UNESP – UNIVERSIDADE ESTADUAL PAULISTA
DCSO – Departamento de Comunicação Social
FAAC – Faculdade de Arquitetura, Artes e Comunicação
Curso de Comunicação Social – Jornalismo

O TEXTO COMO DADO:
A CONVERGÊNCIA ENTRE JORNALISMO E *BIG DATA*

Bauru, 2017

Guilherme Nóbrega Costa

O TEXTO COMO DADO:
A CONVERGÊNCIA ENTRE JORNALISMO E *BIG DATA*

Memorial de Projeto Experimental apresentado em cumprimento parcial às exigências do Curso de Jornalismo da Faculdade de Arquitetura, Artes e Comunicação (FAAC), do Departamento de Comunicação Social (DCSO), da Universidade Estadual Paulista "Júlio de Mesquita Filho" (UNESP), para obtenção do título de Bacharel em Comunicação Social – Jornalismo.

Orientador(a) do Projeto Experimental:
Profa. Adj. Maria Cristina Gobbi

Bauru, 2017

Dedico este trabalho a meus
futuros desafios. Que sirva de recordação
de que nada está além do alcance
da mente decidida e
do enfrentamento diligente.

Agradecimentos

Agradeço aos meus pais pelo apoio
incondicional, aos meus amigos pela
compreensão e incentivo e principalmente
à Profa. Maria Cristina Gobbi e
ao Prof. Juliano Ferreira de Sousa
pela orientação, paciência e atenção.

Esvazie sua mente. Seja amorfo, sem definições, como a água. Se colocarmos a água num copo, ela se torna o copo; se colocarmos numa garrafa, ela se torna a garrafa; se colocarmos numa chaleira, ela se torna a chaleira. A água pode fluir, ou pode colidir. Seja água, meu amigo.

Bruce Lee

RESUMO

Os avanços da tecnologia na década atual têm trazido à luz o termo *Big Data* em várias áreas do conhecimento e do mercado. O termo abrange, basicamente, vários aspectos relacionados a dados que se apresentam em grandes volumes, são gerados de maneira muito rápida, são variados, complexos e estão conectados uns aos outros. O campo do Jornalismo e as empresas de mídia de forma geral, entretanto, ainda não estão suficientemente familiarizadas com as possibilidades geradas pelas análises de *Big Data*. Para contribuir para a reversão desse quadro, o presente trabalho traz uma pesquisa aplicada na área de análise de *Big Data*, focando em análise de dados não-estruturados, ou seja, sem formato ou padrões definidos, mais especificamente no formato de texto. Trazendo um exemplo prático, uma introdução ao *software Knime* e um exemplo de análise utilizando suas funcionalidades foi conduzido e explicado, visando aumentar a proximidade do jornalista ou estudante de jornalismo com esse tipo de ferramenta. As fontes dos dados utilizados consistem em textos de trabalhos de conclusão de curso dos alunos dos cursos de Jornalismo e Relações Públicas da UNESP. A conclusão da análise demonstra que *insights* úteis podem ser tirados de dados em formato de texto e que jornalistas podem utilizar esse conhecimento para melhorar seu próprio trabalho, atualizar-se perante ao mercado e enriquecer seu valor como profissional.

Palavras-chave: *Big Data*; análise de dados; *Knime*; texto e *string*; jornalismo.

LISTA DE FIGURAS

SUMÁRIO

1 INTRODUÇÃO

A consolidação da tecnologia como entidade ubíqua nos espaços pessoais e de convívio social não são mais uma novidade dentro do contexto atual. A segunda década do século XXI trouxe avanços que rapidamente nascem, crescem e permeiam copiosamente as nossas vidas, muitas vezes sem que ao menos notemos sua presença e seus efeitos na nossa rotina e no mundo ao nosso redor.

A necessidade de conhecer cada vez mais essas tecnologias, por conseguinte, tem invadido todos os campos do conhecimento, na proporção em que se torna útil e muitas vezes indispensável para a promoção da inovação e da própria sustentação de atividades, ofícios, profissões, empresas e instituições.

O campo do Jornalismo é, invariavelmente, afetado por essas mudanças. Da invenção da prensa móvel até a ascensão das mídias sociais, a profissão de jornalista têm sido desafiada a se reinventar para adaptar-se aos sucessivos "novos" tempos.

O objetivo primário desse trabalho é introduzir ao estudante e/ou profissional à teoria de uma das tendências mais recentes trazidas pelos avanços tecnológicos da década atual: a análise de grandes quantidades de dados. De prevenção de desastres naturais até análises de tendências sociais; de previsões econômicas precisas até gerenciamento de crises em tempo real; de marketing pessoal direcionado até monitoramento de estados de saúde e humor. Nada escapa ao olhar do chamado *Big Data*. Essa grande quantidade de dados que "fluem" por entre nós nos abre possibilidades infinitas de enxergar o mundo de uma perspectiva que, se não "onisciente", busca ser mais próximo de tal.

Em segundo lugar, com um caráter mais relevante e prático, essa produção busca contextualizar o *Big Data* dentro da profissão de jornalista, dando ao leitor uma base teórica para um entendimento mais aprofundado do funcionamento de ferramentas de análise de *Big Data*, focando em conceitos essenciais para o seu instrumento de trabalho: o texto. Além disso, debruça-se sobre uma ferramenta específica, o software *Knime*, desenvolvido para facilitar a manipulação de dados em grande escala.

A metodologia de pesquisa utilizada não se apoia amplamente em pesquisas bibliográficas, pois essas são raras e ainda não estão largamente consolidadas em livros e publicações, principalmente no que diz respeito à convergência dos campos de análise

de dados e de jornalismo. A pesquisa exploratória foi preferida, aliada à cursos online recentemente concluídos pelo autor nas áreas de Programação, Modelamento e Gerenciamento, Integração e Processamento e Aprendizado de Máquina em *Big Data*.

É importante notar, ainda, que alguns assuntos pertinentes a esse universo como as necessidades computacionais de *hardware* e *software* para o estabelecimento de um sistema de *Big Data* e as capacitações em conceitos anexos como linguagens de programação não serão aqui abordados. Os serviços hoje disponíveis no mercado para a montagem ou aluguel de tais sistemas permitem a utilização e expansão de recursos de maneira autônoma, tornando-se cada vez mais plausível para organizações menores e até indivíduos estabelecerem seu ecossistema de *Big Data*. Notadamente, os recursos de *software*, em específico, são em sua grande maioria gratuitos, uma vez que são desenvolvidos em plataformas *open-source* (código aberto).

Ademais, o programa de *Big Data Knime* foi em grande medida escolhido justamente por reduzir drasticamente a necessidade de conhecimentos complexos na área de programação para sua operação, diminuindo também a necessidade da existência de um time dedicado de profissionais de Tecnologia da Informação para extração de valores dos dados analisados, aumentando, adicionalmente, a independência do profissional de Jornalismo nessa área.

O trabalho está dividido basicamente em três partes. A primeira discorre sobre as origens e o conceito de *Big Data*, para ambientar o leitor no contexto dos avanços que o constituem. Num segundo momento, o leitor é conduzido a conceitos mais específicos sobre dados, visando caracterizá-los, destacando o próprio texto como fonte de dados. Por fim, a experimentação com o software *Knime* visa demonstrar algumas das capacidades dessa ferramenta para manipular dados em grande escala, gerando visualizações da informação e possíveis *insights* a partir de uma coletânea de textos analisada. O tom didático presente em recorrência ao longo do texto visa, por fim, facilitar o contato do jornalista ou estudante de jornalismo com estruturas em forma de dados e tem, em última instância, o objetivo de despertar seu interesse para esse campo diverso e complexo porém cada vez mais essencial para a vida moderna. Espera-se, assim, que essa apresentação o leve a uma busca do aperfeiçoamento pessoal e profissional, contribuindo para cada vez mais enriquecer e destacar o valor de nossa profissão.

2 DADOS E SEUS CONTEXTOS

2.1 O que é *Big Data*?

Big Data é essencialmente um conceito que funciona como um "guarda-chuva" para uma extensa variedade de estratégias e táticas que envolvem um número muito grande de dados e as respectivas tecnologias que os captam, transformam, analisam e deles retiram *insights* (descobertas).

A aparição do termo e a utilização recorrente do mesmo vem alinhada a algumas características marcantes do desenvolvimento tecnológico dos últimos anos, principalmente às que definem as novas capacidades de dispositivos móveis e o universo das redes sociais. Segundo relatório produzido pelo *McKinsey Global Institute*:

> Em vários setores e regiões, várias tendências convergentes têm motivado a geração de dados e continuarão a abastecer a rápida expansão de bases de dados. Essas tendências incluem o crescimento nos *databases* transacionais tradicionais, a contínua expansão do conteúdo multimídia, a crescente popularidade das redes sociais e a proliferação de aplicativos de sensores da Internet das Coisas[1]. (MANYIKA et al., 2011, p.21)

Outro relatório produzido pelo mesmo instituto aponta que existem atualmente mais de 2.5 bilhões de pessoas utilizando a *Internet* e estima-se que esse número será de mais de 5 bilhões em 2025 (MANYIKA et al., 2013, p. 62). Nos primeiros anos da década atual, entretanto, o crescimento de diferentes meios de acesso e conectividade revela uma tendência crucial para a compreensão do fortalecimento do *Big Data*. Há 1.1 bilhões de pessoas atualmente utilizando *smartphones* e *tablets* (2013, p.29). O número de dispositivos conectados, porém, é de cerca de 9 bilhões (2013, p.51). Isso se dá principalmente pela ascensão dos dispositivos *wearable* (vestíveis), como relógios e pulseiras providos de sensores que captam dados do usuário em tempo real, e dos dispositivos da chamada *Internet of Things* (Internet das Coisas), como geladeiras e televisões inteligentes que possuem conectividade à internet e se "relacionam" cada vez mais com o usuário, com outros dispositivos e com os dados disponíveis na rede.

[1] Across sectors and regions, several cross-cutting trends have fueled growth in data generation and will continue to propel the rapidly expanding pools of data. These trends include growth in traditional transactional databases, continued expansion of multimedia content, increasing popularity of social media, and proliferation of applications of sensors in the Internet of Things.

Espera-se que o número de dispositivos conectados cresça drasticamente nos próximos anos, com estimativas para a próxima década de alcançarem dentre 50 bilhões até 1 trilhão. Estima-se que em 2025, aproximadamente 80% de todas as conexões feitas à *Internet* será realizada a partir de dispositivos móveis (MANYIKA et al., 2013, p.32). A finalidade dessas conexões também engloba cada vez mais áreas de nossa vida. A proliferação de aplicativos para virtualmente qualquer fim (trabalho, educação, entretenimento, comunicação, comércio, transporte, dentre outros) e a aquisição de novos formatos e interfaces intuitivas nos mesmos já ocasiona uma mudança generalizada no estilo de vida das pessoas e nos modelos de negócios em todos os setores da economia.

A quantidade de dados que circula em redes sociais também vem crescendo exponencialmente, não só devido ao engajamento dos próprios usuários, mas principalmente aos avanços tecnológicos que intermeiam essas relações, os quais se tornam cada vez mais poderosos e acessíveis. Nos últimos 5 anos, a capacidade de processamento dos dispositivos móveis cresceu em torno de 25% por ano, provendo um aumento significante na capacidade de produção, armazenamento e transmissão de arquivos (MANYIKA et al., 2013, p.30). Seguindo a mesma tendência, o avanço em poder computacional dos sistemas de armazenamento e processamento de dados em "nuvem" reduzem a dependência de grandes investimentos para acesso desses dados.

A acessibilidade à computação em "nuvem" a um público cada vez menos segregado possibilita que pequenas empresas ou até mesmo indivíduos possam manipular esses dados e deles extrair informações úteis para vários fins. Em suma, a conectividade ininterrupta aliada aos diferentes modos de expressão e interação *online* têm criado vias plurais e permanentes de transmissão e captação de dados, o que era impossível ou impraticável há apenas alguns anos. Esse constante fluxo de dados em grande quantidade está cada vez mais acessível, gerando um cenário de oportunidades inéditas e valiosas de análise sobre o mundo em que vivemos.

