibilita uma generalização e permite capacidade de inferências sobre novas teorias a partir deste conglomerado de informações.

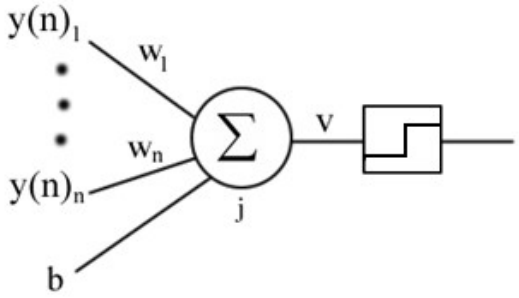
Já para Ferneda (2006), redes neurais são modelos que buscam simular o comportamento do cérebro humano em relação ao processamento de informação ao enviar sinais de um neurônio a outro a fim de processar tais sinais e, como resultado, levar a um estímulo físico.

Deste modo, as redes neurais podem ser definidas como modelos matemático-computacionais baseados no comportamento biológico da estrutura neural de seres humanos entre outros animais com alguma inteligência cognitiva. São formadas pelo combinado de vários neurônios artificiais os quais podem ser entendidos como funções matemáticas compostas por um conjunto de conexões de entrada caracterizadas por seus pesos, além de um somador para acumular os sinais da entrada e uma função de ativação que permite ou não que o sinal de entrada passe aos neurônios posteriores.

Uma das características mais importantes de uma rede neural artificial é sua capacidade de aprender por intermédio de exemplos e fazer inferências sobre o que foi aprendido, ao passo que este processo vai sendo melhorado com o passar do tempo de treinamento desta rede.

Na figura 2, pode-se perceber estes parâmetros em detalhes, os quais são: $\mathbf{y(n)_1}$ a $\mathbf{y(n)_n}$ são as entradas, já $\mathbf{W_1}$ a $\mathbf{W_n}$ são os pesos para estas entradas que se multiplicam pelo peso **W** e, no final, somam-se formando um conjunto de entrada $\boldsymbol{\xi = \sum w * y(n)}$. O resultado passa por uma função de ativação e transmite a saída v. Quando o valor ξ exceder o limite da função de ativação, o neurônio será ativado e retornará um valor.

Figura *2* – Visão geral de uma Redes Neural Artificial



Fonte: elaborado pelo autor (2021)

De acordo com Bengio *et al*. (2003), modelos neurais, ou seja, aqueles baseados em redes neurais são introduzidos no contexto da Recuperação da Informação para propiciar maior generalização (conceito no âmbito do aprendizado de máquina que permite que uma determinada solução seja aplicada a outros conjuntos de dados).

Trazendo este conceito para o foco do trabalho, um modelo usando aprendizado de máquina para processar centenas de milhares de palavras de um *corpus*, ou seja, processar sentenças de múltiplas palavras possibilita que cada palavra analisada possa ter uma conexão com outras palavras. Esta capacidade de analisar palavras, mesmo que desconhecidas a princípio e fazer conexões com outras palavras define este tipo de modelo como essencial para aprender e entender um corpus e permitir inferências.

Nesse sentido, os modelos neurais tendem a melhorar o processo de análise de documentos de texto e, por conseguinte, a recuperação da informação em sistemas digitais computacionais baseados em documentos textuais por conseguirem mapear uma enorme quantidade de ligações entre palavras. Por se tratar de um modelo que “aprende” (faz inferências) com as palavras de um corpus e como estas se relacionam, possibilita análises entre documentos, ou seja, em um corpus e até entre corpus.

Este modelo será utilizado nesta tese através da aplicação do algoritmo Word2Vec que será apresentado e discutido na seção 4.6 e na próxima seção será apresentada uma análise de como a inteligência artificial vem sendo usada na Ciência da Informação desde os anos 1990 até o panorama atual.

4.2 INTELIGÊNCIA ARTIFICIAL E RECUPERAÇÃO DA INFORMAÇÃO NA CIÊNCIA DA INFORMAÇÃO

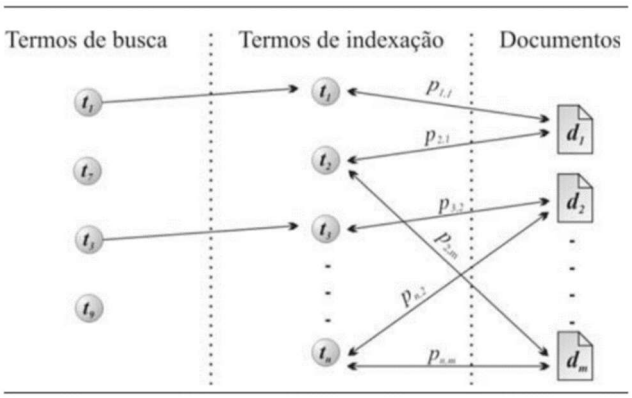
O processo de recuperação da informação vem ganhando aportes de outras ciências desde a década de 1980. Segundo Kobashi (1996), o crescimento do uso da Inteligência Artificial estava presente não somente nos sistemas especialistas, mas também em sistemas de indexação, catalogação e produção automática de resumos.

Já no início dos anos 2000, Ferneda (2006) pontua o uso de redes neurais na recuperação da informação e salienta que, mesmo sendo um campo da ciência da computação ligado à inteligência artificial, as redes neurais são modelos computacionais baseados no sistema nervoso que são capazes de selecionar, agrupar e classificar dados.

Neste sentido, Ferneda (2006) faz uma comparação de que um sistema de Recuperação da Informação (SRI) pode ser comparado com uma rede neural, pois este tipo de sistema é composto por expressões de busca em um primeiro momento, os termos de indexação no centro e os documentos como fim.

Esta estrutura, segundo Ferneda (2006) pode ser comparada a uma rede neural de três camadas, sendo a camada de entrada como as dos termos de busca, a camada de documentos está para a saída e a camada central seriam os termos de indexação, como pode ser visto na figura 3.

Figura 3 - Redes neurais como sistemas de recuperação da informação



Fonte: Ferneda (2006)

Nesta figura 3, é mostrada uma relação entre documentos e termos, esta relação demonstra, em um primeiro momento, que a utilização de conceitos da inteligência artificial para recuperação de informação pode ser um caminho para melhoria dos resultados, e em um segundo momento, mostra que a ciência da informação pode fazer uso dos conceitos e técnicas da ciência da computação e mais especificamente da inteligência artificial tanto conceitualmente como no exemplo do artigo referido, bem como usando as técnicas propriamente ditas como nesta tese, através de algoritmos e processamento em si.

Atualmente, outros processos, técnicas e tecnologias são aplicados para melhoria da recuperação da informação utilizando variações da inteligência artificial. Segundo Coneglian (2020), algumas são as formas de aplicações para recuperação da informação, as quais estão resumidas no quadro 7.

Quadro 7 – aplicações usando IA na recuperação da informação

Técnica de IA	**Aplicação**
Algoritmos Genéticos	Os algoritmos genéticos funcionam baseado nas teorias da biologia evolutiva. Na prática, os algoritmos genéticos irão evoluindo as populações, cruzando, mudando e eliminando os indivíduos, para que a nova geração seja mais adequada. Assim, as melhores opções são aprimoradas e os piores são eliminados, sendo assim, uma Inteligência Artificial que irá aperfeiçoar as opções de processamento. Na recuperação da informação, os algoritmos genéticos podem ser utilizados para definição de parâmetros e pesos para que o processo de recuperação seja melhor ajustado e adequado.
Deep Learning	A abordagem com Deep Learning utiliza pouco pré-processamento dos dados, visando ao aplicar, explorar a maior quantidade de camadas, e um nível profundo de análise dos dados. Na recuperação da informação, a utilização de deep learning para recuperação da informação como músicas e imagens tem-se mostrado bastante vantajosa. Isso ocorre, pois tanto as imagens quanto os áudios contêm grandes quantidades de informações, favorecendo com que o deep learning, por ser um aprendizado de máquinas que tem um grafo de análise mais profundo, seja aderente à recuperação da informação, que exige comparação com grandes quantidades de dados.

Processamento de Linguagem Natural	Uma das áreas da Inteligência Artificial que está afetando a área de recuperação da informação é o processamento de linguagem natural, que está modificando o modo como as pessoas interagem com os dispositivos computacionais, de modo que os indivíduos possam ter uma comunicação mais natural com os dispositivos.

Fonte: adaptado de Coneglian (2020)

As abordagens, técnicas e tecnologias reunidas no quadro 7 mostram novos caminhos para que a área de recuperação da informação, tal fato vai ao encontro deste trabalho, pois, demonstra que o uso de inteligência artificial para aplicação no processo de recuperação da informação é promissor.

4.3 APRENDIZADO DE MÁQUINA E ANÁLISE DE TÓPICOS

O campo do aprendizado de máquina ou ainda pelo termo corrente em inglês, *machine learning*, advém dos anos de 1950 logo após a criação do teste de Turing com os primeiros *softwares* que conseguiam aprender sozinhos a partir de dados de entrada.

Segundo Mitchell (1997), o aprendizado de máquina trata da questão da construção de programas que automaticamente melhorem sua atuação com base na sua própria experiência, ou seja, os programas de computadores conseguem melhorar sua assertividade quanto mais são utilizados dentro de um determinado domínio de aplicação.

Aprendizado de máquina é um subconjunto da inteligência artificial no qual programas são usados para aprender através deles mesmos, ou seja, de forma autônoma a partir de dados e informação.

Basicamente há dois tipos de algoritmos de aprendizado de máquina, o primeiro, chamado de **não supervisionado** é um tipo de algoritmo que não tem conhecimento prévio sobre o conjunto de dados ao qual será processado. Este tipo de algoritmo ao ser aplicado a um documento de texto tem a capacidade, sem conhecimento prévio dos dados, no caso das palavras do documento, de “aprender” sobre as palavras e suas relações dentro do documento.

Já o segundo tipo, chamado de algoritmo de **aprendizado supervisionado** utiliza-

se de um prévio conhecimento do *corpus*, ou seja, houve em algum momento um processamento para anotar quais palavras estão no corpus e como estas se relacionam. Este tipo de algoritmo permite a descoberta de padrões entre os dados (palavras), levando, dessa maneira, a uma melhor compreensão do *corpus* em análise.

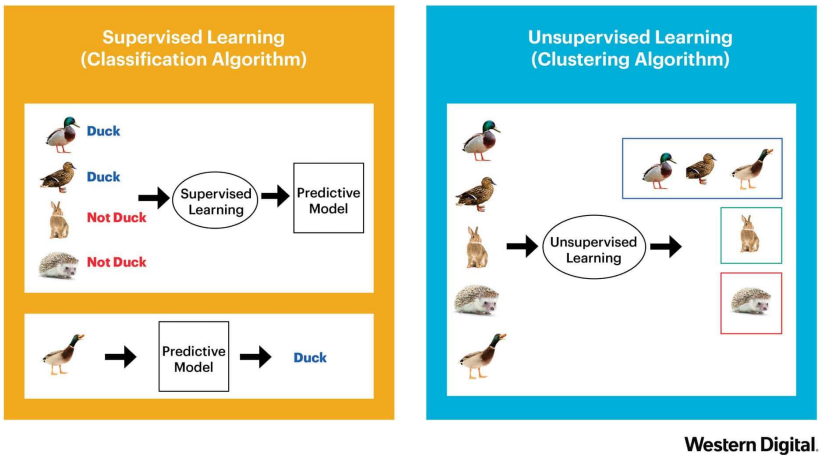
Os algoritmos de aprendizado de máquina são usados para várias aplicações, no entanto para este trabalho o foco está no processamento de linguagem natural aplicado em documentos textuais, a fim de identificar características, padrões e relações entre documentos e em cada documento.

Algumas características como: palavras que mais aparecem no texto, palavras que mais aparecem junto de outras palavras, tamanho das palavras, grupos de palavras que aparecem juntas (*clusters*), permitem que, mesmo sem prévio conhecimento do conteúdo de um repositório de documentos, seja possível fazer inferências quando tais documentos forem buscados em um sistema computacional composto por documentos digitais textuais e, desta forma, podem melhorar a assertividade sobre o conhecimento de um determinado *corpus*.

Para um exemplo e melhor compreensão dos tipos de aprendizado de máquina a figura 8 mostra do lado esquerdo o, **aprendizado supervisionado**, onde há um prévio processamento “ensinando” o algoritmo que determinadas imagens possuem um nome ou formalmente, um *“label”*, ou seja, um rótulo e quando o modelo, usando deste tipo de algoritmo, tenta prever uma imagem desconhecida ele “lembra” que determinada imagem parece com aquelas que ele aprendeu e então tenta inferir.

Do lado direito da figura 4 é mostrado o algoritmo de **aprendizado não supervisionado**, é aquele que não possui nenhum conhecimento prévio do conjunto de dados (dataset), ou seja, não houve um processo prévio “ensinando” rótulos das imagens e desta forma o que o algoritmo faz é um agrupamento por características semelhantes, o que é chamado formalmente de clusterização, ou seja, por não conhecer previamente as imagens este faz uma aglomeração das características comuns e este mesmo processo é aplicado sobre textos.

Figura 4 - Classificação e alocação de tópicos



Fonte: Western Digital (2021)

4.3.1 Aplicação de aprendizado de máquina e análise de tópicos

Os algoritmos de aprendizado de máquina são utilizados no processo de análise de documentos textuais digitais no sentido de usar o poder de processamento das máquinas para que as mesmas, sob a orientação destes algoritmos, possam ter como entrada uma grande quantidade de texto e como resultado desta entrada, estruturas que podem ser analisadas pelo ser humano.

Segundo Braga (2018), *Bag of Words* (*BoW*) é uma das técnicas base que são empregadas em um processo para análise de textos que cria uma estrutura que possa representar as palavras de um *corpus.* Esta técnica é base para os algoritmos de aprendizado de máquina e usa as palavras do texto como uma coluna e a quantidade de vezes que esta palavra aparece no texto (frequência) como outra coluna.

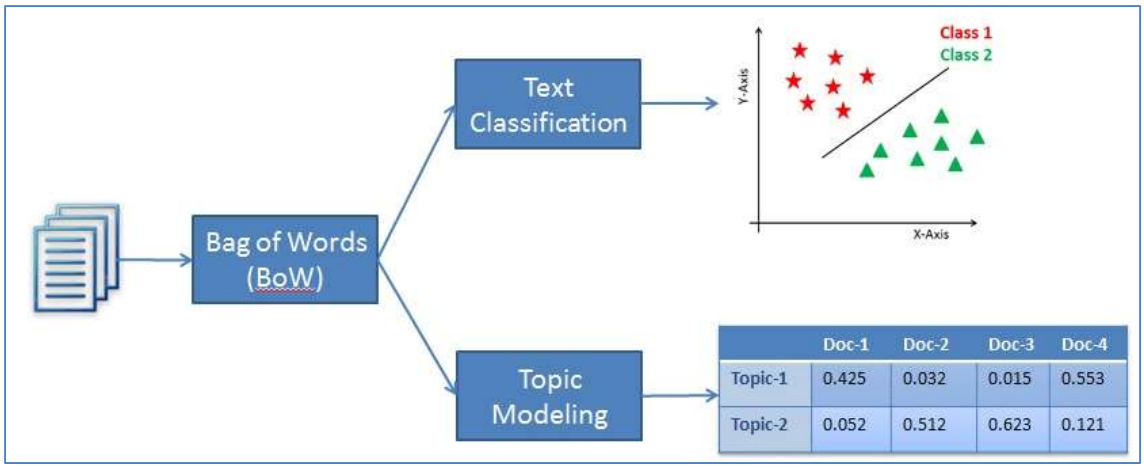
> A maioria dos algoritmos de agrupamento utiliza o modelo de saco de palavras (bag of words) para representar um documento. Este modelo gera alta dimensionalidade dos dados, ignora o fato de que diferentes palavras podem ter o mesmo significado e não considera o relacionamento entre elas, presumindo que as palavras são independentes umas das outras (BRAGA, 2018, p. 175).

Esta estrutura básica, *Bag of Words,* tem esse nome por não levar em conta o posicionamento de cada palavra, ou seja, se uma determinada palavra já foi encontrada

ela não possui precedência sobre a anterior. Esta técnica forma o alicerce para que outros algoritmos e outras técnicas possam ser aplicados no processo de análise de um corpus.

Dentre estas outras técnicas, destacam-se duas: a chamada alocação de tópicos (*Topic Modeling*) e a chamada classificação de texto (*Text Classification*) que podem ser vistas na figura 8.

Figura 5 - Classificação e alocação de tópicos



Fonte: DATACAMP (2018)

A figura 5 apresenta uma aplicação das técnicas supracitadas e destaca que ambas possuem como estrutura base a BoW como preparação para classificação de textos ou ainda para modelagem de tópicos.