A descrição do *Big Data*, portanto, não pode estar desvencilhada desse contexto que permeia seu surgimento. Por isso, ele é comumente descrito como contendo cinco "V"s.

O primeiro "V" é referente à "volume". Um grande volume de dados pode ser atribuído tanto a um único ou múltiplos *datasets* conectados, tanto quanto a vários

fragmentos de dados coletados ao longo do tempo. O volume no âmbito estudado, como o adjetivo *big* sugere, é de grande proporção. Segundo ilustra Stone (2014, p.1) um vídeo de 7 minutos em alta-definição ocupa cerca de 1 *gigabyte* em volume de armazenamento. Essa quantidade de dados é facilmente capturada e armazenada em computadores pessoais atuais. Em contraste, servidores e sistemas de *Big Data* geralmente trabalham com volumes na ordem de *terabytes* (1 mil *gigabytes*) e *petabytes* (1 milhão de *gigabytes*). Um *petabyte* de armazenamento equivale à 13,3 anos de vídeos em alta-definição sendo continuamente reproduzidos. O site *Youtube*, propriedade da empresa *Google*, processa mais de 24 *petabytes* de dados por dia. Para as próximas eras de avanços digitais, já existem termos como *exabytes* (1 bilhão de *gigabytes*), *zettabytes* (1 trilhão de *gigabytes*), *yottabytes* (10 elevado à potência 24 *bytes*) e *brontobytes* (10 elevado à potência 27 *bytes*). De acordo com um relatório produzido através da parceria entre a empresa de inteligência em marketing global *IDC* e a *Dell EMC*, braço da empresa *Dell* para fornecimento de infraestrutura digital de tecnologia da informação, no ano de 2020 a humanidade produzirá 44 *zettabytes* de informação (2014, p.2). Outra previsão, presente no relatório de quantificação e previsão de crescimento de tráfego de dados da empresa *Cisco* aponta que a quantidade de dados de *Big Data* chegará a 247 *exabytes* em 2020, praticamente 10 vezes mais do que o existente no ano de 2015, cerca de 25 *exabytes* (2016, p.3).

O segundo "V" aborda o problema da velocidade da geração de dados. Segundo informações compiladas pela consultoria em tecnologia *Excelacom* (2016) sobre a quantidade de dados que circula na *Internet*, em apenas 1 minuto cerca de 2,4 milhões de buscas são feitas no buscador *Google*, 2,78 milhões de visualizações em vídeos no *Youtube,* mais de 20,8 milhões de mensagens são trocadas no aplicativo *WhatsApp*, mais de 347 mil *posts* são feitos no site *Twitter*, mais de 38 mil fotos são colocadas no aplicativo *Instagram*, 69,4 horas de vídeo são assistidas no serviço de *streaming* de filmes *Netflix* e 38 horas de música armazenadas no serviço de *streaming* de música *Spotify*. É importante notar também que o crescimento desses números anualmente retrata um uso cada vez mais expressivo da *Internet*. Comparativamente, as principais diferenças apontadas pela própria consultoria em suas análises relativas aos anos de 2015 e 2016 foram um aumento de 100% no número de corridas agendadas pelo aplicativo *Uber* (695 corridas a mais por minuto), 70% de aumento nas vendas do site

Amazon (83,836 mil dólares a mais por minuto) e 186% de acréscimo na quantidade de músicas disponíveis no aplicativo *Spotify* (24752 horas disponibilizadas a mais por minuto).

O processamento de dados em tempo real para acompanhar a imensa velocidade da produção dos mesmos é uma característica notável do *Big Data*. Esse tipo de capacidade, por exemplo, permite a personalização de propagandas em páginas na *web* visitadas pelo usuário baseando-se em pesquisas, visualizações e compras recentes. Ao fazer uso de dados em tempo real ou dentro do tempo necessário, empresas e instituições podem se assegurar de que não estão perdendo oportunidades de investimento ou de realização de ações cruciais para atingir seus objetivos e seu público. Informações como as de caráter meteorológico, por exemplo, tem um alto valor quando são analisadas o mais próximo possível do momento em que acontecem, tendo em vista obter respostas imediatas, precisas e atualizadas. A previsão do tempo se encaixa nesse quadro, pois geralmente se procura saber como estará o tempo daqui a algumas horas ou dias. As informações sobre horas, dias ou meses anteriores são, nesse caso, irrelevantes.

Entretanto, não podemos assumir que a velocidade de captação e análise da informação, mesmo dentro do campo do *Big Data*, sejam importantes apenas quando se aproximam ao máximo da velocidade da geração dos seus objetos. O processamento em lotes, conhecido como *batch processing*, era a "norma" do campo de análise de dados antes do desenvolvimento das capacidades de processamento em tempo real. Nesse tipo de processamento, grandes quantidades de dados são captadas e alimentam dispositivos de armazenamento, para só então serem analisadas gradativamente. Um ciclo de análise em etapas é pré-determinado e pode ser demarcado tanto de acordo com um tamanho específico de "pacote" de dados por vez, ou de acordo com períodos de tempo estabelecidos. Apesar de sua natureza, esse tipo de análise ainda é expressivamente relevante no universo do *Big Data* em casos como prevenção e gerenciamento de crises ou catástrofes.

Informações meteorológicas, como no exemplo citado acima, podem ser armazenadas para a criação de um *database* que prevê a formação e expansão de focos de incêndios naturais, por exemplo; algo particularmente relevante em países de clima ou estações secas. Medições de sensores meteorológicos como as de temperatura, direção do vento e umidade relativa do ar, armazenadas por um longo período, podem

revelar padrões como meses e regiões de maior incidência de incêndios de causas naturais. Ainda, cruzando-se essas informações com as obtidas em tempo real, pode-se prever a taxa de expansão e direção provável de um incêndio florestal, possibilitando assim que ações bem-informadas sejam tomadas para solucionar o problema em questão. Além disso, à medida que novas informações chegam, os resultados são adaptados para condizer com as mudanças ocasionadas.

O terceiro "V", "veracidade", refere-se à qualidade dos dados ingeridos. Os objetos contidos num ambiente de *Big Data*, assim como qualquer dado captado, podem estar sujeitos a vários "ruídos". Alterações, erros ou ausência de informações-chave, por exemplo, são comuns em grandes volumes de dados. Em alguns casos, dados podem ser incertos ou simplesmente conter informações não-verdadeiras. Essa característica também pode ser chamada de "validade" ou "volatilidade", em referência adicional ao tempo necessário para que um dado se torne "inútil" ou inválido para análise.

Um caso notável que gerou matérias em *websites* de notícias como *Buzzfeed*[2], *Huffington Post*[3] e *Business Insider*[4] e análises em revistas publicitárias como *Adweek*[5] é o da página do utensílio de cozinha "*Hutzler 571 Banana Slicer*" no site de vendas online *Amazon*. O produto contém, até a presente data, mais de 5,6 mil avaliações de compradores. Em sua maioria, são avaliações positivas, dando ao produto uma pontuação de 4.4 numa escala de 5 pontos satisfatórios. Entretanto, a grande maioria das avaliações mais votadas por outros usuários, ou seja, as que aparecem nos primeiros lugares da página, são comentários feitos com a intenção de serem sarcásticos ou cômicos.

[2] https://www.buzzfeed.com/sirajdatoo/amazon-reviews-of-this-plastic-banana-slicer-are-just-the-be?utm_term=.viYe8Dra5#.gp3av3rpb
[3] http://www.huffingtonpost.com/2013/02/01/ridiciulous-banana-slicer-thousands-mocking-amazon-reviews_n_2600008.html
[4] http://www.businessinsider.com/banana-slicer-funny-amazon-reviews-2016-6
[5] http://www.adweek.com/creativity/maker-stupid-banana-slicer-thrilled-sarcastic-reviews-amazon-146527/

Figura 1 – Seção da página do produto *Hutzler 571 Banana Slicer*

Fonte: Captura de tela feita pelo autor (2017)

Até o presente momento, mais da metade (53%) do total de avaliações do produto foram “5 estrelas”, ou seja, a pontuação máxima dada por um usuário. Todavia, mais de 56 mil pessoas avaliaram o primeiro comentário da página intitulado “Sem mais vitórias para você, senhor banana!” como útil. Esse contraste coloca claramente em cheque a qualidade de uma possível análise da satisfação do consumidor com o produto em questão, visto que muitas das avaliações positivas podem simplesmente não ter relação alguma com a satisfação real do comprador.

Desse modo, o dados perdem seus valores para análise, uma vez que podem gerar conclusões distorcidas sobre a situação analisada e, em consequência, direcionar

possíveis investimentos em direções contraprodutivas. Portanto, as evidências extraídas do *Big Data* dependem diretamente de um dado de qualidade satisfatória. Assim, fatores como a confiabilidade das fontes e a precisão e modo como os dados foram gerados são essenciais para a checagem da veracidade da informação.

O quarto "V" alude à "variedade" de tipos de dados que podem ser encontrados. Dados em forma de texto, imagens, sons, vídeos, informações de rede, mapas e simulações computacionais são alguns tipos que compõem essa diversidade. Essa heterogeneidade pode ser observada dentro de algumas dimensões como diferentes estruturas, meios e semânticas.

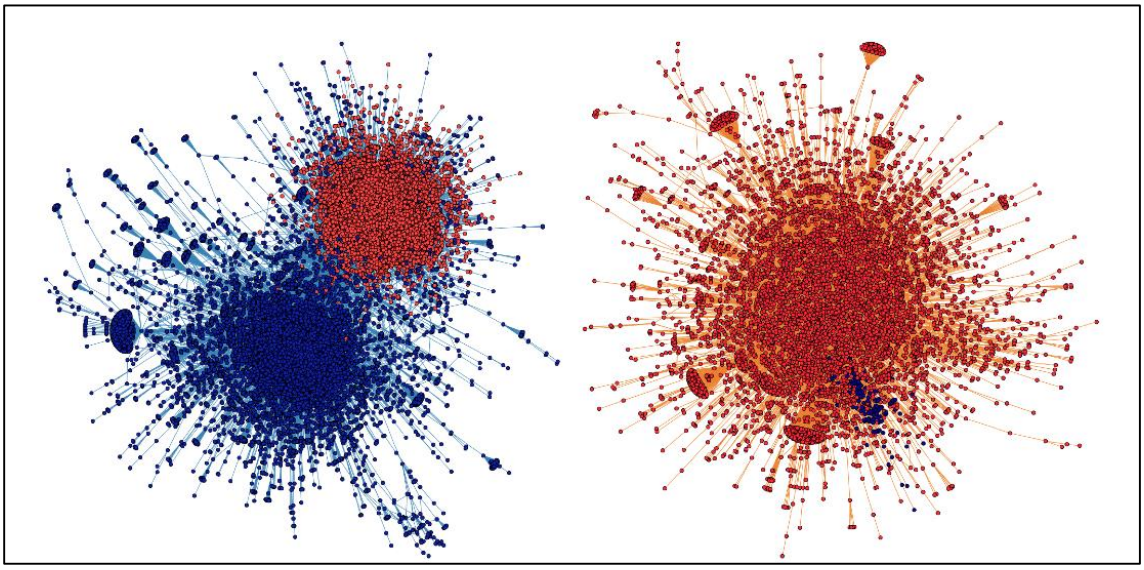
Variações estruturais podem ser observadas quando há diferenças no modo como os dados são representados. A clara diferença entre o modo como sinais de um eletrocardiógrafo e um artigo de jornal disponibilizam uma informação é uma amostra comparativa dessa situação. Já a variedade dos meios pode ser representada na comparação entre o áudio de um discurso e sua transcrição, que apesar de conterem a mesma informação e estrutura, essa está disponível em meios diferentes. A diferença semântica por sua vez é melhor visualizada quando se pensa em atribuições de origem quantitativa e qualitativa. Por exemplo, uma informação sobre o campo "idade" pode ser representada tanto por um número quanto por um termo como "criança", "adolescente" ou "adulto". Convém observar, ainda, que pode haver variedade no acesso aos dados. Uma câmera de tráfego em uma estrada pode ser acessada a qualquer momento, continuamente; já fotos de satélite de uma região específica, apenas quando o mesmo se encontra acima do local de interesse.