A técnica conhecida como Classificação de Texto permite fazer uma segregação das palavras sobre um documento, ou ainda um arranjo com o intuito de apontar determinados clusters de palavras, já a técnica de *Topic Modeling* permite extrair uma análise de quais palavras formam um determinado tópico e quais tópicos são pertinentes ou tem maior predominância em determinado documento.

Segundo Lamba e Madhusudhan (2019), a **Análise de Tópicos** (*Topic Modeling*) é o processo de atribuir tópicos a um grupo de palavras que possuem uma alta frequência no texto em ordem decrescente por um processo automatizado para fins de gerenciamento e organização do texto em documentos.

Segundo Blei (2012), *Topic Modeling* situa-se como uma subárea da área de mineração de texto cunhado nos anos 1960, e ganhou notoriedade pelo aumento da

pesquisa na área de recuperação da informação. Nos anos 1990 teve maior capacidade de expansão e de pesquisa devido ao crescimento da quantidade de documentos digitais e pela capacidade de processamento dos computadores.

Já para Faleiros *et al.* (2016), o modelo de tópicos não é simplesmente um ranqueamento das palavras que mais ocorrem em um documento, mas sim uma análise de tópicos que possuem maior frequência real de aparição dentro de cada documento.

Neste sentido, *Topic Modeling* traz como resultado um conjunto de termos que podem ser relevantes dentro de um documento para serem pesquisados segundo alguma ferramenta em um sistema computacional e os tópicos extraídos são estruturas com valor semântico que formam grupos/*clusters* de palavras que frequentemente ocorrem juntas.

> [...] Esses grupos de palavras quando analisados, dão indícios a um tema ou assunto que ocorre em um subconjunto de documentos. A expressão “tópico” é usada levando-se em conta que o assunto tratado em uma coleção de documentos é extraído automaticamente, ou seja, um tópico é definido como um conjunto de palavras que frequentemente ocorrem em documentos semanticamente relacionados (FALEIROS *et al*., 2016, p. 9).

Para Blei (2012), *Topic Modeling* é base do modelo LDA ou *Latent Dirichlet Allocation*, este modelo permite uma análise de estruturas temáticas não estritamente expostas de um documento textual e, portanto, possibilitam melhor compreensão para análises de grandes volumes de texto.

Para um melhor entendimento do modelo LDA, o mesmo será apresentado na seção seguinte com seus detalhes e conceitos, uma vez que este é parte fundamental do modelo proposto no capítulo 5.

4.4 LATENT DIRICHLET ALLOCATION - LDA

Segundo Lamba e Madhusudhan (2019), o LDA é um modelo que permite encontrar um conjunto de tópicos em um documento. Estes tópicos podem ser entendidos como uma coleção de palavras mais recorrentes no documento e passam a poder representar o documento, pois tais tópicos possuem as palavras de maior frequência no documento e as que ocorrem também em posições próximas às outras e essa estrutura pode ser usada para análise de cada documento.

Para Wang *et al*. (2012), o LDA pode ser usado como componente auxiliar para

classificação textual inclusive em domínios do conhecimento onde a característica do texto em si é denominada de textos curtos (como os de algumas mídias sociais). Isto corrobora a utilização do LDA em vários tipos de aplicações, tanto para textos curtos quanto para textos com mais volume, demonstrando sua capacidade para análises textuais.

Esta característica do LDA, de possibilitar a representação de um documento como um conjunto de tópicos, é um importante passo para que sistemas computacionais possam fazer o reconhecimento de um repositório digital de documentos, pois, cada documento pode ser visto como uma coleção de tópicos que refletem o verdadeiro conteúdo de cada documento e isso possibilita a identificação do *corpus* baseado na análise de cada documento, ou seja, a partir de uma análise de menor granularidade pode-se analisar o corpus como um todo.

4.4.1 Funcionamento do LDA

O LDA tenta descobrir os tópicos de um determinado documento, inicialmente, como se fosse sob a forma de tentativa e erro. O processo para essa descoberta é realizado ao supor que uma determinada palavra deve pertencer a um determinado tópico aleatório em um primeiro momento.

No entanto, ao passo que o algoritmo analisa os documentos e suas palavras, determinadas palavras que ocorrem mais em determinados tópicos do que em outros são alocadas para este tópico em específico e, desta maneira, uma camada latente é descoberta, ou seja, um tópico que representa um conjunto de palavras.

Quadro 8 - Fórmula de probabilidade para tópicos e palavras do LDA

Etapas do processo de análise de Tópicos do LDA
A) Atribua aleatoriamente cada palavra de um documento a um dos possíveis tópicos latentes, exemplo: a palavra “design” de um documento é atribuída aleatoriamente ao Tópico B.
B) Em um segundo momento, analise cada palavra e sua designação para um tópico em cada documento e verifique:

1) com que frequência o tópico ocorre no documento? 2) com que frequência a palavra ocorre no tópico?
C) Com base nas informações acima, atribua a palavra um novo tópico. exemplo: parece que “design” não ocorre frequentemente no Tópico B (que mais se parece com outro tipo de tópico, por exemplo, sobre animais) e portanto, a palavra “design” provavelmente deve ser atribuída ao Tópico A ou outro tópico, mas não ao Tópico B, que possui maior frequência de palavras relacionadas a animais.
D) Passe milhares ou milhões de vezes por esse processo, ou seja, faça várias iterações disso. Eventualmente, os tópicos começarão a fazer sentido de uma maneira que possamos interpretá-los e dar-lhes temas, como "animais", "política", entre outros.

Fonte: elaborado pelo autor (2021)

Como pode ser analisado no quadro 8, o processo para definição dos tópicos latentes do algoritmo LDA passa por inúmeras iterações sob o texto do documento a fim de descobrir um determinado padrão para determinadas palavras. Esta iteração ao ser analisada mais formalmente chega-se a uma fórmula, o que é pertinente, já que o modelo LDA faz esta análise matematicamente através de computadores com grande aferição para processos repetitivos.

A figura 6 apresenta a fórmula para a probabilidade de geração dos tópicos para o modelo LDA que pode ser entendida como:

1) **Z** a representação de um Tópico
2) **W** representando palavras
3) **D** representando um Documento.

Figura 6 - Fórmula de probabilidade para tópicos e palavras do LDA

$$P(Z|W,D) = \frac{\#\ of\ word\ W\ in\ topic\ Z + \beta_w}{total\ tokens\ in\ Z + \beta} * (\#\ words\ in\ D\ that\ belong\ to\ Z + \alpha)$$

Fonte: elaborado pelo autor (2021)

A probabilidade de uma palavra **W** relacionar-se ou pertencer a um Tópico **Z** em um documento **D** é dada pela divisão do “número de palavras **W** no tópico **Z**” e pelo “número de palavras já existentes em **Z**” multiplicado pela quantidade de palavras no

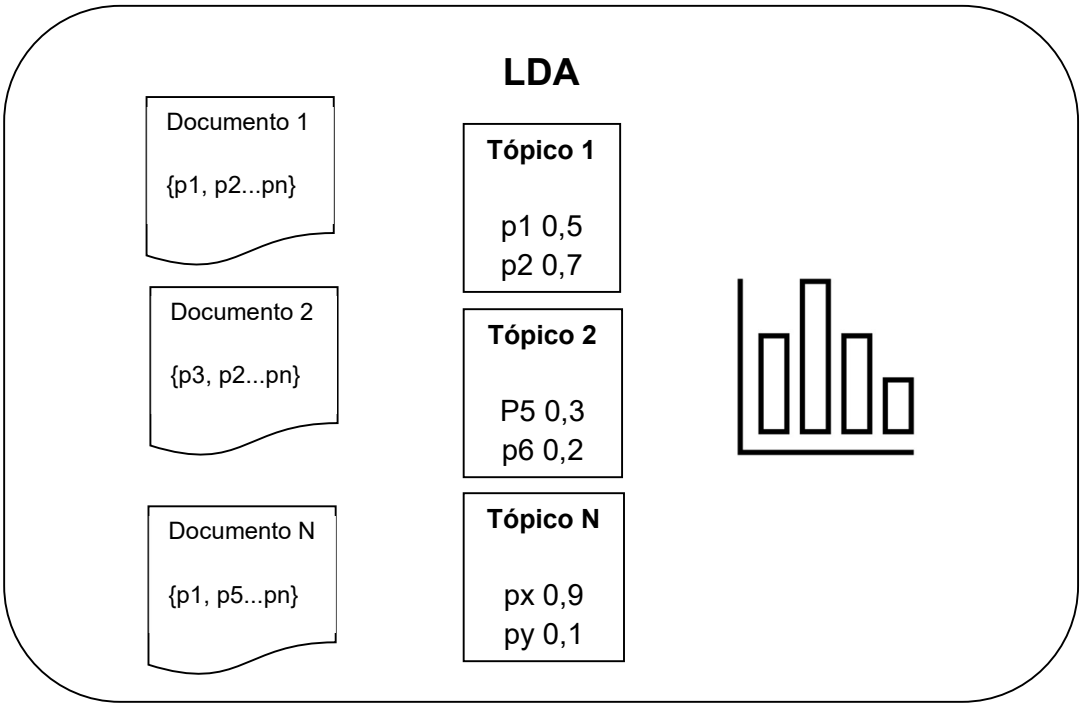
documento (não no tópico) que pertencem a **Z**.

Outros parâmetros fazem parte da fórmula para melhorar a aferição de palavras versus tópicos, são as letras gregas Alfa e Beta. Alfa indica a quantidade de tópicos que os documentos são compostos, que permite uma calibragem de tópicos por documento, ou seja, quanto maior o parâmetro Alfa, maior a quantidade de tópicos e assim mais específica é a distribuição de palavras por tópico e maior a aferição de palavras naquele tópico.

Já o parâmetro Beta indica quantas palavras compõe os tópicos, o que permite uma distribuição de palavras por tópico mais específica, ou seja, quanto a quantidade de palavras, mais específica a distribuição dentro do tópico.

Cada documento é composto por várias palavras, o modelo LDA analisa cada documento como uma combinação de tópicos.

Figura 7 – LDA e seu funcionamento para geração de Tópicos



Fonte: elaborado pelo autor (2021)

A figura 7 apresenta como esta análise é feita, pode-se ver que em um primeiro momento cada documento é apresentado como um conjunto de palavras D1 = {p1, p2, p33, 55}. Em um segundo momento, para cada documento analisado foi gerado uma

mistura de tópicos para ele, ou seja, o documento 1 possui uma mistura dos tópicos 1 e 3 por exemplo. Isto permite gerar análises como, quais documentos possuem maior predominância dos tópicos 1 e 2, quais possuem predominância somente do tópico 3, e desta maneira, selecionar determinados documentos conforme o aprendizado de máquina gerado pelo LDA.

Por fim, uma vez que o LDA identificou os tópicos de um determinado *corpus* e por conseguinte, também identificou as palavras que formam um determinado tópico, então, será possível utilizar estas palavras para comparação com as palavras-chave definidas pelo usuário e, desta maneira, aumentar a capacidade de análise comparativa, ou seja, as comparações serão sobre palavras de destaque em cada tópico e não sobre palavras gerais de todo o corpus.

4.5 WORD2VEC

Segundo Mikolov *et al.* (2013), o algoritmo Word2Vec aprende o significado das palavras ao processar um grande *corpus* de texto, mesmo que o *corpus* seja desconhecido e não tenha nenhuma identificação ou marcação, usando da incorporação de palavras como base para montagem do espaço vetorial.

O Word2Vec foi baseado no modelo neural de Bengio *et al.* (2003) e desenvolvido em 2012 por Thomas Mikolov que usou uma rede neural para predizer palavras similares contextualmente. A ideia é que, dada uma palavra alvo, as palavras que estariam perto teriam maior similaridade e para esta análise foi utilizada uma técnica chamada, vetor de palavras. Em 2013, quando trabalhava no *Google*, Mikolov fez o lançamento do algoritmo e deu o nome Word2Vec.

Em relação ao seu uso no modelo proposto, o algoritmo é utilizado com o objetivo de recuperar palavras similares para cada uma das palavras digitadas pelo usuário. Como pode ser visto na figura 11, a biblioteca aceita como entrada uma determinada palavra e como saída obtém-se dez palavras similares semanticamente baseado no treinamento do modelo NILC.BR.

Figura 8 - Espaço vetorial com a incorporação de palavras

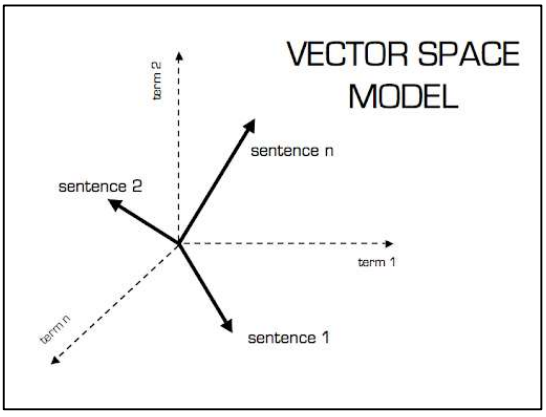
```
#Lista das 10 palavras mais similares usando a semântica do Word2Vec+NILC Model
similares = model.similar_by_word('dpoc')
for i in similares:
  print(i[0])
```

```
nefrótica
obstrutiva
chediak-higashi
hipercalcemia
cardiopatia
celíaca
imunossupressão
isquémica
uretrite
disfunção
```

Fonte: elaborado pelo autor (2021)

Já na figura 9, encontra-se uma generalização da aplicação do algoritmo como realizado no parágrafo acima, ou seja, cada palavra irá se encontrar em um posicionamento espaço vetorial, e cada palavra terá sua coordenada neste espaço e poderá ser alcançada se comparada, através de cálculos matemáticos como a distância do cosseno (ver seção 4.5.2), para mostrar sua relação com as demais palavras.

Figura 9 - Espaço vetorial com a incorporação de palavras



Fonte: adaptado de Mikolov *et al*. (2003)

Para fazer a assimilação das palavras de um determinado corpus o algoritmo Word2Vec utiliza do aprendizado não supervisionado, ou seja, aprende diretamente a partir do conjunto de dados fornecido, no caso, um *corpus* de um repositório digital. Além

disso, habilita grande variação de aplicações, uma vez que grande parte dos textos no mundo real são não estruturados e não possuem marcação para auxílio computacional no momento de extração de informação.

Segundo Mikolov (2013), o Word2Vec foi desenvolvido para trabalhar com grandes corpora, da ordem de bilhões de palavras. Desta maneira consegue aprimorar sua efetividade em categorizar vetores de palavras. Basicamente, possui duas formas de funcionamento, o CBOW e o *Skip-Gram* que serão analisados da seção seguinte.

4.5.1 CBOW e Skip-Gram

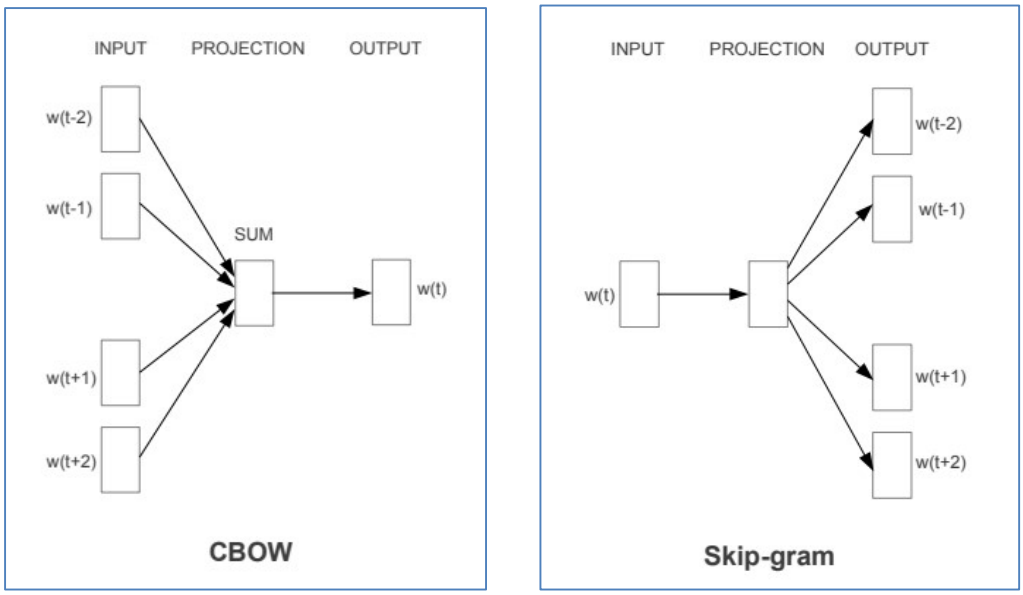
O objetivo da aplicação do Word2Vec é possibilitar ao modelo encontrar palavras semanticamente próximas baseadas em um treinamento executado anteriormente usando, no caso deste trabalho o vetor de palavras do NILC.BR.

Neste sentido, para que de fato extraia-se palavras contextuais dadas palavras de de contexto o algoritmo Word2Vec possibilita duas arquiteturas ou formas de aplicação, a conhecida como *Continuous Bag of Word* (CBOW) e a *Skip-Gram*.