Por último, a "valência" configura o quinto "V". O conceito se refere à quantidade de "conectividade" de um dado com outros dados, assim como no termo "valência" utilizado para caracterizar elétrons no campo da Química.

Objetos de um pacote de dados estão geralmente conectados uns aos outros, direta ou indiretamente. Por exemplo, uma cidade está sempre diretamente conectada ao estado que pertence, ou dois ou mais usuários do *Facebook* estão conectados porque são amigos. Em termos indiretos, podemos traçar uma ligação entre dois usuários do *LinkedIn* que tenham o mesmo tipo de cargo em diferentes empresas, independente se eles estão de fato conectados um ao outro nessa rede social. Nesse contexto, quanto

mais conectados os dados estão, mais "densos" se tornam suas conexões e, por conseguinte, há maior possibilidade de outras conexões surgirem.

Figura 2 – *Clusters* de *tweets* de tema político



Fonte: CONOVER et al., 2011, p.92

A figura acima reproduzida, extraída de um estudo sobre a polarização política no *Twitter,* mostra as diferenças de conectividade entre dados dessa rede social. A aglomeração à esquerda é uma visualização dos *retweets* e suas conexões uns com os outros. À direita temos a visualização de "menções". Podemos observar regiões onde dados se aglomeram ou se tornam esparsos e as áreas onde há conexão entre seus diferentes tipos. No caso, os nódulos coloridos na cor azul representam postagens definidas como contendo cunho liberal; já em vermelho, as de cunho conservador.

Podemos dizer, ainda, que os 5 "Vs" do *Big Data* estão apoiados em um sexto, "valor". O que o torna valiosos todos os aspectos descritos anteriormente são as inúmeras maneiras as quais podemos utilizá-los para obter novas visões sobre o mundo. A análise de *Big Data* proporciona a construção de decisões mais completas e objetivas sobre o mundo, levando a resultados de maior precisão e confiabilidade.

Segundo a analogia publicada pela empresa de *software Tableau*:

> Um lago de dados é como um reservatório feito pelo homem. Primeiramente, você deve construir uma represa (construir uma unidade de alocação) e depois enchê-la de água (dados). Uma vez pronto, pode-se começar a usar a água (dados) para várias funções, como gerar eletricidade, matar a sede ou para recreação (análises preditivas, aprendizado de máquina, segurança cibernética, etc)[6] (TABLEAU, 2017, p.5)

Assim, vemos que uma análise de *Big Data* consiste em um processo que busca a aplicação efetiva desses dados em contextos definidos, visando o desenvolvimento e enriquecimento de uma ou múltiplas áreas, baseando-se nas possibilidades descobertas e nos *insights* que foram obtidos.

[6] A data lake is like a man-made reservoir. First you dam the end (build a cluster), then you let it fill up with water (data). Once you establish the lake, you start using the water (data) for various purposes like generating electricity, drinking, and recreating (predictive analytics, ML, cyber security, etc.).

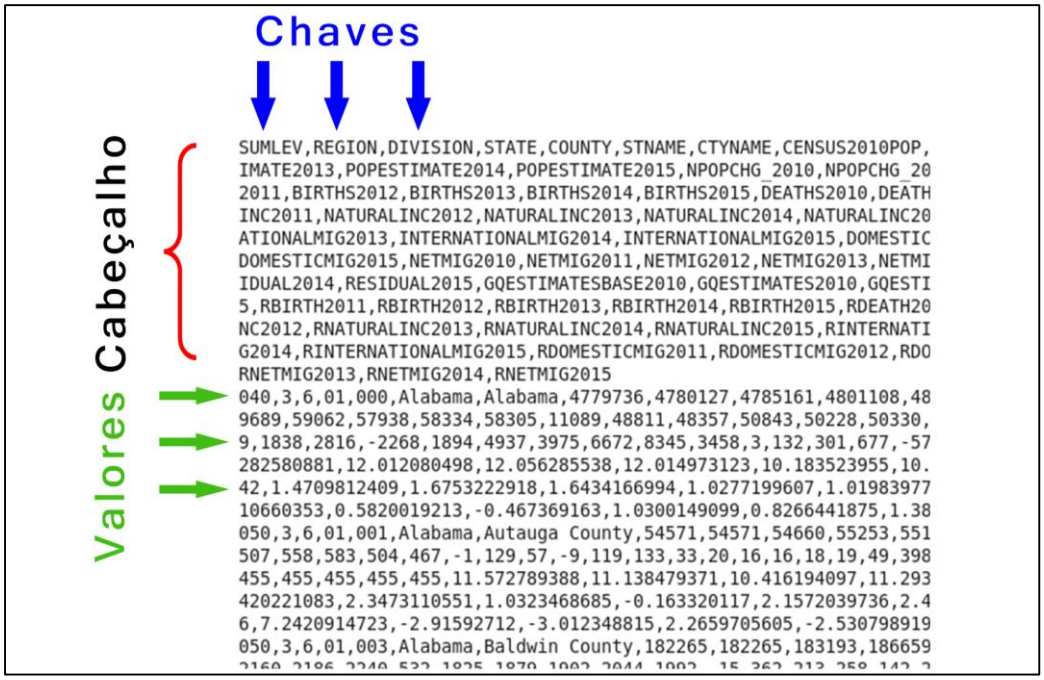
2.2 Quando o texto se torna *string*

Dentro do universo lexical de análise de dados, termos como “dados estruturados”, “semiestruturados” e “não-estruturados” são frequentemente encontrados. Esses três modelos básicos definem como devemos lidar com os dados e determinam, em última instância, as ferramentas necessárias ou recomendadas para manipulá-los.

Como abordado na primeira seção deste trabalho, dados existem em uma infinidade de variações. Uma tabela pode ser estruturada, por exemplo, contendo os campos “nome”, “sobrenome” e “idade”. Essa imposição estrutural impede, logicamente, que números sejam colocados no campo “nome” e “sobrenome”. Por outro lado, operações matemáticas como a média de idade dos referenciados na mesma podem apenas ser realizadas usando-se os dados do campo “idade”. Além disso, podemos inferir que um registro do campo “data de nascimento” contendo um ano de cinco casas (ex: 19799) contém um erro, pois um ano específico deve ser um número de apenas quatro casas. Essas características estruturais, operacionais e constritivas são exemplos de como um modelo de dados se define e varia de caso a caso.

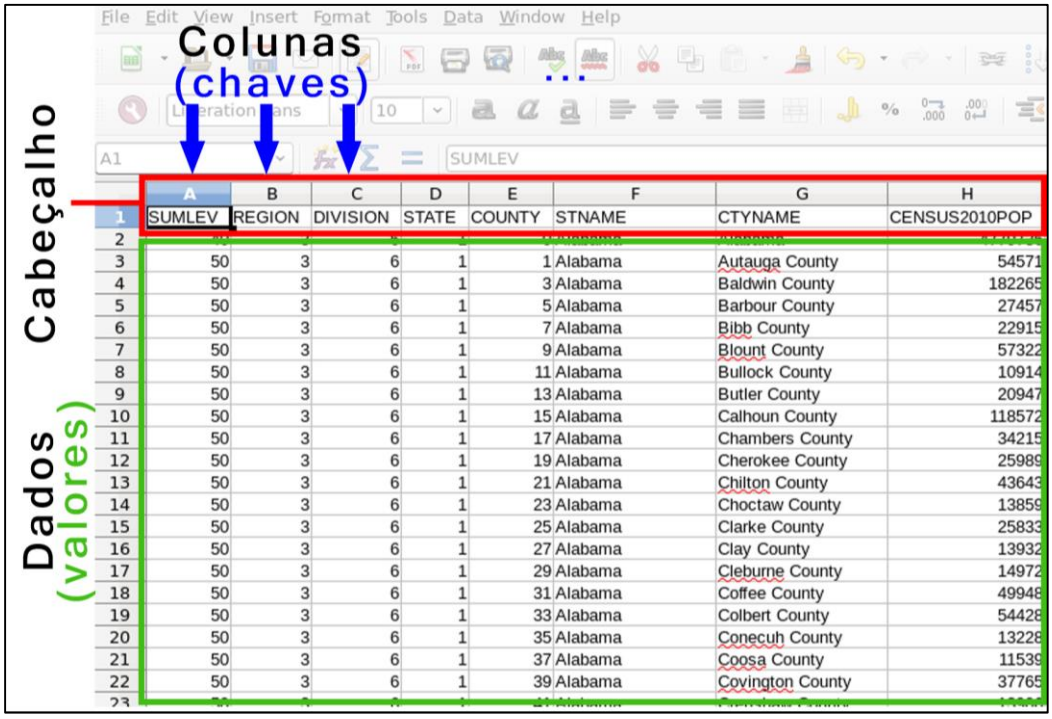
Um modelo de dados estruturado segue um esquema fixo pré-definido (SINT et al., s.d., p.3) e “promove um processamento de dados eficiente e armazenamento e navegação de conteúdo otimizado” (ABITEBOUL et al., 1999, p.122 *apud* SINT et al., s.d., p.3). Considerado simples, é basicamente um modelo “chave-valor” que “permite a visualização do banco de dados como uma grande tabela *hash*” e “que os dados sejam rapidamente acessados pela chave” (LÓSCIO et al., s.d., p.6). Nele a “chave” representa um campo e o “valor” uma instância correspondente atribuída para esse campo. Um arquivo do tipo CSV (*Comma Sepparated Values)* possui essa característica. Na figura 4, observamos a disposição das chaves e valores dentro do arquivo em sequência, como o próprio nome diz, em “valores separados por vírgula”. Na figura 5, o mesmo arquivo aberto em um software apropriado para sua visualização pode ser compreendido com mais clareza. Apesar de softwares como *Microsoft Excel* e *OpenOffice Calc* serem compatíveis com esse tipo de arquivo, não são capazes de manipulá-los quando estes são muito grandes. O arquivo em questão possui 3194 linhas; porém, se tivesse milhões ou mais linhas, não conseguiríamos visualizá-lo em uma tabela. Seria necessário importá-lo para um sistema de *Big Data* como o *Hadoop*.

Figura 3 – Arquivo CSV



Fonte: Captura de tela e edição feita pelo autor (2017)

Figura 4 – Arquivo CSV visualizado no *software OpenOffice Calc*



	A	B	C	D	E	F	G	H
1	SUMLEV	REGION	DIVISION	STATE	COUNTY	STNAME	CTYNAME	CENSUS2010POP
3	50	3	6	1	1	Alabama	Autauga County	54571
4	50	3	6	1	3	Alabama	Baldwin County	182265
5	50	3	6	1	5	Alabama	Barbour County	27457
6	50	3	6	1	7	Alabama	Bibb County	22915
7	50	3	6	1	9	Alabama	Blount County	57322
8	50	3	6	1	11	Alabama	Bullock County	10914
9	50	3	6	1	13	Alabama	Butler County	20947
10	50	3	6	1	15	Alabama	Calhoun County	118572
11	50	3	6	1	17	Alabama	Chambers County	34215
12	50	3	6	1	19	Alabama	Cherokee County	25989
13	50	3	6	1	21	Alabama	Chilton County	43643
14	50	3	6	1	23	Alabama	Choctaw County	13859
15	50	3	6	1	25	Alabama	Clarke County	25833
16	50	3	6	1	27	Alabama	Clay County	13932
17	50	3	6	1	29	Alabama	Cleburne County	14972
18	50	3	6	1	31	Alabama	Coffee County	49948
19	50	3	6	1	33	Alabama	Colbert County	54428
20	50	3	6	1	35	Alabama	Conecuh County	13228
21	50	3	6	1	37	Alabama	Coosa County	11539
22	50	3	6	1	39	Alabama	Covington County	37765

Fonte: Captura de tela e edição feita pelo autor (2017)

Modelos semiestruturados são compostos de “dados que não se encaixam em campos fixos mas contém *tags* e outros marcadores que separam os seus elementos” (MANYIKA et al., 2011, p.33). Isso não significa, porém, que não haja uma certa estruturação; indica, apenas, que o modelo pode acomodar variações em sua estrutura (SINT et al, s.d., p.4).

Figura 5 – Seção do código HTML da página inicial do aplicativo *Spotify*

Fonte: Captura de tela e edição feita pelo autor (2017)

O arquivo em formato HTML (*Hypertext Markup Language*), comumente utilizado para a construção de páginas da *web* pode ser pensado como uma “árvore” de componentes que estão alojados dentro de outros, que por sua vez alojam outra série de componentes e assim sucessivamente (figura 6). Blocos podem ser manipulados, retirados e adicionados, e cada documento conterá um número diferente de blocos e associações. Em outras palavras, é um modelo mais flexível que o anteriormente apresentado.

Arquivos no formato JSON (*Java Script Object Notation*), deixam ainda mais claro a presença de partes estruturadas e não-estruturadas. No excerto mostrado na figura 7, retirado de um conjunto de postagens extraídas da rede social *Twitter*, observamos as estruturas que compõe uma postagem específica (nome do usuário,

número de seguidores, número de menções, etc). A parte grifada representa o texto que foi postado, que é um segmento não-estruturado. Além de elementos de linguagem natural, a postagem pode conter *hashtags*, menções e outras idiossincrasias próprias do funcionamento da rede social, não seguindo nenhum padrão pré-definido. Essa característica nos leva ao terceiro modelo: dados não-estruturados.

Figura 6 – Excerto de conjunto de dados em JSON

Fonte: Captura de tela e edição feita pelo autor (2017)

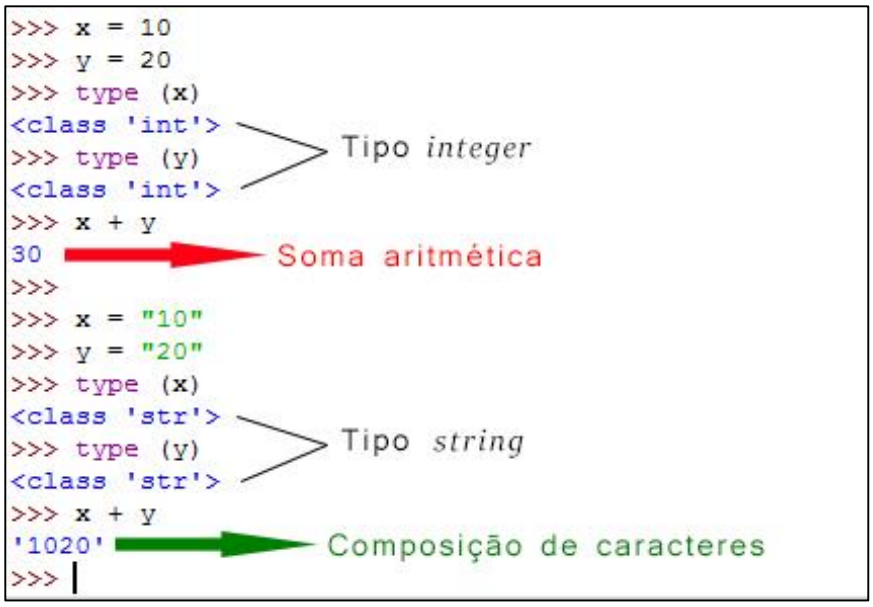
Para entender a natureza da composição de dados não-estruturados é importante ter em mente a distinção entre tipos de dados. O tipo de um dado define que operações podem ser executadas com sucesso para criar, transformar e utilizar uma variável específica dentro do modelo abordado. Modelos de dados podem aceitar um ou mais tipos de dados em sua estrutura.

Adicionando ao exemplo citado no início desta seção, um arquivo que contenha os campos “nome”, “sobrenome” e “idade” pode conter, como efetivamente ocorre no contexto da rede social observada acima, um campo “nome de usuário”. O nome de usuário, neste caso, pode conter números e símbolos em conjunto à letras do alfabeto. Um registro como “Maria_123” seria, então, válido nesse campo.

O tipo de dado em questão é denominado dentro do campo da programação um dado *string*. Em tradução livre, *string* remete a uma *sequência conectada*. Assim, um dado desse tipo “informa” ao modelo que habita que é constituído de uma “sequência de caracteres”. Em contraste, um dado do tipo *integer* (número inteiro) “informa” ao

modelo que é um número, aceitando, assim, operações matemáticas. Da mesma maneira, o tipo *float* (ponto flutuante) vai além, pois representa um número fracionário, aceitando valores "quebrados". Já numa distinção mais radical, um dado do tipo *boolean* aceita apenas dois valores: "verdadeiro" e "falso". Dessa maneira, um registro sobre o status "*online*" e "*offline*" de um usuário de rede social é um registro "booleano", pois não há outras opções possivelmente existentes para esse registro.

Figura 7 – Comportamento de dados do tipo *integer* e *string*



Fonte: Captura de tela e edição feita pelo autor (2017)

As operações realizadas acima em um interpretador da linguagem de programação *Python* demonstram na prática como a determinação de um tipo de dado afeta as operações e análises realizadas. Nas primeiras linhas, as variáveis "x" e "y" foram definidas, respectivamente, como "10" e "20". Por padrão, o interpretador atribui aos números o tipo *integer* (abreviado como *int*). Uma operação "x + y" resulta, então, no valor "30". Na segunda parte, o valor atribuído a "x" e "y" são os mesmos números, contudo, a adição das aspas comunica ao interpretador que esses são apenas "uma sequência de caracteres". Naturalmente, a soma de "x + y" resulta na junção dos caracteres: "1020".

A importância da análise de dados do tipo *string* não reside somente no fato de que esse está presente massivamente em dados não-estruturados, mas sim na observação de que a grande maioria dos dados que são gerados e circulam no mundo atualmente são, justamente, do tipo não-estruturados.

Segundo projeções apresentadas pela empresa de tecnologia computacional *Oracle* (2014, p.2) apenas 12% dos dados existentes são estruturados. Nesse âmbito estão informações geradas por máquinas, sensores e dispositivos, como estações meteorológicas, satélites, ou proveniente das próprias análises computacionais e de bancos de dados.

Assim, o restante (88%) estão no domínio das informações não-estruturadas. Mensagens de e-mail, resultados de enquetes, postagens em redes sociais, mensagens trocadas em aplicativos (incluindo por imagens, áudio e vídeo), conteúdo de *websites* e *blogs*, etc, são parte desse modelo, pois a maioria desses conteúdos não contém atributos ou relações específicas. Em suma, o desafio do modelo não-estruturado é intrínseco, uma vez que uma "navegação controlada dentro de um conteúdo não-estruturado não pode ser feita" (SINT et al., s.d., p.3).

Para lidar com esse obstáculo, principalmente quando os conteúdos analisados estão em formato de texto (*string*), técnicas específicas podem ser usadas. Uma delas é a própria criação de "estruturas" para o texto através do modelo de espaço vetorial. Nesse modelo:

> [...] um documento é representado por um vetor no qual cada elemento determina o peso ou a importância do respectivo termo na representação do conteúdo informacional do documento. Cada vetor descreve a posição do documento em um espaço multidimensional onde cada termo de indexação representa uma dimensão ou eixo (FERNEDA; LOPES, s.d., p.7)

Para melhor visualização do conceito, tomemos como exemplo três arquivos doravante determinados "Doc 1", "Doc 2" e "Doc 3". Visando simplificar a aplicação da técnica, imaginemos, ainda, que cada arquivo contenha apenas três palavras: "Doc 1" contendo "Estado", "São" e "Paulo"; "Doc 2" contendo "Folha", "São" e "Paulo"; "Doc 3" contendo "Folha", "Minas" e "Gerais", como demonstrado na figura 9.

Figura 8 – Tabela de frequência de termos (3 documentos)

	Minas	Gerais	São	Estado	Folha	Paulo
Doc. 1	0	0	1	1	0	1
Doc. 2	0	0	1	0	1	1
Doc. 3	1	1	0	0	1	0

Fonte: Tabela elaborada pelo autor (2017)

A matriz elaborada demonstra quantas vezes cada termo aparece em cada documento. Convenciona-se, no campo de análise de dados, dar a esse tipo de tabela a nomenclatura "TF", abreviação de *term frequency* ou "frequência de termos". Temos, então, que os termos "São", "Paulo" e "Folha" aparecem 2 vezes no total na coleção abordada. Já "Minas", "Gerais" e "Estado" aparecem cada um 1 vez no total.

Numa classificação de milhões de documentos contendo milhões de palavras no total, entretanto, é natural que várias palavras se repitam com uma frequência muito maior do que outras, como conjunções e pronomes. Assim, para equilibrar o uso de termos, é necessária a criação de um fator de "penalidade", que diminua estatisticamente a "importância" dessa palavra em relação a outras. Também se deve, então, elaborar uma matriz "IDF" (*inverse term frequency* ou "frequência inversa de termos"). A frequência inversa de um termo pode ser calculada, convencionalmente, como:

$$\log_2\left(\frac{T}{N}\right)$$

Em que T representa o número total de documentos e N o número de vezes que o termo específico em questão aparece dentro do universo de documentos analisados. Desse modo, uma visualização do número de vezes que um termo aparece no conjunto e o seu respectivo "IDF" pode ser:

Figura 9 – Cálculo do IDF

Termo	Frequência total nos documentos	IDF
Minas	1	$\log_2\left(\frac{3}{1}\right) = 1{,}584$
Gerais	1	$\log_2\left(\frac{3}{1}\right) = 1{,}584$
São	2	$\log_2\left(\frac{3}{2}\right) = 0{,}584$
Estado	1	$\log_2\left(\frac{3}{1}\right) = 1{,}584$
Folha	2	$\log_2\left(\frac{3}{2}\right) = 0{,}584$
Paulo	2	$\log_2\left(\frac{3}{2}\right) = 0{,}584$

Fonte: Tabela elaborada pelo autor (2017)