A arquitetura CBOW possibilita que, dada uma gama de palavras de contexto, como pode ser visto na figura 10, definidas como w(t-2), w(t-1), w(t+1) e w(t+2), este tipo de arquitetura permite predizer uma dada palavra alvo, no caso w(t). Vale ressaltar que mesmo que as palavras de contexto estejam foram de ordem, a arquitetura CBOW possibilita recuperar a palavra contextual alvo, e esta característica torna-se interessante para aplicações nas quais deseja-se obter uma determinada palavra dadas outras em uma frase ou texto.

Já a arquitetura *Skip-Gram* é o inverso, ou seja, dada uma determinada palavra como input, na figura representada por w(t), o modelo consegue gerar as possíveis palavras, w(t-2), w(t-1), w(t+1) e w(t+2), que estão ao redor da palavra w(t). Esta é a arquitetura utilizada no modelo proposto.

Figura 10 - Modelo CBOW e Skip-Gram do Word2Vec

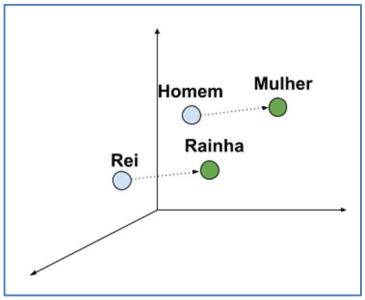


Fonte: Mikolov *et al.* (2013)

Um exemplo de como o Word2Vec pode ser utilizado pode ser visto na figura 11, na qual são apresentados em um espaço vetorial euclidiano as palavras: “Homem”, “Mulher”, “Rei” e “Rainha”. Para compreensão humana, o exemplo está em três dimensões, mas o Word2Vec pode utilizar até mais de trezentas dimensões para representação dos vetores de palavras.

A figura mostra uma relação vetorial entre as palavras e comparações como “Rei” está para “Rainha” assim como “Homem” está para “Mulher” são relações que possibilitam que o modelo retorne que estas palavras estão semanticamente relacionadas. Ou seja, a partir de uma busca do usuário, toma-se as palavras-chave definidas e analisa-se junto ao Word2Vec quais são as palavras de contexto para cada palavra da busca.

Figura 11 - Exemplo de aplicação do Word2Vec



Fonte: adaptado de Mikolov *et al.* (2013)

Na seção seguinte, pretende-se explorar como o modelo Word2Vec possibilita esta análise de similaridade entre estes vetores de palavras. Como o cálculo de similaridade do cosseno permite que matematicamente palavras semelhantes em seu contexto podem ser encontradas.

4.5.2 Similaridade do cosseno

Segundo Esposito *et al.* (2020), a similaridade semântica entre duas palavras é dada por um cálculo de similaridade do cosseno ou, como também pode ser encontrado na literatura, a semelhança do cosseno. Esta é uma medida entre dois vetores de palavras em um espaço vetorial, um vetor pode ser visto na figura 12, a qual representa a palavra gato em um espaço vetorial de dimensões elevadas, ou seja, as palavras são representadas desta forma a qual para compreensão humana pode ser vista na figura 13 em 3 dimensões.

Ao extrapolar essa visualização desta palavra para milhares de palavras dá-se o que é chamado de espaço vetorial e, portanto, quando são analisadas 2 ou mais palavras pode-se verificar o quão perto ou quão longe uma palavra está da outra, permitindo dessa maneira que os computadores consigam ter compreensão da linguagem natural.

Portanto, o cosseno do ângulo formado entre estes dois vetores de palavras é que vai exprimir o quão perto uma palavra está da outra, ou seja, quão semelhantes semanticamente estas palavras são, uma vez que palavras com mesmo contexto, pela hipótese distributiva, possuem significados semelhantes.

Desta forma, matematicamente torna-se possível avaliar um *corpus* com milhares de palavras, documento a documento, possibilitando-se aplicações em vários corpora de texto não estruturado.

Neste contexto é que o algoritmo Word2Vec se encaixa e torna possível fazer analogias entre palavras utilizando a distância do cosseno e trazendo de forma lógico-matemática a relação entre as palavras de um determinado texto, ou seja, de documentos em um sistema computacional.

Figura 12 – Representação vetorial de uma palavra

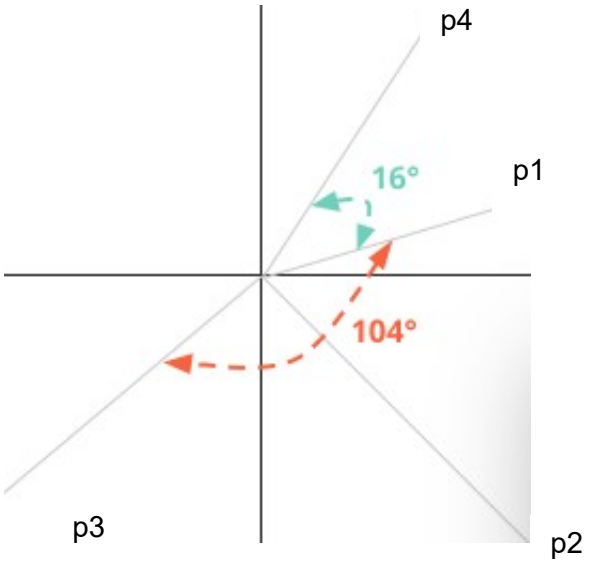
```
model.get_vector('gato')
```

```
array([-0.25352 , -0.085966, -0.088982, -0.113394,  0.008701, -0.150444,
        0.156058, -0.284607,  0.254255,  0.162898,  0.256427, -0.254399,
       -0.017484, -0.137932, -0.210804, -0.199106,  0.433588, -0.246379,
       -0.390673,  0.229393,  0.035672,  0.522252,  0.182892, -0.37677 ,
        0.014336, -0.261313, -0.090331,  0.236965,  0.131862,  0.149726,
        0.240298, -0.118945, -0.077243, -0.183553,  0.120625, -0.025893,
        0.189624, -0.234493, -0.511314, -0.221332,  0.321519,  0.341439,
       -0.118699,  0.447367, -0.252668,  0.118569,  0.296533,  0.184336,
       -0.013855, -0.037908], dtype=float32)
```

Fonte: elaborado pelo autor (2021)

Portanto, o cálculo do cosseno se dá pela análise dos vetores das palavras, como exemplo, seja *d1* a distância entre os vetores das palavras p1 e p4 e seja *d2* a distância entre as palavras p1 e p3, o seguinte cálculo do cosseno dos ângulos formados por estas palavras é mostrado na figura 13.

Figura 13 – Distância do cosseno e semelhança por cosseno

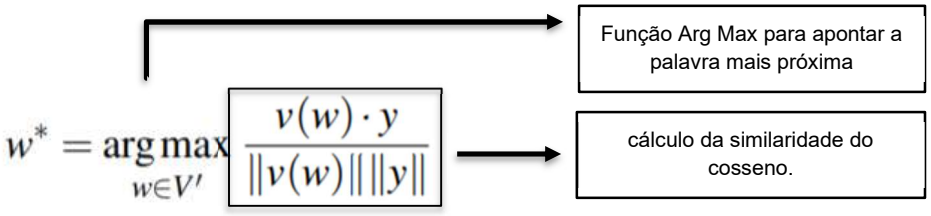


Fonte: adaptado de Maroti (2019)

O cálculo do cosseno pode ser visto na figura 14, o qual se faz necessário para designar se uma palavra está próxima ou não de outra palavra, ou seja, se o resultado de *d1* for maior que o resultado de d2, isso quer dizer que a proximidade das palavras p1 e p4 é maior que das palavras p1 e p3.

Desta forma pode-se inferir através de cálculos matemáticos, que estão embutidos em comandos de linguagens de programação, que a proximidade entre palavras possa exprimir sua relação semântica.

Figura 14 - Fórmula para aproximação de palavra por semelhança de cosseno



Fonte: elaborado pelo autor (2021)

Isto posto, conclui-se que ao utilizar-se do algoritmo Word2Vec, ganha-se a possibilidade de usufruir de uma estrutura vetorial para representação das palavras e, desta maneira, pode-se retornar palavras similares semanticamente, tarefa chave do modelo proposto em sua segunda fase para expandir contextualmente as palavras-chave do usuário.

4.6 CONSIDERAÇÕES SOBRE PLN, HISTÓRICO E ABORDAGENS

Segundo Zhai e Massung (2016), o Processamento de Linguagem Natural é uma área de pesquisa interdisciplinar que está atenta em utilizar-se e desenvolver técnicas computacionais com o intuito de que uma máquina possa “compreender” um texto escrito ou falado em linguagem de um ser humano, linguagem natural.

Ainda conforme Zhai e Massung (2016), o processamento de linguagem natural é uma subárea da Inteligência Artificial que processa linguagens naturais e transforma em números cada palavra, o que possibilita endereçar estes números em vetores, ou seja, estruturas que permitem armazenar múltiplos dados em um único endereço na memória do computador e assim possibilitar acesso mais contextual a estes dados.

Esta é a exata noção do significado de uma palavra para um computador, pois para uma máquina esta não possui a capacidade de entender a diferença entre uma palavra e outra, mas máquinas têm grande capacidade de processar, fazer cálculos e,

ao transformar-se uma palavra em um número, tais cálculos tornam-se possíveis e abrem uma miríade de possibilidades no campo do processamento de linguagem natural.

Segundo Shalaby *et al.* (2016), as estruturas de um texto não estruturado são na verdade mais ricos em relacionamentos do que propriamente um texto estruturado. Pois, as relações entre as palavras de um texto podem ser expandidas semanticamente ao se utilizar de outras bases, como a utilização de bases georreferenciais.

Ainda de acordo com Shalaby *et al.* (2016), este tipo de relacionamento, onde mesmo um texto não estruturado possa ser enriquecido por bases informacionais externas, pode permitir um contexto ainda maior, o que foi chamado de hiperrelacionamento.

O PLN não é uma área tão recente, trabalhos com análises textuais remontam à década de 1950 quando as primeiras aplicações com automação de traduções foi demonstrado, a chamada, máquina de tradução foi um projeto da IBM-Georgetown (HUTCHINS, 2004), no qual envolveu a tradução automática de mais de sessenta frases russas para o inglês.

Na década de 1960, uma variação dos sistemas de processamento de linguagem natural, conhecido como NLU (*Natural Language Understanding*) ou em português Sistemas de Compreensão de Linguagem Natural, foi desenvolvido. Neste projeto, o usuário tenta manter uma conversa com o sistema. Foi considerado um grande avanço na área de inteligência artificial e um dos grandes representantes foi o sistema SHRDLU escrito por Terry Winograd (WINOGRAD, 1971) que possibilitava que blocos de textos em forma de perguntas ou comandos simulava uma interação contextual homem máquina.

Já na década de 1970, houve algumas iniciativas chamadas de “ontologias conceituais” nas quais palavras e conceitos foram estruturados de tal forma que os computadores conseguem “entender” o relacionamento entre estes artefatos e inferir respostas, o que permitiu o crescimento dos *chatterbots* (GRUBER, 1991).

O aprimoramento da PLN começa a surgir a partir da década de 1980 quando algoritmos de aprendizado de máquina começam a ser utilizados no processamento de linguagem natural e, basicamente, estes intentos levaram a classificação atual de que são 3 as abordagens para resolução de problemas relacionados ao Processamento de

Linguagem Natural: a simbólica; a estatística ou também conhecida como abordagem tradicional que se utiliza de vertentes da área de aprendizado de máquina; e a abordagem baseada em redes neurais que se utiliza de aprendizado profundo (*deep learning*).

A abordagem simbólica reúne eventos dos anos de 1960 e 1970 na qual as relações eram fortemente baseadas em regras linguísticas bem estruturadas, mas que já provou ser eficiente e que atualmente consegue atingir bons resultados mesmo em *corpus* com grande quantidade de palavras. Por sua característica de conhecimento sobre classes gramaticais e regras das mesmas é possível se extrair informação somente ao reunir determinados verbos ou entidades como pessoas ou empresas.

Na abordagem estatística é utilizado aprendizado de máquina para aprender sobre um determinado *corpus* para que se permita fazer inferências sobre o domínio do *corpus* através de métodos estatísticos. O que possibilita agrupar conjuntos de palavras ou de documentos que não são visíveis ao primeiro momento, criando oportunidades de aplicação como similaridade de documentos.

4.6.1 CONCEITOS E APLICAÇÕES DO PLN

Alguns conceitos são importantes como referencial teórico para compreensão da área de processamento de linguagens naturais no escopo deste trabalho. Antes são apresentadas algumas aplicações do PLN.

Para Lane, Howard e Hapke (2019) em seu livro *Natural Language Processing in Action*, o processamento de linguagem natural é ubíquo, está por toda parte e na figura 15 destaca-se aplicações de forma categorizada.

Figura 15 - Categorização das aplicações usando PLN

Aplicação	Categoria 1	Categoria 2	Categoria 3
Search	Web	Documents	Autocomplete
Editing	Spelling	Grammar	Style
Dialog	Chatbot	Assistant	Scheduling
Writing	Index	Concordance	Table of contents
Email	Spam filter	Classification	Prioritization
Text mining	Summarization	Knowledge extraction	Medical diagnoses
Law	Legal inference	Precedent search	Subpoena classification
News	Event detection	Fact checking	Headline composition
Attribution	Plagiarism detection	Literary forensics	Style coaching
Sentiment analysis	Community morale monitoring	Product review triage	Customer care
Behavior prediction	Finance	Election forecasting	Marketing
Creative writing	Movie scripts	Poetry	Song lyrics

Fonte: adaptado de Lane, Howard e Hapke (2019)

Pode-se observar que o processamento de linguagem natural está presente nos mais variados âmbitos de aplicação e nas mais variadas áreas, sejam científicas ou já atuantes no mercado. Neste contexto, também a gama de conceitos atrelados são variados, e pertencem às várias áreas como: Estatística, Matemática, Computação, Linguística e Ciência da Informação.

Em resumo, a área de Processamento de Linguagem Natural permite um amplo aspecto quando se olha do prisma de aplicações. No entanto, para que estas aplicações sejam possíveis, o cerne de um sistema que interpreta uma linguagem natural está voltado para o entendimento das palavras por uma máquina.

Para efeito de comparação, um programa de computador deve ser escrito em alguma linguagem para então ser executado por uma máquina. Este texto escrito por um ser humano, é apresentado à máquina que faz uma interpretação deste texto, tornando-o em um tipo de texto específico que é chamado de linguagem de programação.

Segundo Fischer e Grodzinsky (1993), uma linguagem de programação é definida como um conjunto de comandos (palavras) que devem ser escritas seguindo algumas

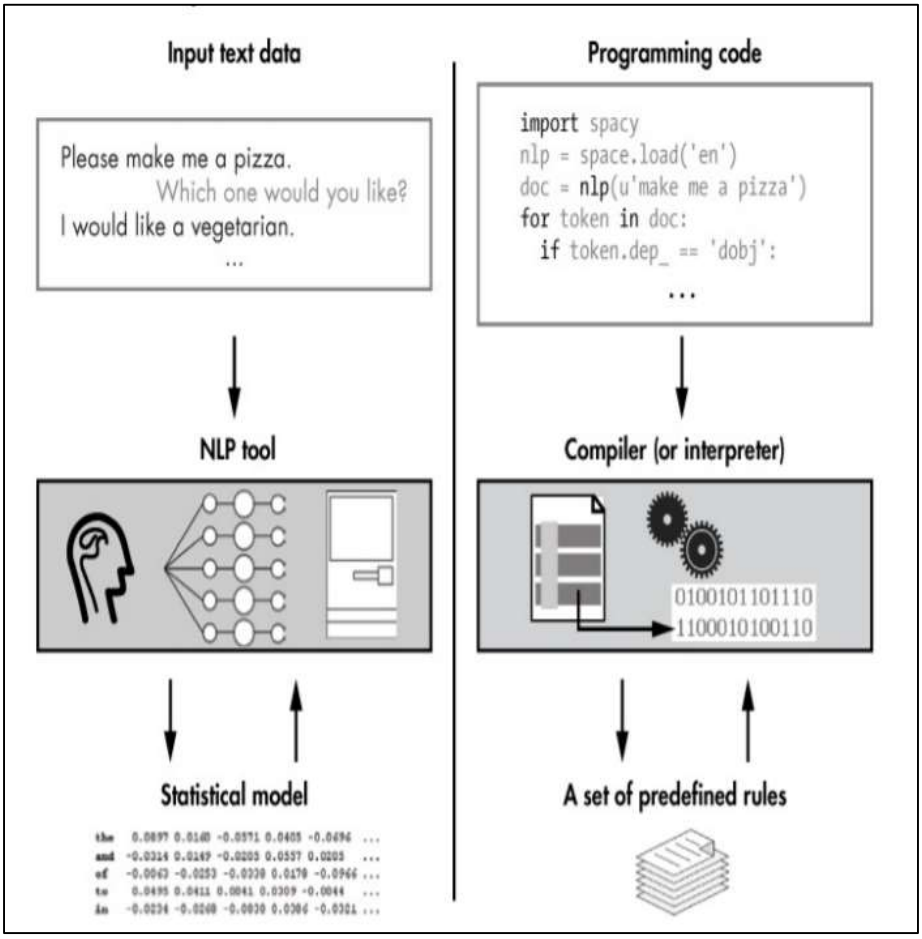
regras sintáticas para que, quando forem empacotadas em um arquivo, possam ser executados em uma máquina que irá interpretar tais comandos e fazer exatamente o que foi escrito.

Este conjunto de comandos de uma linguagem de programação é restrito, como exemplo a linguagem de programação Java em que são somente 61 palavras, as quais são interpretadas pela máquina de maneira satisfatória, ou seja, há uma compreensão pela máquina de cada palavra escrita pelo programador.

No entanto, para um modelo de PLN conseguir entender um texto em algum idioma (outra linguagem) o conjunto de palavras é bem maior que da linguagem de programação Java, por exemplo em 1989, o dicionário Oxford para o Inglês possuía 171.476 palavras distintas. Portanto, uma outra abordagem deve ser empregada para que os modelos de PLN possam ter maior capacidade de entendimento de textos.

A figura 16 explana exatamente estes tipos de abordagens que uma máquina precisa ter para conseguir entender um texto. Do lado direito da figura, a máquina faz uma interpretação de um texto escrito nos comandos de uma linguagem de programação, ou seja, há um conjunto limitado de palavras e regras bem definidas. Já do lado esquerdo da figura, um texto escrito no idioma inglês é apresentado como entrada para que um sistema, utilizando do processamento de linguagem natural, possa interpretá-lo e gerar ações sobre este texto e, portanto, a abordagem para este processo utiliza modelos estatísticos para a interpretação, o qual será apresentado na seção seguinte.

Figura 16 - PLN e regras pré-definidas versus Modelos Estatísticos



Fonte: adaptado de Vasiliev (2020)

4.6.2 Etapas do processamento de linguagem natural

Pode-se observar que são vários os conceitos e procedimentos embutidos ao processar um texto, estes procedimentos seguem um fluxo de instruções e ações ao qual é dado o nome de **pipeline**, ou seja, uma segmentação de instruções para resolução de um problema ligado ao processamento de textos em linguagem natural.

Como pode ser visto, a figura 17 mostra passos para os processos do pipeline de desenvolvimento de um modelo de PLN. Cada processo do pipeline é uma contribuição para a análise estatística do modelo, ou seja, são processos que auxiliam este tipo de

modelo.

Figura 17 - Pipeline de um Processamento de Linguagem Natural

Fonte: elaborado pelo autor (2021)

4.6.2.1 Segmentação de Sentença e Tokenização

Segundo Lane, Howard e Hapke (2019), o primeiro processo no pipeline para o processamento de linguagem natural é a **segmentação de sentenças** na qual o texto é quebrado em pequenas partes chamadas de segmentos, que podem ser um parágrafo que então pode ser quebrado em sentenças e estas sentenças são quebradas em frases as quais são quebradas em palavras.

Ainda segundo Lane, Howard e Hapke (2019), a fase chamada de Tokenização (*Tokenization*) é uma fase especializada da fase de segmentação e nela se segmenta ainda mais o texto recebido. É o processo de “quebrar” cada palavra do texto, um **token**. Este processo permite que o modelo estatístico possa, por exemplo, fazer uma contagem do número de determinada palavra em um texto ou ainda identifique as de maior frequência em conjunto com outra palavra.

4.6.2.2 Part of Speech

Segundo Vasiliev (2020), a fase chamada de *Part of Speech* produz uma análise sintática do texto. Nesta fase são feitas análises de classes gramaticais das palavras como substantivo, verbo, adjetivos, entre outros que auxiliam na contextualização das palavras e inferências através do tempo verbal das palavras, avaliando se as mesmas

estão no infinitivo, passado, futuro ou presente, em qual pessoa e se estão no singular ou plural.

4.6.2.3 Lematização e Stemização

Segundo Vasiliev (2020), a lematização (lemmatization) é o processo de converter todas as variantes de uma palavra para seu lema, ou seja, caso for um substantivo esta é transformada em sua forma no masculino e singular e caso sejam verbos vão para o infinitivo. Por exemplo, se a palavra analisada for “cachorro” ou “cachorra“ ou ainda, “cachorros” a lematização de todas estas palavras seriam o lema: “cachorro”. Já para verbos, quando encontrado alguma inflexão esta é retornada ao seu lema, como “andando” seu lema será “andar”.

Já o processo de stemização (*stemming*), segundo Vasiliev ***et al***. (2020), consiste em reduzir uma palavra ao seu radical, por exemplo, se a palavra encontrada é “menino” ou “menina” ou ainda “menininho” sua stemização ficaria “menin” o que reduz a quantidade de variações da palavra e auxilia também em inferências.

4.6.2.4 Stop Words

Segundo Lane, Howard e Hapke (2019), a fase no pipeline chamada de “*Stop Words*” ou palavras de parada, consiste no processamento para retirar do texto palavras que não vão adicionar valor semântico para uma análise, são exemplos de palavras com alta frequência como “a”, “o”, “de”, “do”.

4.6.2.5 NER - Named Entity Recognition

Conforme Lane, Howard e Hapke (2019), a fase de *Named Entity Recognition* ou reconhecimento de entidades é a fase em que o processo se traduz em tentar reconhecer um nome próprio que no caso pode ser nome de pessoas, de empresas, de localidades, entre outras entidades. O processo de NER é especialmente importante para verificação

de textos nos quais se necessita reconhecer um nome de uma empresa ou de uma pessoa, por exemplo em um *dataset* com posts de redes sociais.

4.7 HIPÓTESE DISTRIBUTIVA E SEMÂNTICA VETORIAL

Segundo Harris (1954, p. 146), a hipótese distributiva diz que “palavras que ocorrem em contextos similares tendem a ter sentidos ou significados similares”. Esta hipótese está diretamente ligada ao processo de recuperação da informação semântica de um *corpus*, já que os procedimentos utilizados no processamento de linguagem natural somados aos conceitos dos modelos estatísticos possibilitam inferir probabilidades de frequência de palavras em contextos similares.

Ainda de acordo com Harris (1954), a Hipótese distributiva é o fundamento principal para os modelos estatísticos, pois as palavras (*tokens*) de um *corpus*, ao serem mapeadas para uma estrutura matricial na qual cada palavra possui um endereço de linha e coluna e palavras de contexto similar estarão em linhas e colunas que estão perto uma das outras, possibilitam avaliar matematicamente relacionamentos entre estas palavras.

A esta estruturação das palavras em linhas e colunas de uma matriz dá-se o nome de formação de **vetores de palavras** (ver 4.6). Este conceito é a base para os chamados modelos baseados em contagem, nestes modelos a formação dos vetores de palavras é baseada na frequência com que as palavras ocorrem.

Para que os vetores de palavras sejam montados, é necessário a criação do **vocabulário** (também encontrado na literatura como **Dicionário**), ou seja, cada *token* extraído na fase de Tokenização vai pertencer ao vocabulário, que forma a base para construção dos vetores de palavras (*word vectors*).

Esta representação de um texto de um documento em um vetor de palavras é conhecida como *Bag of Words* (BoW) ou, em uma tradução direta, saco de palavras. É dado este nome para poder explicar que as palavras, ora tokenizadas e colocadas nesta grande estrutura, não foram colocadas com nenhum tipo de ordenação.

Uma explicação mais detalhada cabe neste momento para dizer que uma frase que contenha números, por exemplo: Elmo Street, 215. Esta frase, ao ser tokenizada permitirá que o modelo que está processando este texto irá entender que não serão

somente 3 *tokens* separados: (Elmo), (Street), (215), mas sim um único *token* (Elmo Street, 215) e esta forma auxilia na inferência e descobrimento de contexto para datas e números aplicados no texto.

Um vocabulário ainda não é forma que uma máquina “entende” uma palavra, esta estrutura deve ainda ser representada em números para que a máquina possa inferir. Esta representação numérica das palavras ora tokenizadas é conhecida como *one-hot vectors*, que são sequências de zero e um para representar cada palavra do texto.

A título de exemplo, pode ser visto na tabela 1 uma representação vetorial para a seguinte frase: **às 19:00 Sr. John foi ao seu escritório.** A frase foi passada para uma estrutura de vetores, isto é, cada palavra é mapeada no vetor (quando se tem um número 1) e todas as outras posições ficam zeradas.

Esta estrutura representa bem as palavras do dicionário e permite que uma máquina a consiga ler e fazer inferências, mesmo assim, esta estrutura é conhecida por ser muito esparsa para textos muito grandes com milhares de palavras, o que se torna de difícil armazenamento em memória.

Todavia, é uma estrutura que pode representar palavras de forma completa para um *corpus* e permite a criação de um modelo de PLN para inferências. Por exemplo, a palavra “John” é representada pelo vetor [0,0,0,1,0,0,0,0].

Tabela 1 - One-hot vectors

	Às	*19:00*	*Sr.*	*John*	*Foi*	ao	seu	escritório
0	0	0	**1**	0	0	0	0	0
1	0	0	0	0	**1**	0	0	0
2	0	**1**	0	0	0	0	0	0
3	0	0	0	0	0	**1**	0	0
4	0	0	0	0	0	0	**1**	0
5	0	0	0	0	0	0	0	**1**
6	0	0	0	**1**	0	0	0	0
7	**1**	0	0	0	0	0	0	0

Fonte: elaborado pelo autor (2021)

Este processo é repetido para todas as palavras do vocabulário, permitindo análises matemáticas, mais especificamente utilizando álgebra linear para fazer inferências dentro das várias aplicações do PLN, uma vez que esta matriz pode ser

plotada em um espaço vetorial, sendo cada palavra um vetor em “n” dimensões de palavras.

No entanto, este processo (BoW) não expõe de fato quais palavras são mais relevantes em um documento, passo importante para o processo de recuperação da informação. O processo feito acima é conhecido como TF (*Term Frequency*) ou Frequência de Termos, no qual é feito uma contagem de quantas são as palavras e quantas vezes ela se repete (no exemplo acima, caso uma palavra fosse repetida teria o valor 2 e assim por diante).

Um termo que ocorre várias vezes em um documento pode não ser o termo ao qual representa um documento. Uma palavra ao ter uma alta frequência em um determinado *corpus* pode ser uma palavra importante, no entanto, pode ser que esta palavra seja tão frequente que não irá explicitar qual sua relevância dentro do documento, ou seja, palavras menos frequentes mas que são únicas, podem expressar melhor sua importância no documento.

Segundo Cabrera Diego (2011), o TF-IDF (*Term Frequency Inverse Document Frequency*) é um método que extrai uma lista de palavras, segundo avaliação do quão relevante uma determinada palavra (termo) é em um documento em relação ao *corpus* como um todo.

O TF-IDF fornece um peso para as palavras dentro do documento e, por conseguinte, dentro do *corpus*, ou seja, representa tanto de forma local quanto global uma palavra e isso permite análises mais precisas e retornos para cenários de Recuperação da Informação ainda mais assertivos por dar um peso a cada palavra contextualmente ao *corpus* e não somente verificar a frequência em que aparece no texto, como a primeira abordagem.

Para melhor esclarecimento de como um documento torna-se um vetor dentro de um espaço vetorial, tomar-se-á um exemplo com duas sentenças que serão representadas como dois documentos (d1 e d2), sendo-os:

$$\begin{cases} \text{d1 = “O entardecer é belo”} \\ \text{d2 = “lindo o sol ao entardecer, sempre ao entardecer”} \end{cases}$$

Usa-se, em um primeiro momento, o conceito de termo frequência para representação de cada termo no espaço vetorial, lembrando que o termo frequência (TF) é uma medida de quantas vezes os termos estão presentes do dicionário. Portanto, o primeiro passo é a criação do dicionário, esta estrutura que vincula a frequência da palavra em cada documento:

	Entardecer	belo	lindo	sol	sempre
d1	1	1	0	0	0
d2	2	0	1	1	1

Portanto, o **vetor** a partir do dicionário para cada documento usando TF é dado como:

d1 = (1, 1, 0, 0, 0)
d2 = (2, 0, 1, 1, 1)

ou para representação do corpus pode-se definir como uma **matriz**:

$$\begin{bmatrix} 1, 1, 0, 0, 0 \\ 2, 0, 1, 1, 1 \end{bmatrix}$$

A apresentação na forma acima é definida como não normalizada, ou seja, deve-se aplicar a normalização calculando-se o TF-IDF de cada palavra no contexto do *corpus* e assim se obtém uma equanimidade em relação às palavras do *corpus*.

Desta maneira, obtém-se vetores de palavras balanceados que podem ser utilizados para verificação de similaridade, ou seja**,** este é o princípio do conceito chamado **semântica vetorial**, na qual palavras estão normalizadas e possuem um endereço em uma matriz que representa cada palavra do *corpus*.

4.8 RECUPERAÇÃO DA INFORMAÇÃO BASEADA EM NOÇÃO DE SIMILARIDADE

Segundo Mitra *et al*. (2017), a recuperação da informação vai além da relação dos termos de busca e sua evidência ou não nos documentos, e sugere que um modelo de

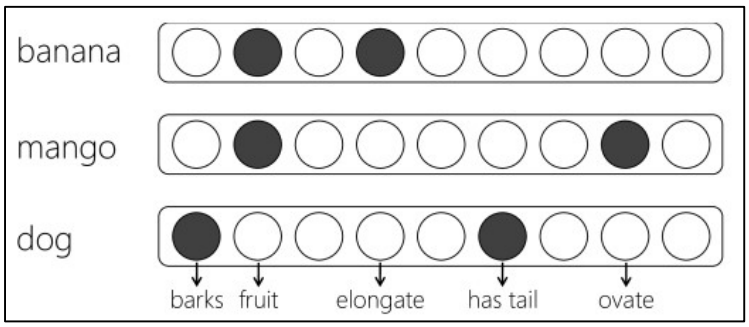
busca deve levar em conta a **noção de similaridade** para que as buscas tenham capacidade de analisar termos adjacentes aos buscados e desta forma aumentar a capacidade de recuperação de informação contextual para o usuário.

A noção de similaridade, segundo Mitra *et al*. (2017), se dá por posicionamento de palavras que estão próximas umas das outras em um vetor de palavras como pode ser visto na figura 18. O vetor de palavras é uma estrutura que armazena cada palavra de um determinado documento e possibilita avaliar quais palavras são mais semelhantes às outras pelo seu posicionamento em um espaço vetorial.

Como pode ser visto na figura 18, as palavras mango e banana estão marcadas para a característica fruta, mas não com a palavra dog. Isto quer dizer que em um contexto (por exemplo um repositório de documentos textuais) haverá documentos que possuem as palavras mango e banana e estas em algum momento estão relacionadas com a palavra fruta, ou seja, no contexto de cada documento estas palavras estão relacionadas e a palavras fruta determina a relação com mango e banana, esta palavra determinante é, no caso, chamada como uma característica.

O mesmo acontece no contexto de documentos que possuem relação com a palavra dog, ou seja, haverá documentos nos quais a palavra aparecerá e contextualmente outras palavras a sua volta determinarão a palavra dog o que permite encontrar documentos nos quais as características determinarão se um determinado documento conterá ou não determinadas palavras, este é o conceito de similaridade semântica, no qual palavras, características e documentos se relacionam e algoritmos como Word2Vec possibilitam fazer uma análise matemática do corpus.

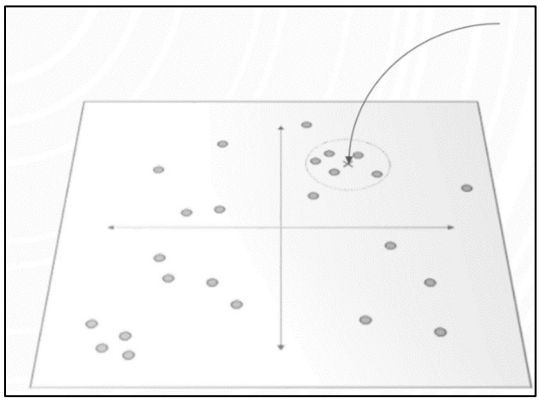
Figura 18 - Vetor de palavras



Fonte: adaptado de Mitra *et al*. (2017)

Neste sentido, como corroborado pela hipótese distributiva de Harris (1954), palavras de contextos semelhantes tendem a ter significados semelhantes, e tendem a pertencer a um mesmo grupo de palavras como pode ser visto na figura 19, ou seja, existem palavras que possuem um relacionamento muito maior do que outras.

Figura 19 - Agrupamento de palavras



Fonte: adaptado de Mitra *et al*. (2017)

Portanto, o processo de recuperação da informação ao se utilizar da noção de similaridade, o qual foi abordado nos parágrafos anteriores e apoiado por relações matemáticas como vetores de palavras e algoritmos como o Word2Vec podem propiciar recuperação da informação de maneira contextual e semântica para os usuários.