Multiplicando-se os valores "TF" e "IDF" de cada termo no documento, tem-se um valor específico resultante para cada um deles, representando a importância relativa de cada termo dentro do universo de palavras analisado. Ainda, os valores dos termos podem dar a cada documento singular um "valor total" específico, comumente chamado

de *length* (comprimento), calculado como a raiz quadrada da soma dos quadrados dos valores de cada termo presente no documento. Dessa maneira:

$$Doc\ 1 = \sqrt{(0{,}584)^2 + (1{,}584)^2 + (0{,}584)^2} = 1{,}786$$

$$Doc\ 2 = \sqrt{(0{,}584)^2 + (0{,}584)^2 + (0{,}584)^2} = 1{,}011$$

$$Doc\ 3 = \sqrt{(1{,}584)^2 + (1{,}584)^2 + (0{,}584)^2} = 2{,}315$$

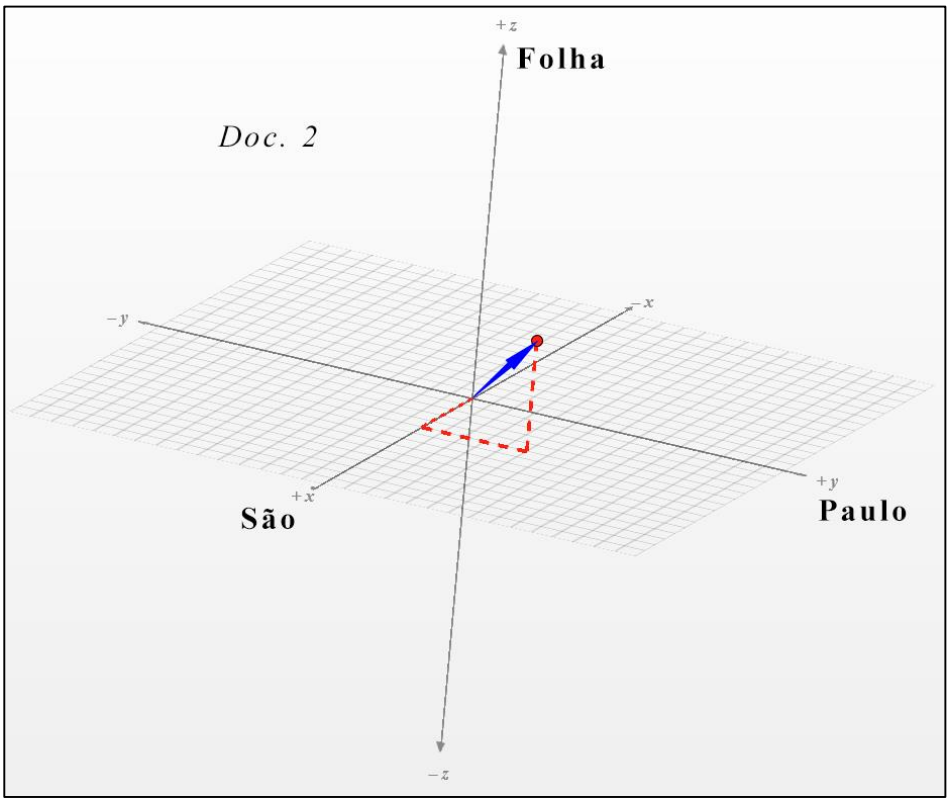
Os valores da matriz "TF-IDF" e o seu comprimento podem ser observados na figura 11. Cada linha da tabela representa, então, um vetor, e cada palavra uma dimensão do espaço vetorial. À título de simplificação de visualização, o vetor representado pela linha "Doc. 2" foi desenhado de maneira rudimentar na figura 12, utilizando-se apenas 3 eixos: "Folha" como "z", "São" como "x" e "Paulo" como "y".

Figura 10 – Tabela TF-IDF

	Minas	Gerais	São	Estado	Folha	Paulo	Comprimento
Doc. 1	0	0	0,584	1,584	0	0,584	1,768
Doc. 2	0	0	0,584	0	0,584	0,584	1,011
Doc. 3	1,584	1,584	0	0	0,584	0	2,315

Fonte: Tabela elaborada pelo autor (2017)

Figura 11 – Vetor "Doc 2" representado em um espaço vetorial



Fonte: Gráfico elaborado pelo autor (2017)

A geração de vetores para cada documento cria, então, uma maneira de compará-los uns aos outros ou de trazer ao usuário (de um sistema de *Big Data*, por exemplo) os resultados mais relevantes em uma busca ou análise. Segundo Ferneda e Lopes, "A utilização de uma mesma representação tanto para os documentos como para as expressões de busca permite calcular o grau de similaridade (distância vetorial) entre uma determinada busca e cada um dos documentos do *corpus*" (s.d., p.7). O cálculo dessa similaridade pode ser feito utilizando-se várias expressões diferentes, como uma função cosseno do ângulo formado entre dois vetores analisados. Uma observação relevante é que um usuário de um referido sistema de manipulação de dados pode atribuir "pesos" diferentes aos termos, ocasionando assim a alteração nos cálculos e nos valores atribuídos a cada palavra. Também pode, ainda, elaborar uma lista de palavras

ou expressões para serem excluídas totalmente. Desse modo, ele poderá direcionar melhor sua análise de acordo com seus objetivos.

O armazenamento de dados não-estruturados e a análise dos mesmos utilizando técnicas como a descrita anteriormente não são, por si só, novidades. Os campos da Tecnologia e Ciências da Informação e as empresas desses ramos lidam com esse tipo de problema há décadas. Contudo, o interesse crescente desencadeado pela explosão das informações disponíveis na rede tem trazido esses problemas e oportunidades cada vez mais para próximo de outros campos do conhecimento e indústrias.

Abordando essa questão de maneira pioneira, a autora Martha Stone, pesquisadora do Instituto *Reuters* para Estudo do Jornalismo da Universidade de Oxford produziu um relatório que problematiza a posição das indústrias de mídia e jornalismo dentro do advento do *Big Data*. Durante os anos de 2013 e 2014, Stone liderou duas conferências sobre "*Big Data* para a Mídia", as primeiras conferências no mundo voltadas exclusivamente para esse tema.

Palestras dadas nesses eventos apontam que a indústria de mídia já se move em direção à tendência de análise de dados em grandes quantidades. Empresas como o grupo *Financial Times* mantêm "impressões digitais" de seus assinantes para entender as necessidades de seus consumidores, uma estratégia de modelo de negócios. Já o grupo *CNN* amplia o uso do *Big Data* para a própria produção de notícias. Em parceria com a rede social *Twitter* e a empresa de tecnologia *Dataminr*, analisa o fluxo de 500 milhões de *tweets* diários para criar um sistema de "alerta" para possíveis notícias de primeira mão que possam estar na iminência de acontecerem ou tenham acabado de surgir (STONE, 2014, p.13). Porém, ainda segundo Stone:

> O *Big Data* é uma imensa oportunidade para empresas de mídia. Entretanto, em geral, elas ainda estão longe de extrair o máximo de benefícios de uma estratégia de *Big Data*. Se uma estratégia de *Big Data* pudesse ser dividida como uma pirâmide [...] as empresas de mídia estariam aproximadamente num nível intermediário, entre os níveis de "informação" e "conhecimento"[7] (STONE, 2014, p.2)

[7] Big Data is a huge opportunity for media companies, but media companies in general are still far away from extracting maximum benefit from a Big Data strategy. If a Big Data strategy could be broken down like a […] pyramid, media companies' level would fall somewhere in the middle, between the information and knowledge stages.

De fato, como outras oportunidades do mundo digital que vieram antes dela como a *Internet*, as próprias mídias sociais e os aplicativos móveis, a oportunidade do *Big Data* é muitas vezes tratada com certa resistência e ceticismo (STONE, 2014, p.9). Com o intuito de "quebrar" essa resistência, a visão apresentada nesse trabalho se assemelha à de Wilfried Runde, gerente de inovações e projetos da empresa de mídia alemã *Deutsche Welle*. Segundo Stone, Runde defende que todas as redações deveriam investir em *Big Data* tanto em posicionamento, como no desenvolvimento de habilidades e na exposição a diferentes cenários.

Seguindo a mesma tendência, algumas publicações de referência dentro do universo jornalístico já alertam para a necessidade dessa mudança. Como nota a versão mais atualizada do manual de Jornalismo 2017 da *Associated Press*:

> O Jornalismo de Dados se tornou um pilar da reportagem em várias áreas e plataformas. A habilidade de analisar informações quantitativas e apresentar conclusões de maneira estimulante e precisa não é mais responsabilidade de especialistas somente. Governos, empresas e outras organizações se comunicam em linguagem de dados e estatísticas. Para cobrí-los, jornalistas devem se tornar fluentes nessas linguagens também[8] (ASSOCIATED PRESS, 2017)

A conclusão, então, é estimular o jornalista a aprender a "nadar" nesse "oceano" de dados. Como Stone (2014) resume em sua análise, dois dos desafios-chave para trazer o *Big Data* para as redações são a aplicação de recursos e treinamento (p.14). Da vasta gama de *softwares* e sistemas disponíveis para análise de dados, o programa *Knime*, de agora em diante abordado, traz algumas vantagens que, na opinião do autor, são ideais para um primeiro porém profícuo "mergulho" no universo do *Big Data*.

[8] Data journalism has become a staple of reporting across beats and platforms. The ability to analyze quantitative information and present conclusions in an engaging and accurate way is no longer the domain of specialists alone. Government agencies, businesses and other organizations alike all communicate in the language of data and statistics. To cover them, journalists must become conversant in that language as well.

3 O SOFTWARE *KNIME*

O software *Knime*, abreviação de *Konstanz Information Miner* é um programa de análise de dados desenvolvido por engenheiros de *software* da Universidade de Konstanz, em Konstanz, Alemanha. Lançado em 2006 e inicialmente um produto voltado para análise de dados e pesquisas da indústria farmacêutica, seu uso se expandiu ao longo dos anos para análises de CRM (*Customer Relationship Management*), inteligência empresarial, dados financeiros e outros. Segundo descrição disponível no *website* da empresa o produto foi desenvolvido para:

> [...] inovação orientada por dados, desenvolvido para descobrir o potencial escondido em dados, mineração para insights inovadores ou prever novos futuros. Organizações podem alcançar um novo patamar de colaboração, produtividade e performance [...] em todo tipo de dados: de números à imagens, moléculas à pessoas, sinais à redes complexas e de simples estatísticas até análises de *Big Data*[9] (KNIME, 2017)

Como descrito acima, uma das grandes vantagens do uso do software consiste em sua capacidade para a análise de variados tipos de dados, sejam eles estruturados, semiestruturados ou não-estruturados. Desse modo, apresenta-se como uma ferramenta versátil que permite ao usuário explorar dados de diferentes fontes, contextos e formatações.

É imprescindível notar, também, que como a maioria dos softwares e plataformas voltadas para o *Big Data*, o software *Knime* é uma ferramenta em desenvolvimento *open source*. Em outras palavras, qualquer pessoa tem licença gratuita para obter, utilizar e modificar sua estrutura, sendo inclusive encorajada a adição de melhorias voluntárias ao design do *software*. Por esse motivo, a empresa *Knime.com AG*, responsável oficial por sua atualização, disponibilização e desenvolvimento mantém um fórum *online* para ensino, perguntas, sugestões e relatos de análises feitas pelos usuários, formando uma comunidade ativa e colaborativa.

O programa também aceita ser executado em vários sistemas operacionais como *Windows*, *Mac OSX* e *Linux*, seja para acessar sistemas de arquivos de *Big Data* como o

[9] […] data-driven innovation, designed for discovering the potential hidden in data, mining for fresh insights, or predicting new futures. Organizations can take their collaboration, productivity and performance to the next level […] on every kind of data: from numbers to images, molecules to humans, signals to complex networks, and simple statistics to big data analytics.

Hadoop ou sistemas de arquivos do próprio computador pessoal comum do usuário. Apesar de seu código interno ser escrito em linguagem *Java*, aceita adições de funcionalidades para ler outras linguagens como *Python* e *Perl*.

O principal diferencial desse *software* em comparação a outros *softwares* similares voltados para a manipulação de dados é a sua interface. As funções e operações presentes em um *software* desse tipo normalmente são executadas através de um comando direto do usuário utilizando um código digitado em uma linguagem de programação específica, como mostra a figura 12 (uma visualização de comandos de uma biblioteca *Spark* em um sistema de arquivos *Hadoop*, ou HDFS). No caso, a chave "IN" corresponde à função descrita pelo usuário e "OUT" a resposta obtida. A primeira linha informa ao programa o documento a ser acessado, e a segunda linha chama a função "*count*" e o argumento "( )", o que em linguagem de programação *Python* significa, basicamente, a "contagem de todas as linhas do arquivo" de texto mencionado.