No entanto, tais conceitos, tecnologias e algoritmos precisam se relacionar para que a recuperação da informação semântica de fato aconteça. Para isso, na seção posterior, o capítulo 5, será apresentado o modelo que proporciona a correlação destes artefatos e que também utiliza do algoritmo LDA e intermédio de codificação com a linguagem de programação para implementação do modelo.

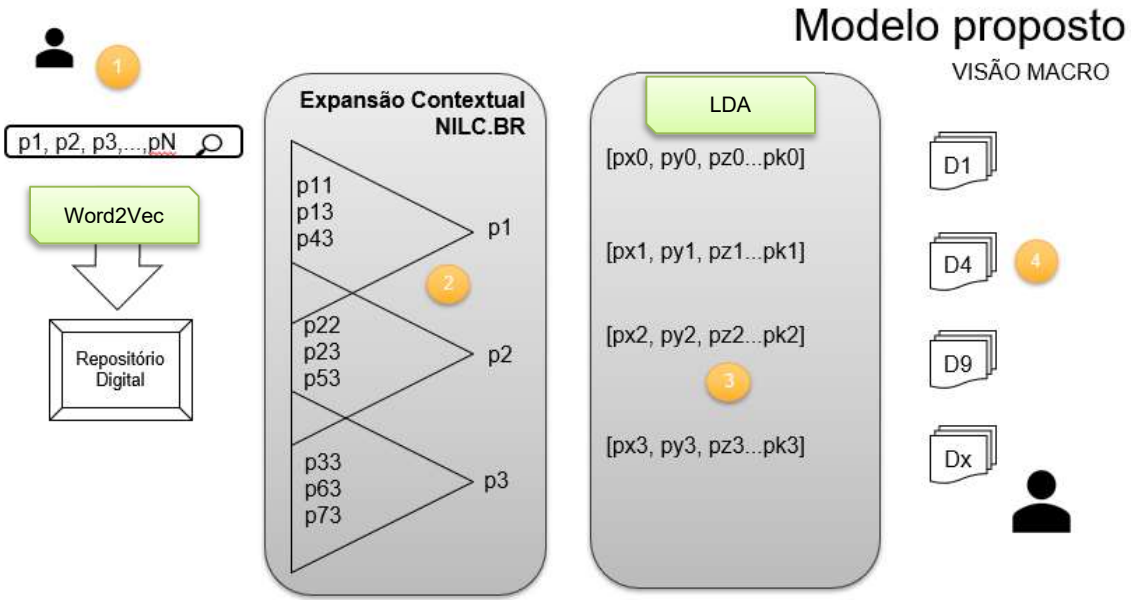
5 MODELO DE RECUPERAÇÃO DA INFORMAÇÃO COM ABORDAGEM SEMÂNTICA

Nesta seção é apresentada a proposta do modelo para recuperação da informação com abordagem semântica baseado nas técnicas de Inteligência Artificial discutidas, tais como: *topic modeling*, incorporação de palavras, o algoritmo LDA, o algoritmo Word2Vec e como estes interagem dentro do modelo. Ressalta-se a importância desta seção por ser a estrutura que permitirá o retorno de documentos de forma semântica, seguindo o processo interno e funcional da Recuperação da Informação de enriquecimento semântico localizado anteriormente à interação do usuário.

5.1 VISÃO GERAL DO MODELO

O modelo tem como principal objetivo habilitar uma estrutura sistêmica que permita, de forma autônoma, um ambiente no qual o usuário possa proceder sua busca e que esta estrutura proporcione o retorno de documentos com relevância contextual semântica sobre um determinado corpus. Nesse sentido, a figura 20 permite analisar, esquematicamente, este objetivo através de suas fases. A **fase 1** é composta pela entrada das palavras-chave.

Figura 20 – Esquema geral do Modelo Proposto



Fonte: Elaborado pelo autor (2021)

Estas palavras-chave vão ser a expressão de busca do usuário que então é ampliada usando o conceito de expansão contextual, definido no capítulo 3, que permite à cada palavra-chave da expressão de busca ser aumentadas através de similaridade contextual (seção 4.9) usando o algoritmo Word2Vec, finalizando a **fase 2**.

Desta maneira, as palavras digitadas pelo usuário serão suplementadas por outras palavras similares contextualmente obtidas pelo uso do algoritmo Word2Vec, e como corroborado pela hipótese distributiva de Harris (1954), que palavras de contextos semelhantes tendem a ter significados semelhantes, as palavras retornadas da aplicação do algoritmo, representadas na fase 2, por exemplo, por (p11, p12, p23) formam um conjunto de novas palavras que podem então serem utilizadas para maior capacidade de recuperação de documentos pertinentes contextualmente para o usuário.

A fase 3 faz a análise do corpus em questão aplicando-se o algoritmo LDA (seção 4.4), o qual possibilita que, de um determinado documento, possa ser extraído tópicos que representam aquele documento. Estes tópicos são representados por um conjunto de palavras as quais serão confrontadas pelas palavras das fases 1 e 2, ou seja, as palavras do usuário foram suplementadas contextualmente e então comparadas com as palavras dos tópicos do LDA encontradas na fase 3, aumentando-se assim a capacidade de “match” de palavras contextuais.

A fase 4 corresponde ao retorno dos documentos encontrados levando-se em conta as fases anteriores, os quais são retornados ao usuário agora usando-se as capacidades contextuais adquiridas e permitindo, desta maneira, maior capacidade de satisfazer as necessidades informacionais do usuário.

5.2 MODELO DE RECUPERAÇÃO DA INFORMAÇÃO SEMÂNTICO

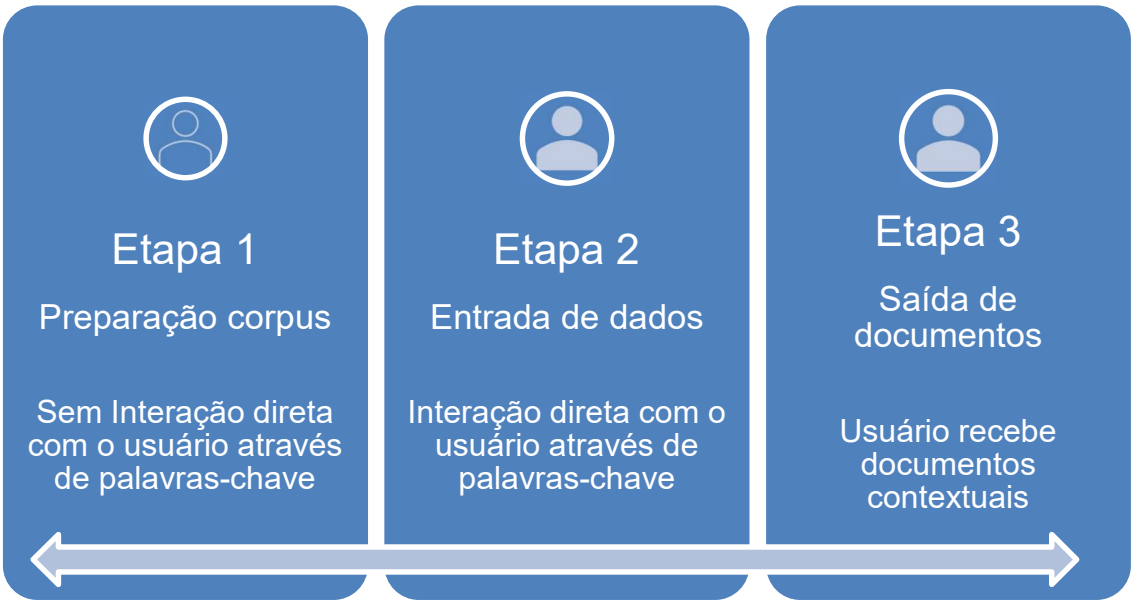
Na seção anterior, foi descrito de forma mais abstrata como o modelo funciona, no entanto, nesta seção será apresentado o modelo, de fato, suas fases bem como a relação entre seus módulos. O modelo está estruturado sob conceitos do processamento de linguagem natural, expansão contextual e uso de algoritmos como Word2Vec e LDA para recuperação da informação contextual para o usuário.

O modelo possui três etapas base, como pode ser visto na figura 21 onde em duas

delas a interação com o usuário é direta, a primeira é a etapa onde não há interação do usuário de forma direta e pode ser entendida como uma fase de preparação e análise do repositório ao qual o modelo será aplicado. Nessa fase o processo consiste em extrair o conteúdo dos documentos, ou seja, ler todos os documentos e montar um dicionário com as palavras extraídas de cada um deles.

A segunda etapa contempla a interação do usuário ao capturar as palavras-chave digitadas no campo de busca do repositório, estas palavras capturadas serão usadas dentro do módulo 6 que trata da expansão contextual dos termos digitados, o qual utiliza o algoritmo Word2Vec para descobrir palavras semanticamente relacionadas pelos conceitos de hipótese distributiva.

Figura 21 – Etapas do Modelo em relação à sua interação com o usuário



Fonte: Elaborado pelo autor (2021)

A figura 22 mostra a relação do modelo com o usuário, pois sabe-se que o usuário é parte fundamental do processo de recuperação da informação, e uma maior discussão sobre ele faz-se necessária. O objetivo do modelo é recuperar documentos com valor contextual para o usuário final, no entanto, como definido no objetivo não é foco desta tese a avaliação do modelo em relação à porcentagem de melhoria ou não do processo de RI usando o modelo proposto.

Nesse sentido, são apresentados na figura 22, os módulos componentes do modelo e como eles interagem a fim de fazer uma recuperação da informação contextual para o usuário. O **módulo 1** é utilizado para extrair e normalizar cada documento, pois os formatos dos arquivos podem ser variados ou até mesmo não ser um formato texto, o que desabilita as etapas seguintes de processamento de linguagem natural aplicada ao corpus presente no repositório.

Uma vez extraídos os documentos e transformados em formato ".TXT", estes poderão então ser analisados pelos próximos módulos, a começar pelo **módulo 2**, o qual é responsável pelo carregamento do arquivo e importação das bibliotecas utilizadas no processo de análise dos documentos.

No **módulo 3**, as palavras extraídas são compiladas em um *dataset* (um conjunto de dados, neste caso as palavras dos documentos analisados). Este *dataset* passa pela execução da fase de *Exploratory Data Analysis* (EDA), esta fase pode ser entendida como um processo de análise exploratória dos dados, ou seja, como não há um conhecimento total dos dados se faz necessário um reconhecimento do Corpus com algumas análises preliminares, além de normalizar o texto para preparação para as fases seguintes.

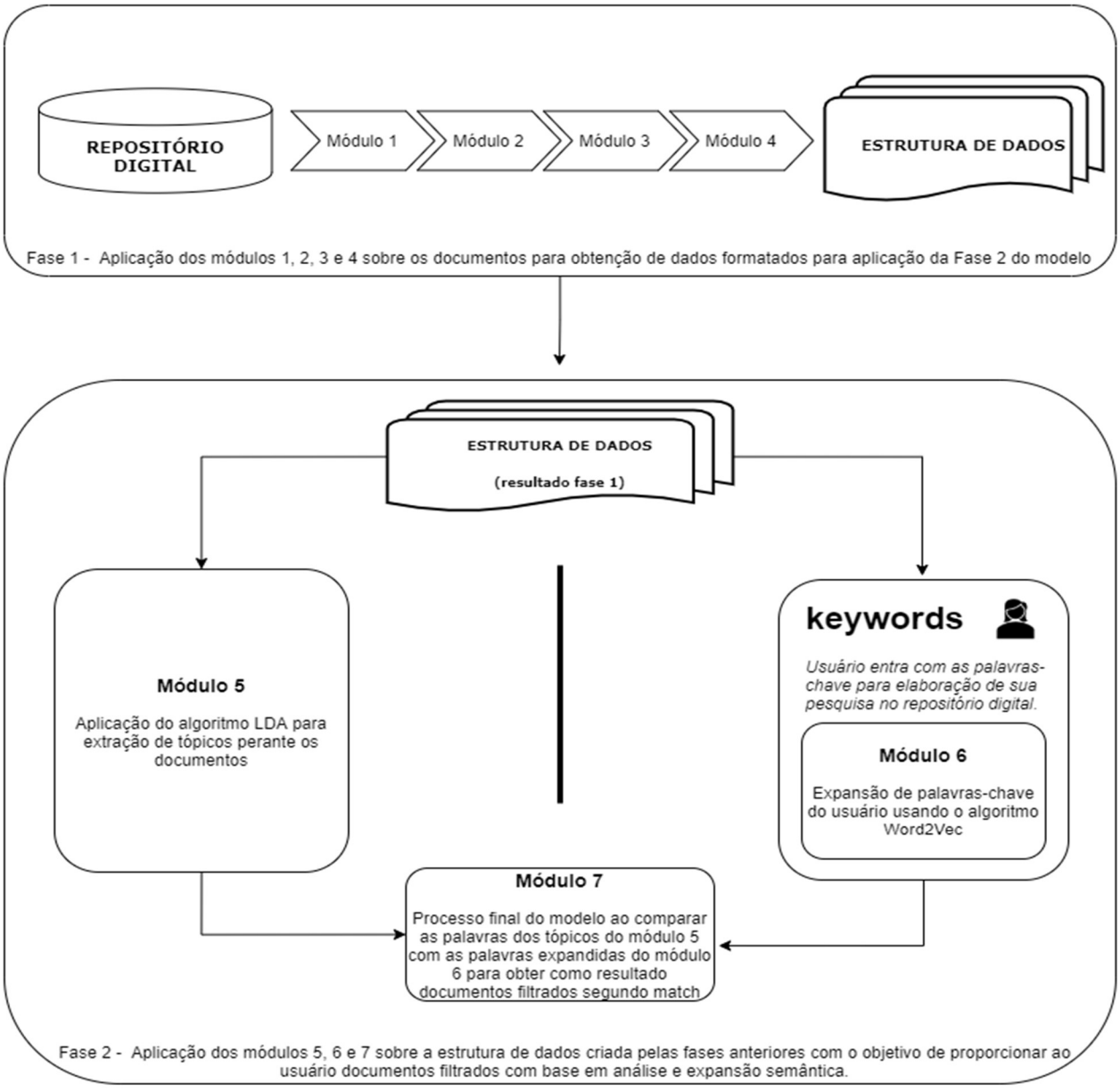
A normalização consta de procedimentos fundamentais do Processamento de Linguagem Natural como: transformação das palavras para minúsculo (*lowercase*) de todos os documentos, exclusão de pontuação e dois outros processos, que são a remoção de *stop words* (palavras que possuem alto grau de repetição no texto, como "a", "o", "que", "de" e não contribuem para um estado informacional sobre seu contexto) e lematização (processo de converter as palavras para sua inflexão no infinitivo dos verbos e masculino singular dos substantivos e adjetivos).

No **módulo 4** tem início a preparação para a aplicação do algoritmo LDA (*Latent Dirichlet Allocation*) que é um modelo que permite encontrar um conjunto de tópicos, ou seja, um conjunto de palavras mais recorrentes a serem descobertos (latentes) pela leitura de cada documento. Porém, dentro deste módulo faz-se necessário um pré-processo chamado de *Topic Modeling*, ou seja, transformar um conjunto de documentos em uma matriz de termos e documentos. A matriz termo documento é essencial para

relacionar cada palavra dos documentos e cada documento do *corpus* para, desta forma, permitir a algoritmos, como o LDA, que possam fazer inferências sobre co-ocorrências de determinadas palavras dentro de uma sentença.

Neste sentido, o LDA atua no modelo proposto para poder fazer a análise de quais os tópicos estão latentes dentro de cada documento do corpus, e desta forma permite através desta redução de dimensionalidade, que milhares de palavras sejam representadas por seus tópicos, algumas dezenas, o que permite melhor compreensão.

Figura 22 - Modelo proposto para recuperação da informação com abordagem semântica



Fonte: Elaborado pelo autor (2021)

A primeira fase termina no **módulo 4**, que produz como resultado uma estrutura de dados especializada para que a aplicação do algoritmo LDA seja executada. Esta execução explicita o **módulo 5**, que faz parte da segunda fase do modelo e tem por objetivo obter uma nova estrutura de dados (tabela 2 – página 93) que servirá de base para comparação das palavras contidas nesta estrutura.

O **módulo 6** é responsável pela expansão contextual semântica das palavras digitadas pelo usuário, por expansão contextual semântica entende-se da necessidade de se utilizar vetores de palavras pré-treinadas, que são estruturas que representam cada palavra matematicamente em um espaço vetorial e desta forma se obter palavras semanticamente vinculadas.

Para aplicação no modelo, foi utilizado o algoritmo Word2Vec que fora treinado usando os vetores de palavras do Núcleo Interinstitucional de Linguística Computacional da USP (NILC.BR), que é um repositório destinado ao armazenamento e compartilhamento de vetores de palavras em língua portuguesa.

Ainda no **módulo 6**, utilizando-se da semelhança de cosseno, explicado na seção 5.7, permite-se encontrar palavras que são semanticamente relacionadas às que foram digitadas pelo usuário e, portanto, a saída do **módulo 6** são as palavras semanticamente relacionadas pela aplicação do algoritmo Word2Vec já treinado.

Finalmente, na terceira fase, o **módulo 7** tem como objetivo fazer o *match* das palavras de saída do **módulo 6** que são palavras semelhantes semanticamente às palavras digitadas pelo usuário com as palavras contidas nos tópicos gerados pelo **módulo 5**. Estas palavras são confrontadas às palavras dos tópicos latentes do LDA, ou seja, para cada palavra pertencente a um determinado tópico faz-se a comparação a fim de se obter palavras semelhantes. Por fim, neste módulo são selecionados, portanto, os documentos que possuem em seus tópicos palavras que são coincidentes com as palavras expandidas pelo algoritmo word2vec do **módulo 6** e retornados ao usuário documentos mais refinados pelo modelo.

Na seção seguinte será demonstrada a prova de conceito usando os conceitos, técnicas e tecnologias que fazem dão sustentação ao modelo proposto na prática para comprovação do mesmo.

5.3 PROVA DE CONCEITO

Nesta seção são apresentados os resultados sobre como o modelo se comporta em uma aplicação real de seu uso. Para isso, apresenta-se a prova de conceito passo a passo, com cada um dos seus scripts e resultados de cada seção

O modelo, em um primeiro momento, tem como entrada uma massa de dados que

no caso são teses do repositório UNESP. Para a aplicação da prova de conceito foi feita uma extração de 50 teses de dois campi diferentes, porém, de mesma área. As áreas em questão são dois programas de pós-graduação na área de biotecnologia, o primeiro é o IQ – Instituto de Química - campus de Araraquara e outro é o programa de pós graduação IBB – Instituto de Biociência do campus de Botucatu. Além destes filtros foi também aplicado o filtro para obtenção de somente teses de doutorado.

Ressalta-se que a obtenção das teses de mesma área tem o objetivo de propiciar um estudo mais amplo em relação ao conteúdo processado pelo modelo desta tese, pois mesmo sendo campis de mesma área e mesma universidade, poder-se-á analisar ainda se os conteúdos produzidos nestes programas são unificados, se possuem correlação.

Destas teses extraídas, que são basicamente arquivos em PDF, foi aplicado um script que permite que possa ser realizado um processamento sobre cada arquivo e de forma automática, ou seja, depois da execução deste script ocorre a transformação dos arquivos de PDF para TXT, que é o formato base para que o script da segunda fase do modelo possa processar.

Uma consideração a ser feita é a de que, embora fora extraído uma quantidade bem menor de artigos do repositório o processamento e aplicação do modelo se comporta da mesma forma, e este limite foi imposto pela capacidade de processamento de máquinas locais (PCs) e não a utilização de máquinas em nuvem, que possibilitam maior capacidade de processamento.

Uma vez que os arquivos estejam transformados inicia-se a segunda fase, que é a responsável por, de fato, realizar a recuperação da informação de forma semântica, ou seja, retornar conteúdo com relevância contextual para o usuário, para isso, inicia-se o carregamento do corpus em si, fase 1 do modelo apresentado.

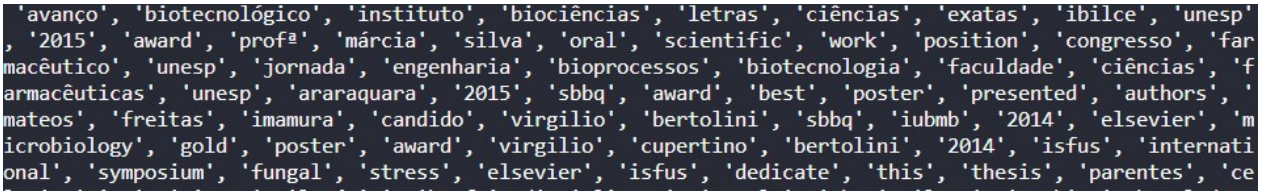
Figura 23 – Carregamento do Corpus

```
import re, nltk, spacy, string

import pandas as pd
import numpy as np


from sklearn.decomposition import LatentDirichletAllocation
from sklearn.feature_extraction.text import CountVectorizer, TfidfVectorizer
from pprint import pprint

import pyLDAvis
import pyLDAvis.sklearn
import matplotlib.pyplot as plt
%matplotlib inline

from plotly.offline import plot
import plotly.graph_objects as go
import plotly.express as px

df = pd.read_csv('/content/drive/My Drive/Colab Notebooks/AUTO2.csv')
```

Fonte: Elaborado pelo autor (2021)

No passo seguinte é apresentada uma amostra do arquivo carregado, como pode ser visto na figura 24, que apresenta a impressão dos artigos em um formato propicio para utilização do script. Durante a execução do script, vários passos ligados ao processamento de linguagem natural (PLN) usam várias bibliotecas (arquivos com comandos que podem ser reutilizados pelos programadores) as quais são instaladas ao longo do desenvolvimento.

Figura 24 – Impressão das primeiras linhas do corpus



```
 'avanço', 'biotecnológico', 'instituto', 'biociências', 'letras', 'ciências', 'exatas', 'ibilce', 'unesp'
, '2015', 'award', 'profª', 'márcia', 'silva', 'oral', 'scientific', 'work', 'position', 'congresso', 'far
macêutico', 'unesp', 'jornada', 'engenharia', 'bioprocessos', 'biotecnologia', 'faculdade', 'ciências', 'f
armacêuticas', 'unesp', 'araraquara', '2015', 'sbbq', 'award', 'best', 'poster', 'presented', 'authors', '
mateos', 'freitas', 'imamura', 'candido', 'virgilio', 'bertolini', 'sbbq', 'iubmb', '2014', 'elsevier', 'm
icrobiology', 'gold', 'poster', 'award', 'virgilio', 'cupertino', 'bertolini', '2014', 'isfus', 'internati
onal', 'symposium', 'fungal', 'stress', 'elsevier', 'isfus', 'dedicate', 'this', 'thesis', 'parentes', 'ce
```

Fonte: Elaborado pelo autor (2021)

5.3.1 Processamento de linguagem natural aplicado

O processamento de linguagem natural é parte essencial ao se analisar textos, essencialmente, são realizados vários processos, como definido na seção 4.6.1, entre eles, o processo chamado de *lowercase*, que é a mudança de cada palavra para minúscula e posteriormente, faz-se a extração das pontuações e remoção das *stopwords*, ou seja, palavras que não contribuem para uma identificação de um artigo e, por fim, foi

aplicado o processo de lematização.

Esta fase do modelo, ou seja, o processamento de linguagem natural, é um processo que demanda alto poder computacional e que normalmente exige-se uma máquina em nuvem. Para a execução do script foi utilizada uma instância do *Google* Colab[2] com aproximadamente 12 GB de memória RAM e 107 GB de disco e somente este processo levou cerca de 12 minutos.

5.3.2 Aplicação do algoritmo LDA no modelo e estruturas de dados

O objetivo desta etapa é atingir uma estrutura de dados como uma matriz a qual mantém em suas linhas os tópicos gerados pelo LDA e em suas colunas as palavras que formam este tópico. Para tal, aplicam-se comandos fornecidos pela biblioteca *Scikit-Learn*[3] que permite aplicação do algoritmo LDA que, por sua vez, aplica o aprendizado de máquina para gerar automaticamente a estrutura conhecida como matriz de tópicos e palavras (PEDREGOSA *et al*., 2011).

Esta matriz de tópicos e palavras, ou *Topic-Word Matrix*, na aplicação da prova de conceitos gerou dez tópicos gerados pelo LDA e para cada tópico suas palavras que o representam, como poder ser visto na figura 25. Esta matriz é importante, pois atua como base para que o modelo possa analisar quais as palavras mais frequentes dentro de cada tópico e por conseguinte, proporcionar estrutura de dados base para comparação com cada termo de busca do usuário.

Figura 25 - Matriz tópico-palavra, estrutura base para relação com termos de busca

[2] https://colab.research.google.com/
[3] https://scikit-learn.org/

df Topic keywords

	W0	W1	W2	...	W47
T0	glycogen	that	genes	...	human
T1	this	binding	nidulans	...	further
T2	were	expression	metabolism	...	detected
T3	accumulation	stress	promoter	...	higher
T4	crassa	clock	glucose	...	formation
T5	strain	factor	carbon	...	8807
T6	gene	signaling	promoters	...	organisms
T7	protein	regulation	wild	...	transduction
T8	transcription	with	chapter	...	present
T9	from	circadian	cell	...	extracts

Fonte: Elaborado pelo autor (2021)

Uma vez que a matriz tópico-palavra esteja definida, pode-se seguir para o passo seguinte, que é permitir através de outra matriz, a de Tópico por Documento (ver figura 26), fazer a análise de quais tópicos estão presentes dentro de cada documento. Desta forma, cria-se uma estrutura de dados na qual pode-se atingir os documentos aos quais as palavras de um determinado tópico tenham relação com os termos de busca do usuário.

Vale ressaltar que, para a etapa de *match* com os termos de busca e as palavras de cada tópico, esta ação não ocorre diretamente e estritamente da pesquisa do usuário, mas sim pela expansão semântica pela aplicação do Word2Vec treinado pelo modelo NILC.BR.

Tabela 2 - Matriz de tópico por documento e porcentagem dos tópicos que mais se destacam

DOCUMENTOS (TESES)	*TÓPICOS(PALAVRAS)*
Documento 1	**T1**(p1, p2, p3...), **T2**(p11, p22, p33...), **T3**(p111, p222, p333...)
Documento 2	**T1**(p1, p2, p3...), **T2**(p11, p22, p33...), **T3**(p111, p222, p333...)
...	
Documento N	**T1**(p1, p2, p3...), **T2**(p11, p22, p33...), **T3**(p111, p222, p333...)

Fonte: Elaborado pelo autor (2021)

No passo seguinte, como pode ser visto pela figura 26, são analisados os tópicos mais relevantes em cada documento, este passo permite analisar quais as palavras que têm maior capacidade de representação do documento, dentro daquele determinado tópico. Esta análise permite que os termos de busca do usuário possam ser confrontados

com as palavras deste tópico e não para o documento todo, e então possibilitar a recuperação de documentos com maior assertividade contextual.

Figura 26 - ordenação de tópicos por documento do corpus

```
# Get dominant topic for each document
dominant_topic = np.argmax(df_document_topic.values, axis=1)
df_document_topic['dominant_topic'] = dominant_topic
df_document_topic['dominant_topic']
```

```
Doc0      5
Doc1      5
Doc2      4
Doc3      9
Doc4     11
         ..
Doc84    11
Doc85    10
Doc86    10
Doc87    10
Doc88     0
Name: dominant_topic, Length: 89, dtype: int64
```

Fonte: Elaborado pelo autor (2021)

Por fim, ou seja, posterior à geração das matrizes, tópico por palavras e tópicos por documentos, e a aplicação do algoritmo LDA sobre estas, finaliza-se a etapa de extração das palavras mais usadas em cada documento, ou seja, a etapa fornece ao modelo a matéria-prima para análise frente aos termos de busca do usuário.

Estes, portanto, serão objetivo da próxima fase do modelo propostos, ao qual serão submetidas a uma expansão semântica pelo uso do algoritmo Word2Vec como apresentado na seção seguinte.

5.3.3 Aplicação do algoritmo Word2Vec e similaridade semântica

Como passo final da aplicação do modelo é utilizado o conceito de vetores de palavras representado pelo algoritmo Word2Vec explicado na seção 4.5. Para utilização

do algoritmo Word2Vec faz-se necessário informar um modelo pré-treinado de vetores de palavras. Para esta prova de conceito foi utilizado o repositório do Núcleo Interinstitucional de Linguística Computacional da USP (NILC), que é um repositório destinado ao armazenamento e compartilhamento de vetores de palavras em língua portuguesa.

O repositório possui um conjunto de vetores pré-treinados a partir de um grande corpus para o português do Brasil e de Portugal baseado em fontes e gêneros variados, totalizando 1.395.926.282 *tokens* (palavras ou símbolos distintos). O objetivo do repositório é fomentar e tornar acessível recursos vetoriais prontos para Processamento de Linguagem Natural e Aprendizado de Máquina no idioma português.

O corpus utilizado no treinamento destes modelos pré-treinados de vetores é bem vasto e variado e pode ser visto na figura 28, a qual descreve um quadro com as bases nas quais os vetores foram gerados. Estas bases estão abalizadas em revistas científicas como FAPESP, livros de literatura brasileira, além de uma amostra de toda a *Wikipedia* extraída em 2016.

Além de proporcionar os vetores treinados para o algoritmo Word2Vec, o NILC.BR também oferece suporte para outros algoritmos com a mesma finalidade do Word2Vec como *FastText* do *Facebook*[4], *Glove* da Universidade de Stanford[5] e Wang2Vec mantido pela comunidade sob o endereço[6].

Figura 27 - Corpus base informacional do NILC-USP

[4] https://fasttext.cc/
[5] https://nlp.stanford.edu/projects/glove/
[6] https://github.com/wlin12/wang2vec

Corpus	Tokens	Types	Genre	Description
LX-Corpus [Rodrigues et al. 2016]	714,286,638	2,605,393	Mixed genres	A huge collection of texts from 19 sources. Most of them are written in European Portuguese.
Wikipedia	219,293,003	1,758,191	Encyclopedic	Wikipedia dump of 10/20/16
GoogleNews	160,396,456	664,320	Informative	News crawled from GoogleNews service
SubIMDB-PT	129,975,149	500,302	Spoken language	Subtitles crawled from IMDb website
G1	105,341,070	392,635	Informative	News crawled from G1 news portal between 2014 and 2015.
PLN-Br [Bruckschen et al. 2008]	31,196,395	259,762	Informative	Large corpus of the PLN-BR Project with texts sampled from 1994 to 2005. It was also used by [Hartmann 2016] to train word embeddings models
Literacy works of public domain	23,750,521	381,697	Prose	A collection of 138,268 literary works from the Domínio Público web-site
Lacio-web [Aluísio et al. 2003]	8,962,718	196,077	Mixed genres	Texts from various genres, e.g., literary and its subdivisions (prose, po-etry and drama), informative, scientific, law, didactic technical
Portuguese e-books	1,299,008	66,706	Prose	Collection of classical fiction books written in Brazilian Portuguese crawled from Literatura Brasileira website
Mundo Estranho	1,047,108	55,000	Informative	Texts crawled from Mundo Estranho magazine
CHC	941,032	36,522	Informative	Texts crawled from Ciência Hoje das Crianças (CHC) website
FAPESP	499,008	31,746	Science Communication	Brazilian science divulgation texts from Pesquisa FAPESP magazine
Textbooks	96,209	11,597	Didactic	Texts for children between 3rd and 7th-grade years of elementary school
Folhinha	73,575	9,207	Informative	News written for children, crawled in 2015 from Folhinha issue of Folha de São Paulo newspaper
NILC subcorpus	32,868	4,064	Informative	Texts written for children of 3rd and 4th-years of elementary school
Para Seu Filho Ler	21,224	3,942	Informative	News written for children, from Zero Hora newspaper
SARESP	13,308	3,293	Didactic	Text questions of Mathematics, Human Sciences, Nature Sciences and essay writing to evaluate students
Total	1,395,926,282	3,827,725		

Fonte: (NÚCLEO INTERINSTITUCIONAL DE LINGUÍSTICA COMPUTACIONAL, 2017)

O repositório NILC.BR ainda disponibiliza vetores de palavras em várias dimensões, ou seja, quanto maior a dimensão mais palavras são abordadas e pré-treinadas em vetores que fornecerão capacidade de inferência semântica. No entanto, quanto maior a dimensão maior também é a necessidade de poder computacional.

Além disso, o NILC.BR ainda suporta as variações *CBOW* e *Skip-Gram*, que como explicado na seção 4.5 são abordagens para contextualização semântica. O mecanismo CBOW determina como entrada algumas palavras de contexto e a partir destas o algoritmo infere uma determinada palavra alvo, já no *Skip-Gram* é o contrário, dada uma palavra alvo o algoritmo infere as “n” palavras contextuais à palavra alvo.

Nesse sentido, a opção pelo mecanismo *Skip-Gram* foi o escolhida para aplicação na prova de conceito do modelo, pois necessita-se descobrir palavras contextuais dada uma determinada palavra do termo de busca do usuário. Portanto, foi realizado o download do modelo pré-treinado chamado “skip_s100.txt”, ou seja, modelo em *skip-gram* já que, para cada palavra-chave digitada pelo usuário, deseja-se expandi-la para

outras palavras semanticamente pertinentes sob o treinamento do modelo do NILC.