Figura 12 – Comandos em Python para acessar arquivos de *Big Data*

Fonte: Captura de tela feita pelo autor (2017)

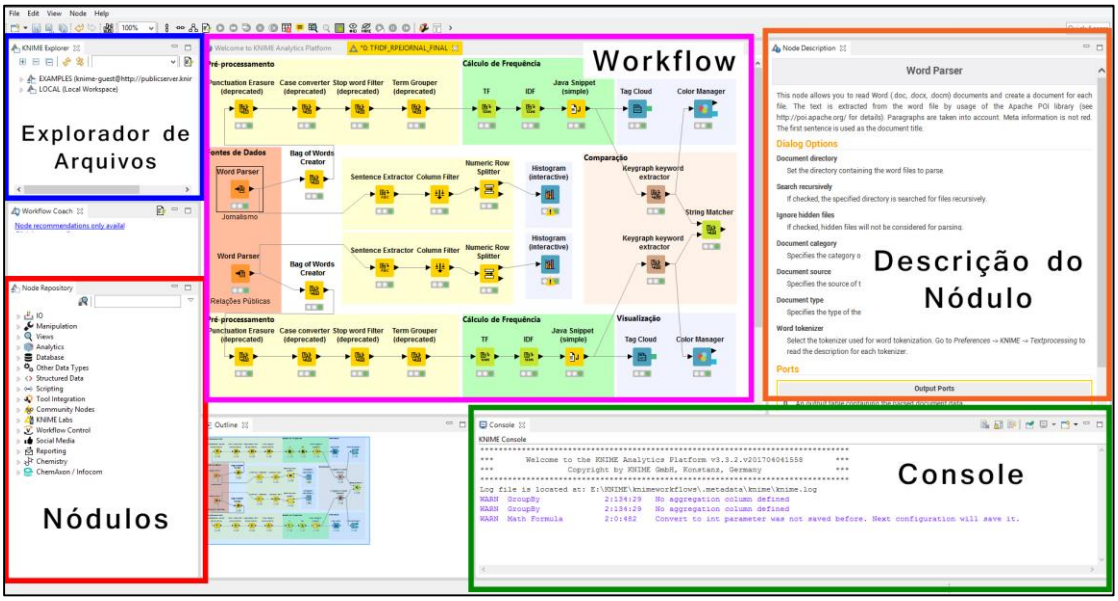
A interface do *software Knime* porém apresenta-se de maneira muito mais intuitiva e simples para um usuário que não domine nenhuma linguagem de programação. O programa possibilita que o usuário crie os chamados *workflows*, representações visuais de funções que são executadas nos arquivos e dados escolhidos. O usuário vai seletivamente criando, conectando e executando alguns ou vários passos

de uma análise na ordem em que desejar, montando visualmente uma "teia" de nódulos (representação visual de linhas de comandos específicos) e conexões em formatos de seta (representação visual da próxima função que deve ser executada pelo programa). Cada nódulo ainda pode ser configurado independentemente e pode gerar visualizações também independentes de cada passo do processo.

Unindo todas as características descritas anteriormente, podemos inferir que o público a que o *software* se destina é bastante amplo. Abrangendo múltiplas necessidades em variados níveis de complexidade, executando em diferentes capacidades de *hardware* e voltado para dados de diferentes tamanhos, modelos e tipos, *Knime* é capaz de atender grandes empresas até pequenas organizações ou indivíduos.

Para melhor compreender a interface e as funções do programa, uma análise-exemplo foi montada pelo autor, utilizando uma amostra de um conjunto de dados em formato de texto. A amostra foi obtida do *database* do catálogo Athena da Unesp de Bauru, mais especificamente retirada aleatoriamente dos trabalhos de conclusão de curso dos alunos dos cursos de Relações Públicas e Jornalismo. Os arquivos foram obtidos em formato PDF (*.pdf*) e transformados em arquivos Word (*.docx*) para melhor leitura do programa. O equipamento utilizado para realizar a análise foi o computador pessoal do próprio autor, executando o sistema operacional *Windows 8.1*, com um processador *Intel Core i7-2700K* e memória RAM de 16 GB.

Figura 13 – Interface do *software Knime*



Fonte: Captura de tela e edição feita pelo autor (2017)

A interface do programa consiste em 5 “janelas” básicas. A área em azul (canto superior esquerdo) mostra o explorador de arquivos do *workspace*, ou seja, a pasta definida pelo usuário em que são armazenados os *workflows* criados ou obtidos de outros lugares (como baixados da *Internet*).

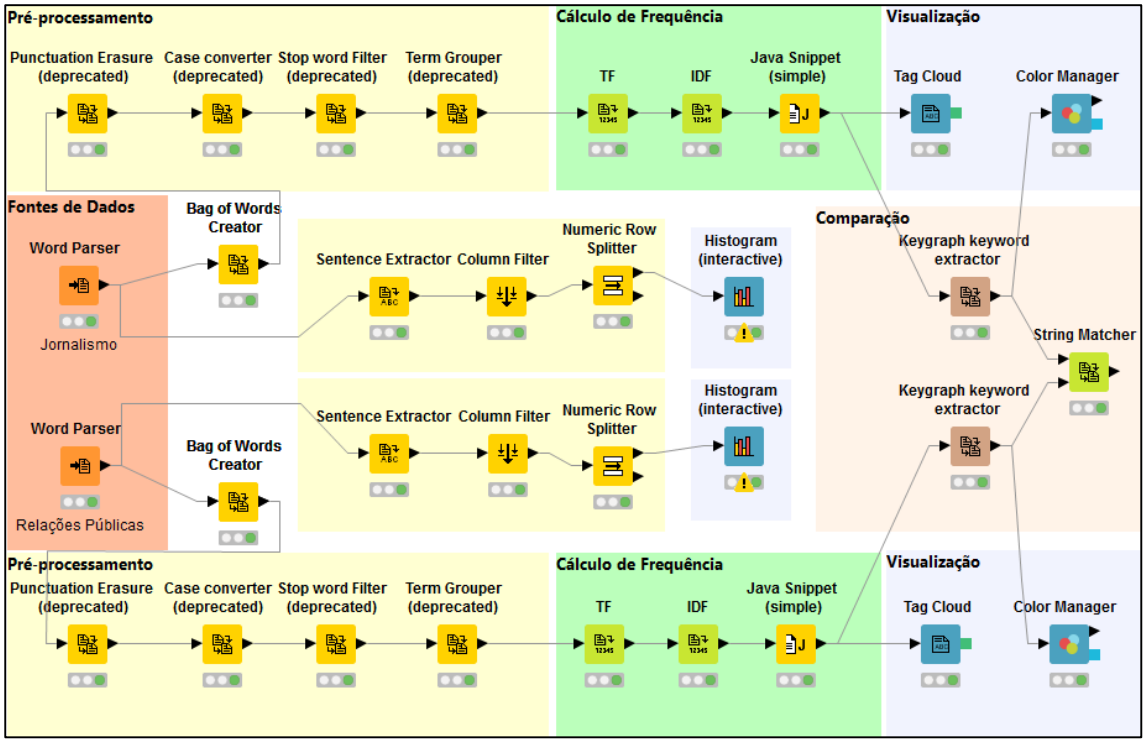
O canto inferior esquerdo, em vermelho, é o repositório de nódulos. Nesse espaço o usuário tem acesso a diferentes e específicos tipos de nódulos para montar seu *workflow*. Os nódulos são separados por categorias, como transformação (filtros, conversores, combinadores) de dados e outros métodos para análise e visualização dos mesmos.

Do repositório, o usuário escolhe e “arrasta” as funções escolhidas até a área central (área em magenta), onde o *workflow* pode ser montado e modificado. Os nódulos normalmente contém um ponto de entrada (*input*) onde recebem os dados, e um ponto de saída (*output*), para onde os dados modificados por ele vão. Ao clicar em um nódulo no *workflow*, uma descrição detalhada de sua operação pode ser visualizada na parte direita superior da tela (área em laranja).

O sistema montado no *workflow* pode ser executado, pausado ou reiniciado utilizando-se os ícones da parte superior da janela do *software*. O *status* de sua execução pode ser visualizado no console na parte inferior direita (área em verde). Ali aparecem em forma de texto os erros que foram encontrados durante a execução e notificações, por exemplo.

No exemplo criado (figura 14), o *workflow* contém 2 cadeias de processos independentes entre si. Cada cadeia realiza 2 tipos de análises diferentes em paralelo, que serão descritas em detalhes adiante. Os nódulos foram posicionados de maneira a facilitar o entendimento e o “fluxo” da informação, utilizando-se retângulos coloridos ao fundo de cada segmento do processo para melhor visualização. Ao final, os resultados de duas das cadeias são comparados.

Figura 14 – *Workflow* completo



Fonte: Captura de tela feita pelo autor (2017)

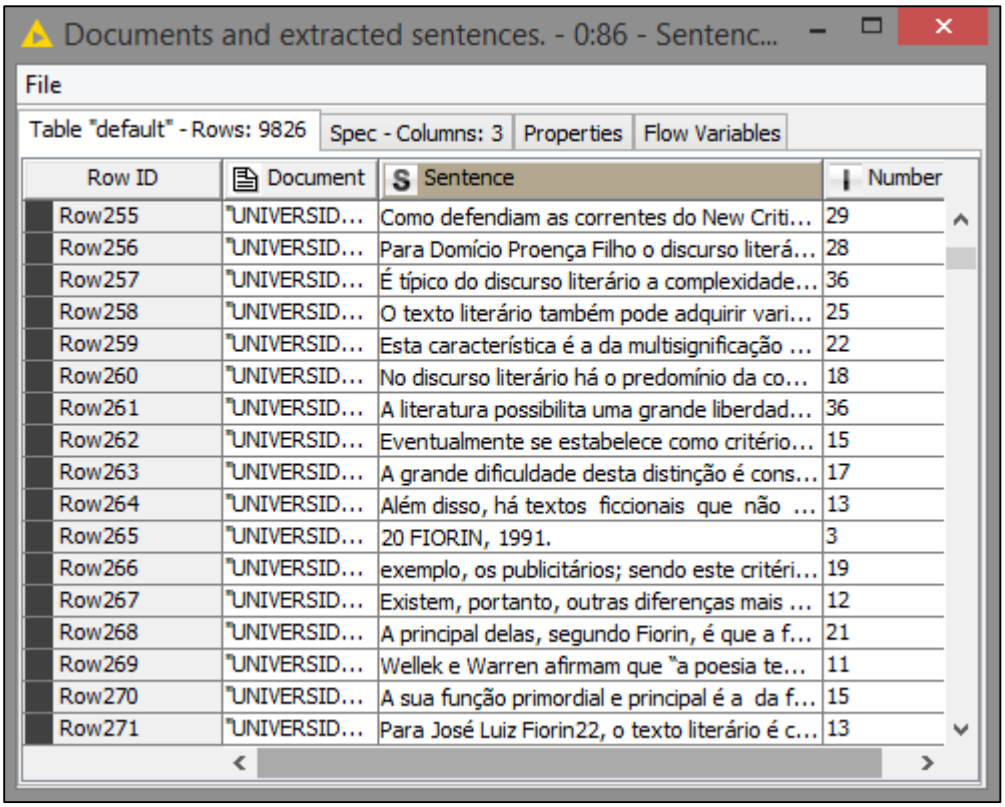
Os processos são iniciados pelos dois primeiros nódulos à esquerda (fundo de cor salmão). No caso, os nódulos escolhidos são do tipo "*Word Parser*", ou seja, um leitor de arquivos provenientes do programa *Word* (*.doc ou .docx*). No caso, ao nódulo de cima foi atribuído uma pasta do sistema operacional do autor onde estão armazenados 5 arquivos *.docx* contendo o texto de 5 TCCs de alunos do curso de Jornalismo. O mesmo foi feito para o nódulo de baixo, porém especificando outra pasta do sistema contendo 5 TCCs de alunos do curso de Relações Públicas. Somando os valores de todos os documentos dos dois cursos, o sistema analisará 1130 páginas, contendo 344.834 palavras num total de 2.210.673 caracteres.

Dos dois nódulos são "disparados" 4 fluxos de processos, 2 para Jornalismo e o outros 2 para Relações Públicas, sendo os fluxos paralelamente idênticos e contendo as mesmas configurações internas.