Posteriormente a esta fase, carrega-se este vetor de palavras junto ao algoritmo Word2Vec para que se possa utilizar o conceito de similaridade semântica, definido na seção 4.8 para então se obter a expansão contextual com o objetivo de melhorar a recuperação da informação baseada em palavras-chave expandidas semanticamente.

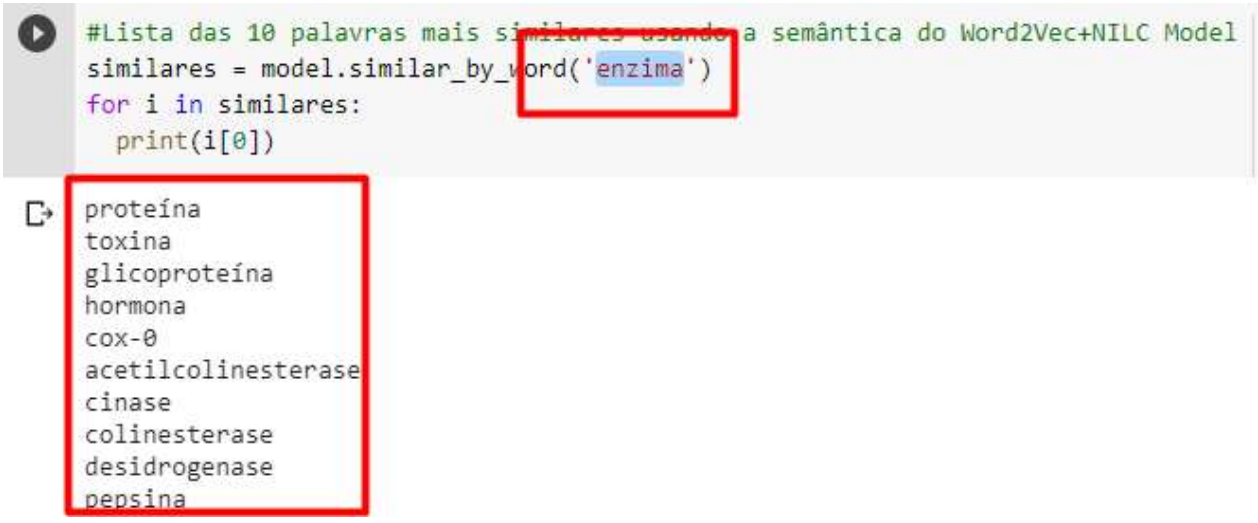
Além do carregamento do modelo pré-treinado, necessita-se de uma outra biblioteca conhecida como Gensim, implementada em linguagem Python, e de código aberto, esta biblioteca carrega o modelo pré-treinado do NILC e possibilita análise de similaridade.

Para exemplificar a expansão das palavras de forma contextual foi realizado, dentro da prova de conceito, uma simulação de busca de documentos usando a palavra “dpoc.”, esta palavra ou acrônimo significa **doença pulmonar obstrutiva crônica** e nesta busca o usuário poderia estar interessado na obtenção de informação sobre este termo.

No entanto, por ser um termo ligado à medicina pode ocorrer que seja necessário conhecer outros termos específicos tão importantes relacionados a este para se obter um resultado satisfatório, e por desconhecimento por parte do usuário de tais termos o resultado poderia ser reduzido em relação aos documentos retornados.

Neste sentido, o modelo apresentado preenche esta lacuna de conhecimento ao se utilizar das técnicas e tecnologias propostas e utilizadas nesta prova de conceito. A figura 29 mostra que após a execução de comandos passando-se como parâmetro a palavra buscada entre aspas e parênteses, o retorno é dado com a expansão de mais dez palavras que possuem o mesmo contexto da palavra em questão.

Figura 28 - Similaridade usando Word2Vec e NILC.BR



Fonte: Elaborado pelo autor (2021)

Ressalta-se que este contexto não é dado por sinônimos ou baseado em uma ontologia ou algum outro instrumento do mesmo tipo, mas sim através da utilização do modelo espaço vetorial, no qual tais palavras estão dispostas.

Este modelo espaço vetorial, no qual tais palavras estão organizadas, foi construído a partir da análise das várias bases do repositório NILC.BR, o que permite uma inferência, usando-se das técnicas e tecnologias propostas no modelo, o retorno de palavras que se encontram em menores distâncias no modelo vetorial uma das outras (ver seção 4.7).

Destaca-se ainda que, este é o ponto no qual acontece de fato a **aplicação da semântica de busca**, ação que ocorre de forma automática, usando as estruturas e algoritmos do modelo proposto as quais permitem retornar palavras pertinentes ao contexto semântico das palavras utilizadas nos termos de busca do usuário.

5.3.4 Finalização da aplicação do algoritmo Word2Vec e uso da similaridade semântica para busca dos documentos

Até o momento duas grandes etapas do modelo foram cumpridas, a etapa para se encontrar os tópicos formadores de um documento usando o algoritmo LDA e a outra etapa, a da geração das palavras expandidas segundo Word2Vec e NILC.BR. Agora pretende-se juntar estas duas grandes etapas em uma única para simular de fato uma

pesquisa de um usuário.

Para isso, a figura 29 explicita como de fato esse processo ocorre, ou seja, como exemplo, toma-se que o usuário tenha digitado o termo “dpoc” e então o modelo, ao processar as variações semânticas desta palavra uma a uma, segundo aplicação do algoritmo word2vec, leva a análise de quais tópicos, ou ainda, quais palavras daquele determinado tópico, que por conseguinte pertence a um determinado documento, possuem relacionamento com a palavra digitada pelo usuário, este processo é iterado para cada palavra expandida como exemplo da figura 29.

Figura 29 – Identificação do Tópico com o termo de busca

```
#simulação de um termo de busca do usuário
#representação da mesma na variável "query"
#deve-se retornar todos os TÓPICOS que possuam vínculo com esta palavra
query = "dpoc"
```

```
['dpoc']
```

	Topic 0	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9
0	False	False	False	False	False	False	True	False	False	False

Fonte: Elaborado pelo autor (2021)

Desta maneira, pode-se recuperar quais documentos estão relacionados com os termos de busca do usuário, de tal forma que, para cada palavra digitada pelo usuário, esta será expandida para 10 outras semanticamente contextuais segundo utilização do modelo e assim prover uma recuperação da informação com capacidade semântica.

Para ilustrar este momento, a figura 30 representa o processo final, no qual, dado que o tópico 6 é o que mais identifica a palavra-chave do usuário, então são recuperados os documentos que estão relacionados ao tópico 6. Como pode-se ver pela figura seriam os documentos 0, o documento 7, o documento 19 e os documentos 26 e 42.

Figura 30 – Relacionamento de tópicos com documentos

	Topic 0	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9
Doc0	0.05	0.05	0.05	0.05	0.05	0.05	0.55	0.05	0.05	0.05
Doc7	0.05	0.05	0.05	0.05	0.05	0.05	0.55	0.05	0.05	0.05
Doc19	0.05	0.05	0.05	0.05	0.05	0.05	0.55	0.05	0.05	0.05
Doc26	0.05	0.05	0.05	0.05	0.05	0.05	0.55	0.05	0.05	0.05
Doc42	0.05	0.05	0.05	0.05	0.05	0.05	0.55	0.05	0.05	0.05

Fonte: Elaborado pelo autor (2021)

Disto posto, conclui-se que essa prova de conceito poderia retornar ao usuário que utilize um sistema baseado no modelo proposto os documentos que possuem maior relacionamento contextual semântico baseado no modelo proposto formado pelas estruturas e algoritmos. Ou seja, o modelo permite que documentos possam ser recuperados para o usuário levando-se em conta uma abordagem semântica que por trás está apoiada sobre inteligência artificial, aprendizado de máquina e processamento de linguagem natural.

No capítulo seguinte são feitas as considerações finais e possíveis trabalhos futuros gerados em decorrência desta tese. Tais trabalhos poderão ser explorados tanto nas áreas relacionadas aos processos tecnológicos para recuperação da informação quanto como parte final da avaliação da ferramenta junto a um grupo focal.

6 CONSIDERAÇÕES FINAIS

Nesta seção, serão discutidos os resultados do modelo proposto quando aplicado a um cenário real. Também serão discutidos o comportamento do mesmo perante os objetivos definidos, além dos trabalhos futuros identificados ao longo do desenvolvimento deste trabalho.

A Ciência da Informação (CI) por seu caráter intrínseco pautado na interdisciplinaridade com várias outras ciências, especialmente à ciência da computação (CC) forma um complexo e poderoso *framework* para ampliar, sobremaneira, a capacidade de alcance na solução de problemas pertinentes aos cenários dessas ciências, especialmente os processos da recuperação da informação.

Neste sentido, esta tese contribui com a Ciência da Informação ao aventar possibilidades de proporcionar uma melhoria na abordagem da Recuperação da Informação em ambientes digitais como os sistemas computacionais baseados em documentos textuais.

Foram reunidos técnicas e conceitos da linguística, da ciência da computação, e da ciência da informação sob o campo da recuperação da informação aplicados à repositórios digitais, para que, através de abordagens de inteligência artificial este trabalho propicie melhoria na forma de recuperação da informação contextual para o usuário e assim possibilitar ampliação das suas necessidades informacionais.

Baseado neste cenário, foi proposto, discutido e implementado um modelo de recuperação da informação com abordagem semântica que, ao reunir algoritmos como o LDA para modelagem de tópicos e Word2Vec para incorporação de palavras, foi possível aumentar a contextualização das palavras de busca usando os modelos treinados pelo Núcleo Interinstitucional de Linguística Computacional da USP (NILC.BR).

No campo social esta tese aspira ter contribuído, pois foi possível prover aos usuários, através de seu modelo e conjunto de conceitos, técnicas e tecnologias uma melhor recuperação da informação em sistemas computacionais digitais baseados em documentos textuais. Permitindo, desta maneira, acesso à informação de forma mais abrangente e contextual aos aparatos informacionais de armazenamento de

conhecimento, o que fornece à sociedade a faculdade de pesquisa, análise, reflexão e tomada de decisão sobre os mais variados domínios do conhecimento.

Em relação ao objetivo geral do trabalho, o qual está definido em ampliar a capacidade Recuperação da Informação nos sistemas computacionais baseados em documentos textuais, para que o usuário, em suas necessidades informacionais, possa expandir o acesso à informação contextual nestes ambientes, assinala-se que o objetivo foi alcançado ao fazer uso do modelo e sua prova de conceito em conseguir ampliar as capacidades informacionais do usuário usando-se das técnicas, tecnologias e conceitos empregados.

Sobre os objetivos específicos, o primeiro estava em analisar os processos de recuperação da informação de forma geral, já o segundo estava em avaliar os processos de recuperação no contexto da inteligência artificial e o terceiro deveria aprofundar sobre o conceito de incorporação de palavras e demais conceitos chave para a melhoria na recuperação da informação como: algoritmos, tecnologias e conceitos matemáticos envolvidos.

Estes três primeiros objetivos específicos foram atingidos a partir dos capítulos de recuperação da informação e expansão contextual, bem como com o capítulo que traz conceitos sobre inteligência artificial. Estas seções abarcam várias das técnicas concernentes às ciências da informação e da computação, integração essencial ao desenvolvimento de todo o trabalho.

Já os objetivos específicos seguintes que versam sobre desenvolver uma arquitetura que abarque *softwares* e bibliotecas pertinentes à área de Processamento de Linguagem Natural para comprovação da utilização na análise de documentos e apurar quais as técnicas de Processamento de Linguagem Natural no sentido de conhecer e avaliar os *frameworks* utilizados para processamento de linguagem natural estão descritos no capítulo sobre inteligência artificial e processamento de linguagem natural.

Neste capítulo as técnicas de processamento de linguagem natural estão, especificamente, descritas nas seções sobre considerações iniciais sobre processamento de linguagem natural, seus conceitos e aplicações e mais detalhadamente nas seções que abarcam as etapas de processamento de linguagem natural.

Já o objetivo específico que trata sobre detalhar de forma prática o algoritmo *Latent Dirichlet Allocation* (LDA) como analisador dos documentos para redução de dimensionalidade é tido como um conceito importante para capacidade de processamento do modelo e foi estabelecido na seção 4.4 que versa sobre o algoritmo em si, seus conceitos e fórmulas matemáticas que o compõe.

Por fim, o objetivo específico que corresponde ao modelo proposto e fala sobre utilizar o modelo construído através de uma prova de conceito que visa a melhoria da recuperação da informação em sistemas computacionais compostos por documentos digitais textuais através de um *script*, com o objetivo de avaliar a recuperação contextual da informação para o usuário está definido no capítulo específico sobre o modelo chamado de modelo de recuperação da informação com abordagem semântica.

O capítulo referido no parágrafo anterior, além de compreender a arquitetura do modelo também contempla a prova de conceito que começa em carregar as teses extraídas do repositório da Unesp e transformadas em formatos de arquivos que podem ser processados. Logo em seguida foi realizada uma análise sobre como este conjunto de arquivos se encontrava e então foi aplicado o processamento de linguagem natural aos documentos para transformá-los a uma forma na qual pudessem ser processados pelos algoritmos de inteligência artificial.

A prova de conceito foi realizada e seus resultados comprovam a tese levantada de que seria possível desenvolver um modelo de recuperação da informação com abordagem semântica, para que a resposta deste modelo pudesse proporcionar informação contextualizada para o usuário. O que corrobora a hipótese desta pesquisa que é a de que ao se utilizar o aprendizado de máquina para maior reconhecimento do grande volume de textos, e ao mesmo tempo, utilizar incorporação de palavras através de algoritmos de inteligência artificial para propiciar similaridade semântica.

Destaca-se, portanto, que a aproximação da ciência da computação com seu arcabouço técnico e tecnológico de fato contribui, sobremaneira, com ciência da informação, sobretudo em áreas como a recuperação da informação e este trabalho pôde comprovar esta integração através do seu modelo definido em proposta de pesquisa.

No entanto, ainda mais importante do que a integração em si, segue-se o fato de que a contribuição maior está na possibilidade de a sociedade utilizar-se deste trabalho

para maior compreensão de mundo através de inúmeras pesquisas realizadas em sistemas computacionais que abarcam o conhecimento armazenado sob a forma de arquivos digitais textuais e nem sempre possuem a capacidade sistêmica de retorno de informação armazenada a contento.

São registradas algumas limitações deste trabalho como o fato de não ter sido tratada uma maior quantidade de documentos, este fator se deu ao fato que há uma limitação de processamento de máquinas. Outra limitação do trabalho está no fato de não ter sido realizada uma prova de utilização humana do modelo proposto a fim de avaliar o modelo qualitativamente e quantitativamente.

Neste sentido, como trabalho futuro fica a proposição de um grupo focal para avaliação do modelo visando garantir que o mesmo possa ser utilizado em larga escala e possivelmente ser criado um módulo de extensão de software, um plugin, para ser adicionado aos sistemas computacionais que armazenam conhecimento sob a forma de arquivos textuais digitais.

REFERÊNCIAS

ABU-SALIH, B.; PORNPIT, W.; CHAN, K. Y. Twitter mining for ontology-based domain discovery incorporating machine learning. **Journal of Knowledge Management**, Kempston, v. 22, n. 5, p. 949-981, 2018. Disponível em: encurtador.com.br/glzEN. Acesso em: 14 fev. 2021.

AHMAD, A. *et al*. A survey on mining stack overflow: question and answering (Q&A) community. **Data Technologies and Applications**, Bingley, v. 52, n. 2, p. 190-247, 2018. Disponível em: encurtador.com.br/pYZ48. Acesso em: 14 fev. 2021.

ARUN, K. S.; GOVINDAN, V. K.; MADHU, K. S. D. Enhanced bag of visual words representations for content based image retrieval: a comparative study. **The Artificial Intelligence Review**, Dordrecht, v. 53, n. 3, p. 1615-1653, 2020. Disponível em: encurtador.com.br/otwV1. Acesso em: 14 fev. 2021.

ÁVILA, B. T.; SILVA, M.; CAVALCANTE, L. Uso de Repositórios Digitais como Fonte de Informação por Membros das Universidades Federais Brasileiras. **Informação & Sociedade: Estudos**, v. 27, n. 3, 24 dez. 2017. Disponível em: https://periodicos.ufpb.br/ojs2/index.php/ies/article/view/31514. Acesso em: 11 ago. 2020.

BAEZA-YATES, R.; RIBEIRO-NETO, B. **Modern Information Retrieval**. New York: ACM Press, 1999. 511p.

BENGIO, Y. *et al*. A Neural Probabilistic Language Model. **Journal of Machine Learning Research**, [*s.l.*], v. 3, n. 1, 2003. Disponível em: https://www.jmlr.org/papers/volume3/bengio03a/bengio03a.pdf. Acesso em: 11 ago. 2020.

BEPPLER, F. D. **Um modelo para recuperação e busca de informação baseado em ontologia e no círculo hermenêutico**. 2008. 123 f. Tese (Doutorado em Engenharia e Gestão do Conhecimento) – Universidade Federal de Santa Catarina, Florianópolis, 2008. Disponível em: https://repositorio.ufsc.br/xmlui/handle/123456789/90972. Acesso em: 10 fev. 2021.

BERNERS-LEE T.; LASSILA, O.; HENDLER, J. The semantic web. **Scientific American**, New York, v. 5, mai. 2001. Disponível em: https://www-sop.inria.fr/acacia/cours/essi2006/Scientific%20American_%20Feature%20Article_%20The%20Semantic%20Web_%20May%202001.pdf. Acesso em: 17 fev. 2021.

BLEI, D. M. **Introduction to Probabilistic Topic Modeling**. Communications of the ACM, [s.l], abr. 2012. Disponível em: https://dl.acm.org/doi/10.1145/2133806.2133826. Acesso em: 17 fev. 2021.

BLUMENTHAL, S. C. **Management information systems**. Englewood Cliffs: Prentice-Hall, 1969.

BRAGA, F. R. Extração semiautomática de taxonomia para domínios especializados usando técnicas de mineração de textos. **Ciência da Informação**, Brasília, v. 45, n. 3, 23 fev. 2018. Disponível em: http://revista.ibict.br/ciinf/article/view/4056/3577. Acesso em: 31 jul. 2020.

CABRERA DIEGO, L. A. **TF-IDF para la obtención automática de términos y su validación mediante Wikipedia**. 2011. Trabalho de Conclusão de Curso (Licenciatura en Ingeniería en Computación) **-** Facultad de Ingeniería, Universidad Nacional Autónoma de México, México, 2011. Disponível em: https://repositorio.unam.mx/contenidos/tf-idf-para-la-obtencion-automatica-de-terminos-y-su-validacion-mediante-wikipedia-3505934?c=yNz8pe&d=false&q=*:*&i=4&v=1&t=search_0&as=0. Acesso em: 31 jul. 2020.

CONEGLIAN, C. S. **Recuperação da informação com abordagem semântica utilizando linguagem natural:** a inteligência artificial na ciência da informação. 2020. 195 f. Tese (Doutorado em Ciência da Informação) - Faculdade de Filosofia e Ciências, Universidade Estadual Paulista, 2020. Disponível em: https://www.marilia.unesp.br/#!/pos-graduacao/mestrado-e-doutorado/ciencia-da-informacao/publicacoes-academicas/teses/ Acesso em: 01 fev. 2021.

CHEN, X. *et al*. **A bibliometric analysis of event detection in social media**. Information Review, Bradford, v. 43, n. 1, p. 29-52, 2019. Disponível em: encurtador.com.br/yAY16. Acesso em: 01 fev. 2021.

CRISTOVÃO, H. M.; DUQUE, C. G.; SERQUEIRA, L. D. Recuperação de informação: uma aplicação na criação e configuração automáticas de cursos virtuais a distância*. In:* ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 13., Rio de Janeiro, 2012. **Anais [...],** Rio de Janeiro, UFS, 2012. Disponível em: http://repositorios.questoesemrede.uff.br/repositorios/handle/123456789/2043. Acesso em: 01 fev. 2021.

FALEIROS, T. de P.; LOPES, A. de A**. Modelos Probabilísticos de Tópicos:** desvendando o Latent Dirichlet Allocation. São Paulo: USP, 2016.

ESPOSITO, M. *et al*. Hybrid query expansion using lexical resources and word embeddings for sentence retrieval in question answering. **Information Sciences**, [s. l.], v. 514, p. 88–105, 2020. Disponível em: https://www.x-mol.com/paper/1308253918814507008?recommendPaper=5778150. Acesso em: 01 fev. 2021.

EVERITT, B. S. **The Cambridge dictionary of statistics**. 3. ed. Cambridge: Cambridge University Press. 2006.

FACHIN, G. R. B. Recuperação inteligente da informação e ontologias: um levantamento na área da Ciência da Informação. **BIBLOS**, [*s.l.*], v. 23, n. 1, p. 259-283, mar. 2010. Disponível em: https://periodicos.furg.br/biblos/article/view/1282. Acesso em: 18 ago. 2020

FERNEDA, E. **Recuperação de informação**: análise sobre a contribuição da ciência da computação para a ciência da informação. 2003. Tese (Doutorado em Ciência da Informação e Documentação) – Escola de Comunicações e Artes, Universidade de São Paulo, São Paulo, 2003. Disponível em: https://www.teses.usp.br/teses/disponiveis/27/27143/tde-15032004-130230/pt-br.php. Acesso em: 18 ago. 2020.

FERNEDA, E. Redes neurais e sua aplicação em sistemas de recuperação de informação. **Ciência da Informação**, Brasília, v. 35, n. 1, 2006. Disponível em: https://www.scielo.br/pdf/ci/v35n1/v35n1a03.pdf. Acesso em: 23 jul. 2020.

FISCHER, A. E.; GRODZINSKY, F. S. **The Anatomy of Programming Languages**. Englewood Cliffs, New Jersey: Prentice Hall, 1993.

GIL, A. C. **Como elaborar projetos de pesquisa**. 4. ed. São Paulo: Atlas, 2002.

GRUBER, T. R. The Role of Common Ontology in Achieving Sharable, Reusable Knowledge Bases. *In***:** ALLEN, J. A.; FIKES, R.; SANDEWALLE. (Eds.). **Principles of Knowledge Representation and Reasoning**: Proceedings of the Second International Conference. San Mateo: Morgan Kaufmann, 1991.

HARRIS, Z. S**.** Distributional structure. **Word**, [*s.l.*], v. 10, p. 146–162, 1954. Disponível em: https://www.tandfonline.com/doi/pdf/10.1080/00437956.1954.11659520. Acesso em: 17 fev. 2021.

HUANG, J. **A tale of two paradigms**: Disambiguating extracted entities with applications to a digital library and the Web. Pennsylvania: University Ann Arbor, 2010.

HUTCHINS, J. The first public demonstration of machine translation: the Georgetown-IBM system. **Semantic Scholar**, 2004. Disponível em: https://www.semanticscholar.org/paper/The-first-public-demonstration-of-machine-%3A-the-%2C-7-Hutchins/ad014a7b7a3142e6f17eedddf4218489b56ab18e. Acesso em: 17 fev. 2021.

KOBASHI, N. Y. Análise documentária e representação da informação. **Informare**, [*s.l.*], v. 2, n. 2, 1996. Disponível em: http://hdl.handle.net/20.500.11959/brapci/40976. Acesso em: 22 jul. 2020.

LAMBA, M.; MADHUSUDHAN, M. Mapping of etds in proquest dissertations and theses (pqdt) global database (2014-2018). **Cadernos BAD,** Portugal, n. 1, p. 169-182, 2019. Disponível em: http://hdl.handle.net/20.500.11959/brapci/134567. Acesso em: 31 jul. 2020.

LANE H.; HOWARD, C.; HAPKE H. Natural Language Processing in Action. **Manning Publications,** 2019. Disponível em: https://www.manning.com/books/natural-language-processing-in-action#reviews. Acesso em: 22 jul. 2020.

MARON, M.; KUHNS, J. L. **Na relevância, na indexação probabilística e na recuperação da informação**. Journal of the ACM, v. 7, n. 3, 1960.

MARTINS, D. L. *et al*. Repositório digital com o software livre Tainacan: revisão da ferramenta e exemplo de implantação na área cultural com a revista filme cultura. ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 18., São Paulo. **Anais [...]**, São Paulo: UNESP, 2017. Disponível em: https://pesquisa.tainacan.org/wp-content/uploads/2019/01/repositorio-digital.pdf. Acesso em: 04 ago. 2020.

MARTINS, D.; SILVA, M.; SIQUEIRA, J. Comparação entre sistemas para criação de acervos digitais: análise dos softwares livres DSpace, EPrints, Fedora, Greenstone e Islandora a partir de novas dimensões analíticas. **InCID**, São Paulo, v. 9, n. 1, p. 52-71, 1 jun. 2018. Disponível em: http://www.revistas.usp.br/incid/article/view/134333. Acesso em: 04 ago. 2020.

MIKOLOV, T. *et al*. Eficient estimation of word representations in vector space. **Cornell University**, 2003. Disponível em: https://arxiv.org/abs/1301.3781. Acesso em: 04 ago. 2020.

MITCHELL, T. M. Machine Learning. **McGraw-Hill Science**, 1997. Disponível em: http://profsite.um.ac.ir/~monsefi/machine-learning/pdf/Machine-Learning-Tom-Mitchell.pdf. Acesso em: 16 jul. 2020.

MITRA, B.; CRASWELL, N. Neural Models for Information Retrieval. **Cornell University**, 2017. Disponível em: http://arxiv.org/abs/1705.01509. Acesso em: 4 ago. 2020.

MOOERS, C. Zatocoding applied to mechanical organization of knowledge. **American Documentation**, [*s.l.*], v. 2, n. 1, 1951.

MULLER, O. *et al*. Utilizing big data analytics for information systems research: challenges, promises and guidelines. **European Journal of Information Systems**, Abingdon, v. 25, n. 4, p. 289-302, 2016. Disponível em: https://www.tandfonline.com/doi/full/10.1057/ejis.2016.2. Acesso em: 14 fev. 2021.

NAVARRO, F. P.; CONEGLIAN, C. S.; SANTARÉM SEGUNDO, J. E. Big data no contexto de dados acadêmicos: o uso de machine learning na construção de sistema de organização do conhecimento. ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 19., Londrina. **Anais [...]**, Londrina: UEL, 2018. Disponível em: http://hdl.handle.net/20.500.11959/brapci/103396. Acesso em: 17 jul. 2020.

NÚCLEO INTERINSTITUCIONAL DE LINGUÍSTICA COMPUTACIONAL. **Repositório de Word Embeddings do NILC.** Disponível em: http://nilc.icmc.usp.br/nilc/index.php/repositorio-de-word-embeddings-do-nilc#. Acesso em: 17 fev. 2021.

PARK, J. D. **Automated question triage for social reference**: A study of adopting decision factors from digital reference. Pittsburgh: University of Pittsburgh, 2013.

PINHEIRO, L.; LOUREIRO, J. Traçados e limites da ciência da informação. **Ciência da informação**, Brasília, v. 24, n. 1, jan./abr. 1995.

RAMACHANDRAN, D.; RAMASUBRAMANIAN, P. Event detection from Twitter – a survey. **International Journal of Web Information Systems**, Bingley, v. 14, n. 3, p. 262-280, 2018.

RUAS, T. *et al*. Enhanced word embeddings using multi-semantic representation through lexical chains. **Information Sciences**, [s. l.], v. 532, p. 16-32, 2020. Disponível em: https://arxiv.org/abs/2101.09023. Acesso em: 12 fev. 2021.

RUSSELL, S. J.; NORVIG, P. **Artificial intelligence: a modern approach**. Malaysia: Pearson Education Limited, 2016.

SAAR K.; SHTOK, A.; KURLAND, O. **Query Expansion Using Word Embeddings**. In: CONFERENCE ON INFORMATION AND KNOWLEDGE MANAGEMENT, 16., 2016. New York: Association for Computing Machinery, 2016.

SANCHEZ, F. A.; VECHIATO, F. L.; VIDOTTI, S. A. B. G. Encontrabilidade da Informação em Repositórios de Dados: uma análise do DataONE. **Informação & Informação**, [*s.l.*], v. 23, p. 51-79, 2019. Disponível em: http://www.uel.br/revistas/uel/index.php/informacao/article/view/30725. Acesso em: 01 fev. 2021.

SANGIORGI, O. Redes neurais naturais, redes neurais artificiais e habilidades de aprendizagem: sob o ponto de vista cibernético. **Cajueiro**, [*s.l.*], v. 1, n. 2, p. 181-196, 2019. Disponível em: http://hdl.handle.net/20.500.11959/brapci/125504. Acesso em: 01 fev. 2021.

SANT'ANA, R. C. G. Ciclo de vida dos dados e o papel da Ciência da Informação. *In:* ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 14., 2013, Florianópolis. **Anais [...],** Florianópolis:UFSC, 2013. Disponível em: http://enancib.sites.ufsc.br/index.php/enancib2013/XIVenancib/paper/viewFile/284/319. Acesso em: 23 set. 2017.

SANTAREM SEGUNDO, J. E. **Representação Iterativa**: um modelo para Repositórios Digitais. 2010. 244f. Tese (Doutorado em Ciência da Informação) – Faculdade de

Filosofia e Ciências, Universidade Estadual Paulista, Marília, 2010. Disponível em: https://repositorio.unesp.br/handle/11449/103346. Acesso em: 11 fev. 2021.

SANTOS, P. L. V. A. da C.; SANT'ANA, R. C. G. Dado e Granularidade na perspectiva da Informação e Tecnologia: uma interpretação pela Ciência da Informação. **Ciência da Informação**, Brasília, v. 42, n. 2, 27 jan. 2015. Disponível em: http://revista.ibict.br/ciinf/article/view/1382/1560. Acesso em: 11 ago. 2020.

SANTOS, R. F. D.; NEVES, D. A. B. **Práticas de indexação em repositórios digitais de acesso aberto**: análise do metadado assunto do repositório institucional da UFRN. *In*: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 19., 2018, Londrina. **Anais [...],** Londrina: UEL, 2018. Disponível em: http://hdl.handle.net/20.500.11959/brapci/102563. Acesso em: 04 ago. 2020.

SARACEVIC, T. Ciência da informação: origem, evolução e relações. **Perspectivas em Ciência da Informação**, Belo Horizonte, v. 1, n. 1, p. 41-62, jan./jun. 1996. Disponível em: http://portaldeperiodicos.eci.ufmg.br/index.php/pci/article/view/235. Acesso em: 05 ago. 2018.

SARACEVIC, T. Interdisciplinary nature of information science. **Ciência da informação**, v. 24, n. 1, p. 36-41, 1995. Disponível em: https://brapci.inf.br/_repositorio/2011/05/pdf_aac5068b8b_0016893.pdf. Acesso em: 05 ago. 2020.

SARMENTO, L. **Simpósio Doutoral Linguateca**. Lisboa: Linguateca, 2005. Disponível em: http://www.linguateca.pt/documentos/SimposioDoutoral2005.html. Acesso em: jul. 2020.

SHALABY, W. *et al.* Entity Type Recognition Using an Ensemble of Distributional Semantic Models to Enhance Query Understanding. *In*: INTERNATIONAL COMPUTER SOFTWARE AND APPLICATIONS CONFERENCE. 1., 2016. **Anais […],** [*s.l*], IEEE Computer Society. Disponível em: https://doi.org/10.1109/COMPSAC.2016.109. Acesso em: 05 ago. 2020

SILVA, E. M. da; SOUZA, R. R. Fundamentos em processamento de linguagem natural: uma proposta para extração de bigramas**. Encontros Bibli,** [*s.l.*], v. 19, n. 40, p. 1-32, mai./ago., 2014. Disponível em: encurtador.com.br/gnz29. Acesso em: 22 jul. 2020.

SOUZA, R. R. Sistemas de recuperação de informações e mecanismos de busca na web: panorama atual e tendências. **Perspectivas em Ciência da Informação**, [s. l.], v. 11, n. 2, p. 161–173, 2006. Disponível em: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1413-99362006000200002&lng=pt&tlng=pt. Acesso em: 19 ago. 2020.

TOMMASEL, A.; GODOY, D. Short-text feature construction and selection in social media data: a survey. **The Artificial Intelligence Review**, Dordrecht, v. 49, n. 3, p. 301-338, mar. 2018.

VALENTIM, M. L. P. **Métodos qualitativos de pesquisa em Ciência da Informação**. São Paulo: Polis, 2005.

VASILIEV, Y. **Natural Language Processing with Python and spaCy**: A Practical Introduction. San Francisco: No Starch Press, 2020. Disponível em: https://books.google.com.br/books/about/Natural_Language_Processing_with_Python.html?id=Au-_DwAAQBAJ&redir_esc=y. Acesso em: 23 jul. 2020.

WANG, B*. et al*. Short text classification based on strong feature thesaurus. **Journal Zhejiang Univ. Sci. C**., v. 13, n. 9, p. 649–659, 2012. Disponível em: https://www.bibsonomy.org/bibtex/1fe057628762a7bea9aa869ccad525efa. Acesso em: 15 fev. 2021.

WINOGRAD, T. **Procedures as a Representation for Data in a Computer Program for Understanding Natural Language**. 1971. Disponível em: https://dspace.mit.edu/handle/1721.1/7095. Acesso em: 23 jul. 2020.

YAN, E. **Towards a systematic approach for studying scholarly communication through scholarly networks**. Indiana: Indiana University, 2013.

ZHAI, C. X.; MASSUNG, S. **Text Data Management and Analysis**: A Practical Introduction to Information Retrieval and Text Mining. New York: Association for Computing Machinery and Morgan, 2016.

ZHANG, W. *et al*. Improving effectiveness of mutual information for substantival multiword expression extraction. **Expert Systems with Applications**, Elsevier, v. 36, 2009.