Começando pela análise do fluxo menor (no centro da figura 14), vemos que o próximo nódulo é um "*Sentence Extractor*", responsável por extrair cada frase do

universo de documentos buscados e colocá-las individualmente em uma tabela. Ao clicar com o botão direito do *mouse* nesse nódulo, temos a opção de visualizar o resultado de seu processamento. No nódulo "*Sentence Extractor"* dos TCCs do curso de Jornalismo, temos:

Figura 15 – Frases Extraídas



Row ID	Document	Sentence	Number
Row255	"UNIVERSID...	Como defendiam as correntes do New Criti...	29
Row256	"UNIVERSID...	Para Domício Proença Filho o discurso literá...	28
Row257	"UNIVERSID...	É típico do discurso literário a complexidade...	36
Row258	"UNIVERSID...	O texto literário também pode adquirir vari...	25
Row259	"UNIVERSID...	Esta característica é a da multisignificação ...	22
Row260	"UNIVERSID...	No discurso literário há o predomínio da co...	18
Row261	"UNIVERSID...	A literatura possibilita uma grande liberdad...	36
Row262	"UNIVERSID...	Eventualmente se estabelece como critério...	15
Row263	"UNIVERSID...	A grande dificuldade desta distinção é cons...	17
Row264	"UNIVERSID...	Além disso, há textos ficcionais que não ...	13
Row265	"UNIVERSID...	20 FIORIN, 1991.	3
Row266	"UNIVERSID...	exemplo, os publicitários; sendo este critéri...	19
Row267	"UNIVERSID...	Existem, portanto, outras diferenças mais ...	12
Row268	"UNIVERSID...	A principal delas, segundo Fiorin, é que a f...	21
Row269	"UNIVERSID...	Wellek e Warren afirmam que "a poesia te...	11
Row270	"UNIVERSID...	A sua função primordial e principal é a da f...	15
Row271	"UNIVERSID...	Para José Luiz Fiorin22, o texto literário é c...	13

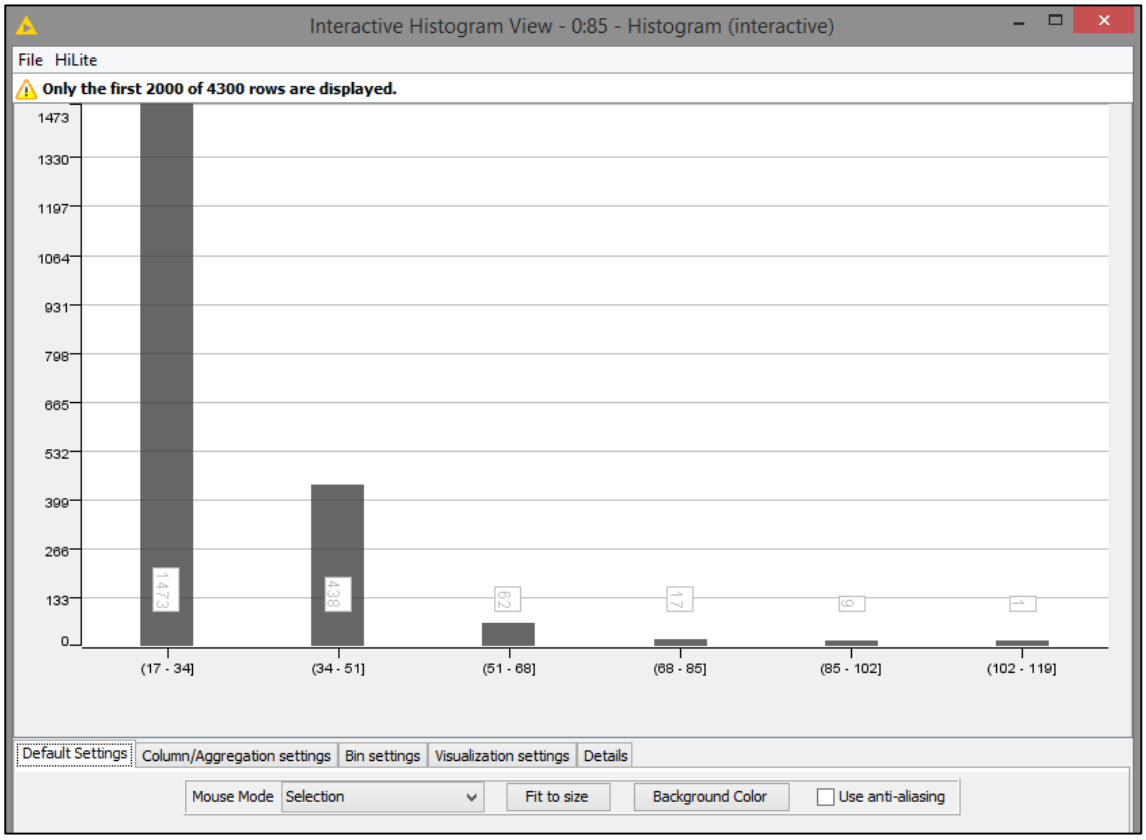
Fonte: Captura de tela feita pelo autor (2017)

O próximo nódulo "*Column Filter*" foi adicionado para retirar da tabela produzida a coluna "*Document*", que mostra de qual documento específico dentro universo de documentos analisado a frase veio. Essa coluna não é relevante para a análise. Já o nódulo "*Numeric Row Splitter"* tem a função de separar uma tabela de números em duas porções: a que se conforme a uma regra definida, e a que não se conforme. Nesse caso, a regra definida foi a exclusão de frases que contenham um número de termos muito baixo para serem informativas, como visto na linha número

265 em que a "frase" é apenas uma referência ao autor e ano de uma citação. Arbitrariamente, o valor mínimo decidido para a regra foi de 17 termos ou mais. Em outras palavras, o nódulo passará para a próxima etapa apenas os registros, no caso, frases, que contenham 17 termos ou mais. Como a parte que sobrou (17 termos ou menos) não é útil nesse caso, foi puxada para próxima etapa apenas a seta de cima, que são os dados aprovados na regra. Se a parte excluída fosse útil (seta de baixo do nódulo), também poderia ser usada para seguir um outro fluxo de análises.

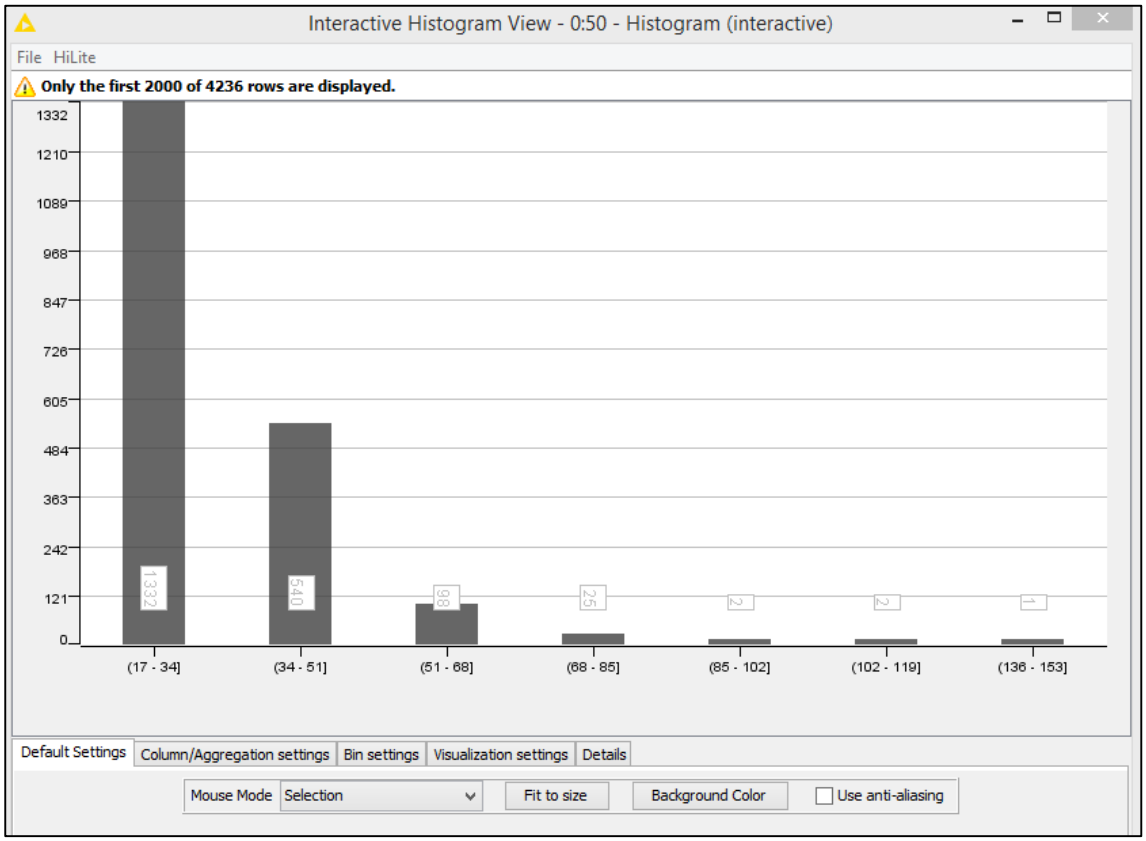
O último nódulo desse fluxo "*Histogram (Interactive)"* apresenta os resultados visualmente em formato de histograma (gráfico de barras), contendo algumas configurações interativas como mostrar a distribuição das informações ao clique do *mouse*. Sendo assim, o resultado das duas análises foi:

Figura 16 – Histograma de Jornalismo



Fonte: Captura de tela feita pelo autor (2017)

Figura 17 – Histograma de Relações Públicas



Fonte: Captura de tela feita pelo autor (2017)

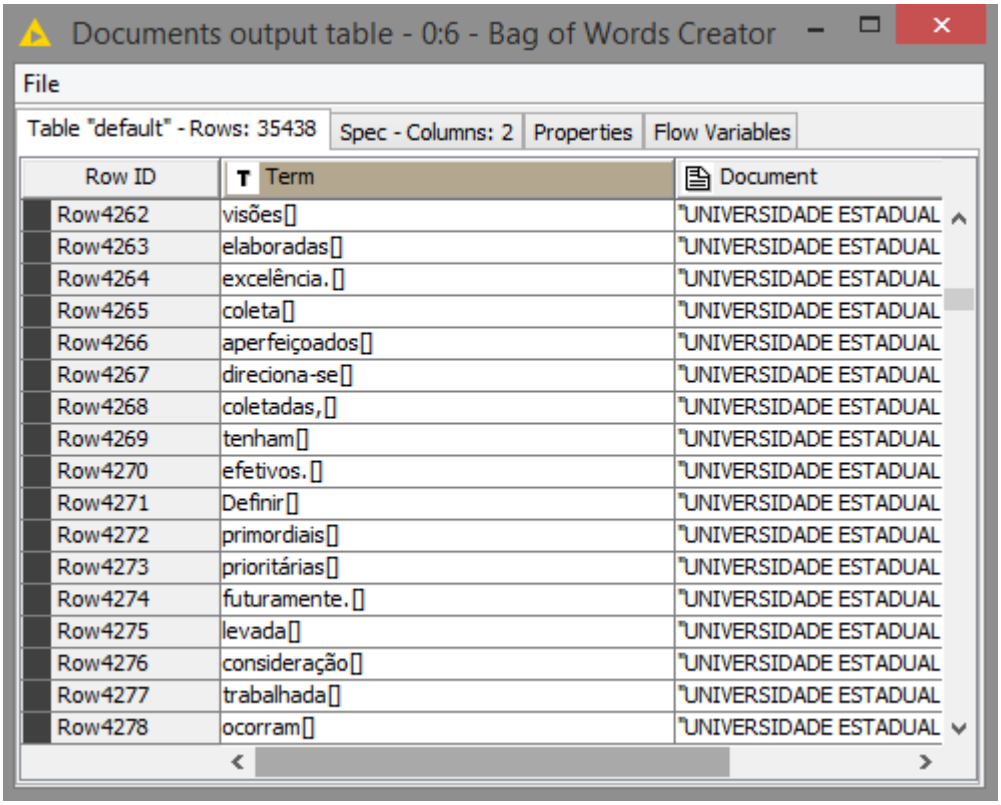
Escolhendo a visualização das 2000 primeiras linhas da tabela anterior, podemos observar que no caso dos trabalhos de Jornalismo, há 1473 sentenças de 17 a 34 palavras e 62 sentenças que possuem de 51 a 68 palavras. Já nos trabalhos de Relações Públicas, são 1332 sentenças de 17 a 34 palavras e 98 sentenças de 51 a 68 palavras. Podemos concluir, então, que nos trabalhos de Jornalismo predominam sentenças mais curtas, e nos de Relações Públicas as sentenças são geralmente mais longas.

Voltando à figura 14, podemos semelhantemente descrever o fluxo mais longo do *workflow* apresentado. Esse novo caminho é iniciado nos mesmos nódulos "*Word Parser*", mostrando que trabalham exatamente com os mesmos arquivos. Entretanto, agora, o fluxo é mais complexo.

A primeira etapa após a captação dos dados da fonte é o nódulo "*Bag of Words*". Diferentemente do nódulo "*Sentence Extractor*", ele cria uma tabela em que cada linha contém uma única palavra; por isso, o nome do nódulo se traduz em "bolsa de

palavras". Para demonstrar essa função, utilizando-se agora os textos do curso de Relações Públicas:

Figura 18 – Palavras Extraídas



Row ID	Term	Document
Row4262	visões[]	"UNIVERSIDADE ESTADUAL
Row4263	elaboradas[]	"UNIVERSIDADE ESTADUAL
Row4264	excelência.[]	"UNIVERSIDADE ESTADUAL
Row4265	coleta[]	"UNIVERSIDADE ESTADUAL
Row4266	aperfeiçoados[]	"UNIVERSIDADE ESTADUAL
Row4267	direciona-se[]	"UNIVERSIDADE ESTADUAL
Row4268	coletadas,[]	"UNIVERSIDADE ESTADUAL
Row4269	tenham[]	"UNIVERSIDADE ESTADUAL
Row4270	efetivos.[]	"UNIVERSIDADE ESTADUAL
Row4271	Definir[]	"UNIVERSIDADE ESTADUAL
Row4272	primordiais[]	"UNIVERSIDADE ESTADUAL
Row4273	prioritárias[]	"UNIVERSIDADE ESTADUAL
Row4274	futuramente.[]	"UNIVERSIDADE ESTADUAL
Row4275	levada[]	"UNIVERSIDADE ESTADUAL
Row4276	consideração[]	"UNIVERSIDADE ESTADUAL
Row4277	trabalhada[]	"UNIVERSIDADE ESTADUAL
Row4278	ocorram[]	"UNIVERSIDADE ESTADUAL

Fonte: Captura de tela feita pelo autor (2017)

Os nódulos que se seguem, contidos na área de cor amarela denominada "Pré-processamento" são responsáveis por "limpar" os dados de modo que algumas peculiaridades sejam removidas para facilitar o processamento no geral. Resumidamente, o nódulo "*Punctuation Erasure*" apaga os sinais e pontuações; o nódulo "*Case Converter*" converte todas as letras para caixa baixa; o nódulo "*Stop Word Filter*" filtra palavras comuns demais na língua selecionada (na língua portuguesa, palavras como artigos "o/os" e "a/as") e o nódulo "*Term Grouper"* que agrupa termos semelhantes em um único registro.

Após essa seção, a etapa do "Cálculo de Frequência" elabora os valores TF e IDF explicados no capítulo 2.2 deste trabalho através dos respectivos nódulos de mesmo

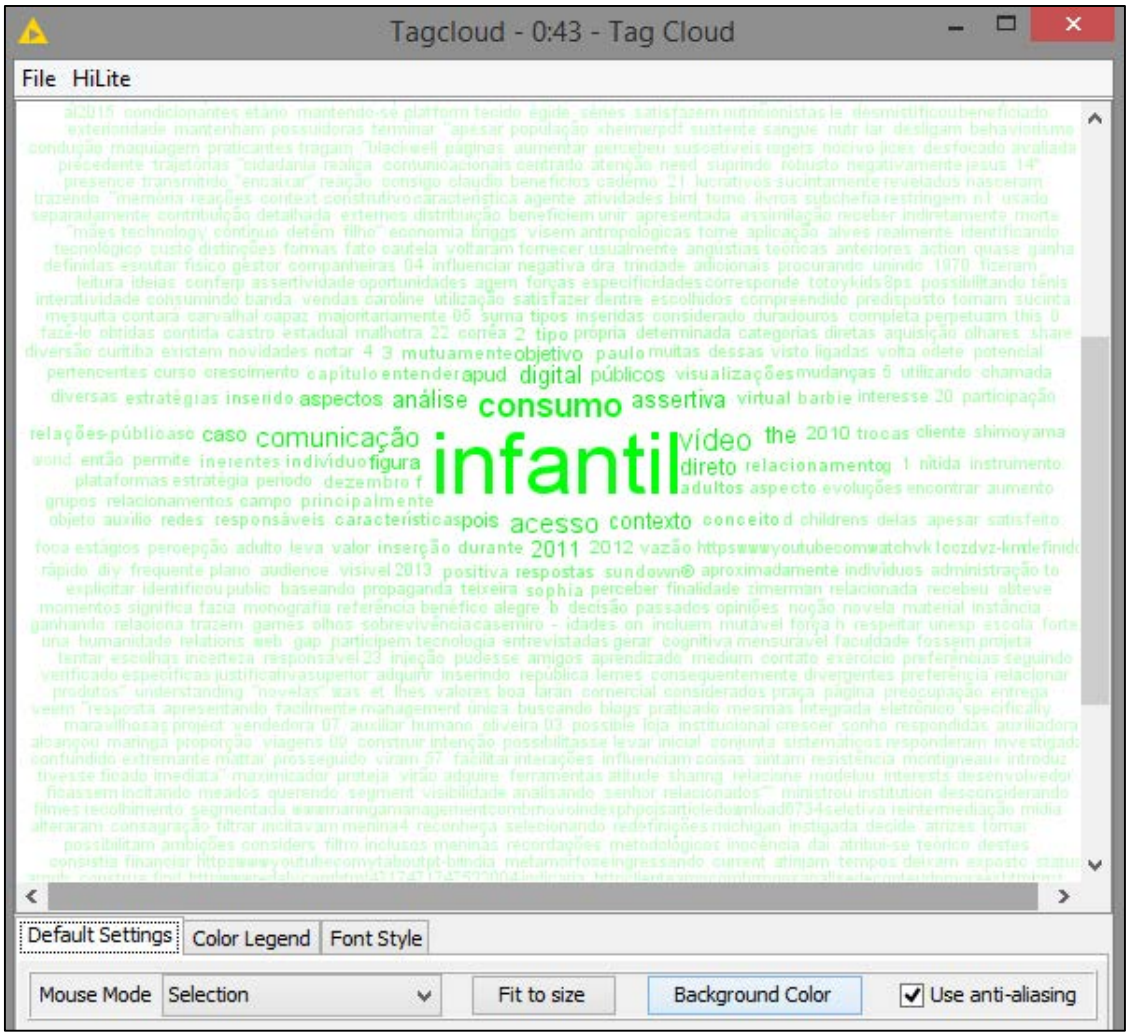
nome. O terceiro nódulo desta parte, "*Java Snippet"* contém um argumento em linguagem de programação *Java* para realizar o cruzamento desses valores e criar uma coluna com os valores da matriz "TF-IDF".

Finalmente, o último passo é a visualização dos resultados obtidos através do nódulo "*Tag Cloud*", que cria uma nuvem de palavras para demonstrar visualmente os valores dos termos que chegaram ao fim do fluxo. Temos que mais ao centro e maiores, então, estão os valores mais elevados dentro do espaço vetorial. No caso dos TCCs de Jornalismo, a palavra que se destaca é "discurso", competindo em tamanho com a palavra "intertextualidade". Já nos TCCs de Relações Públicas, a palavra em proeminência é "infantil", tendo um destaque muito mais significativo em comparação com outras palavras. A palavra "consumo" vem em segundo lugar, mas dessa vez com uma diferença mais expressiva em relação à palavra central.

Figura 19 – Nuvem de Palavras (Jornalismo)

Fonte: Captura de tela feita pelo autor (2017)

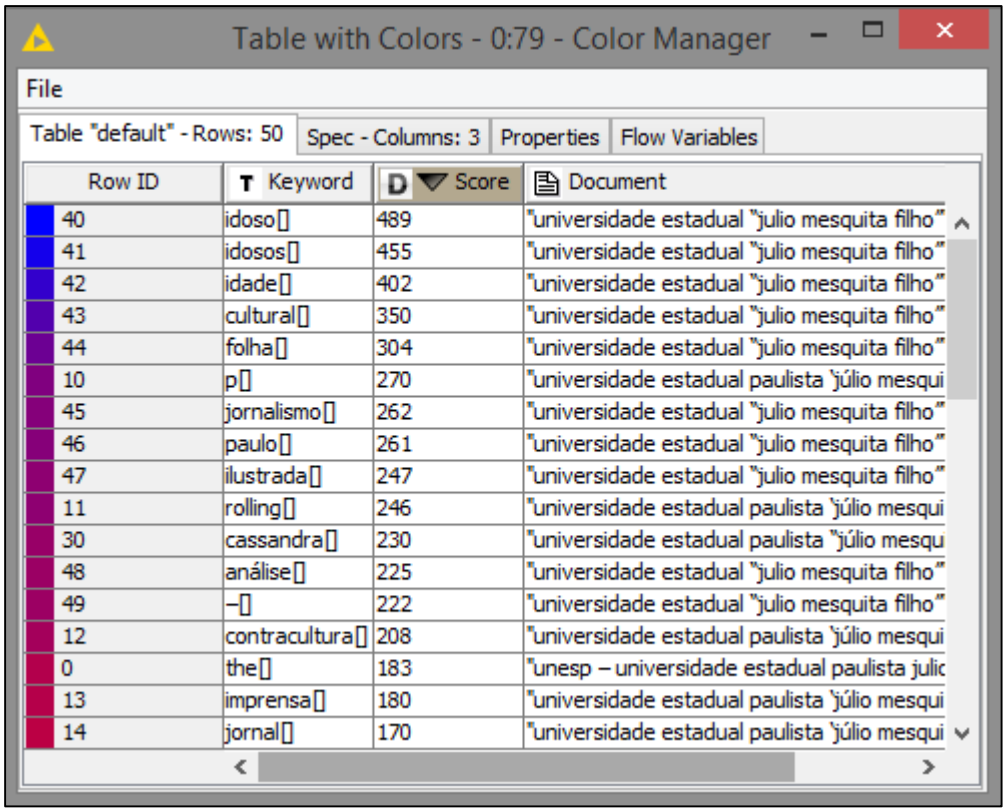
Figura 20 – Nuvem de Palavras (Relações Públicas)



Fonte: Captura de tela feita pelo autor (2017)

Ao final do *workflow* ainda foram adicionados os nódulos "*Keygraph Keyword Extractor*". Seu processo utiliza um outro tipo de técnica de análise baseada em gráficos para gerar uma lista de "palavras-chave" dentro do *corpus* estudado. Após ser processado, foi adicionado o nódulo "*Color Manager*" para colorir os resultados de acordo com seus valores, sendo configurado para valores maiores representados na cor azul, gradativamente se tornando vermelho até atingir o patamar mínimo. O resultado da visualização são tabelas coloridas mostradas nas figuras 21 e 22.

Figura 21 – Palavras-chave (Jornalismo)



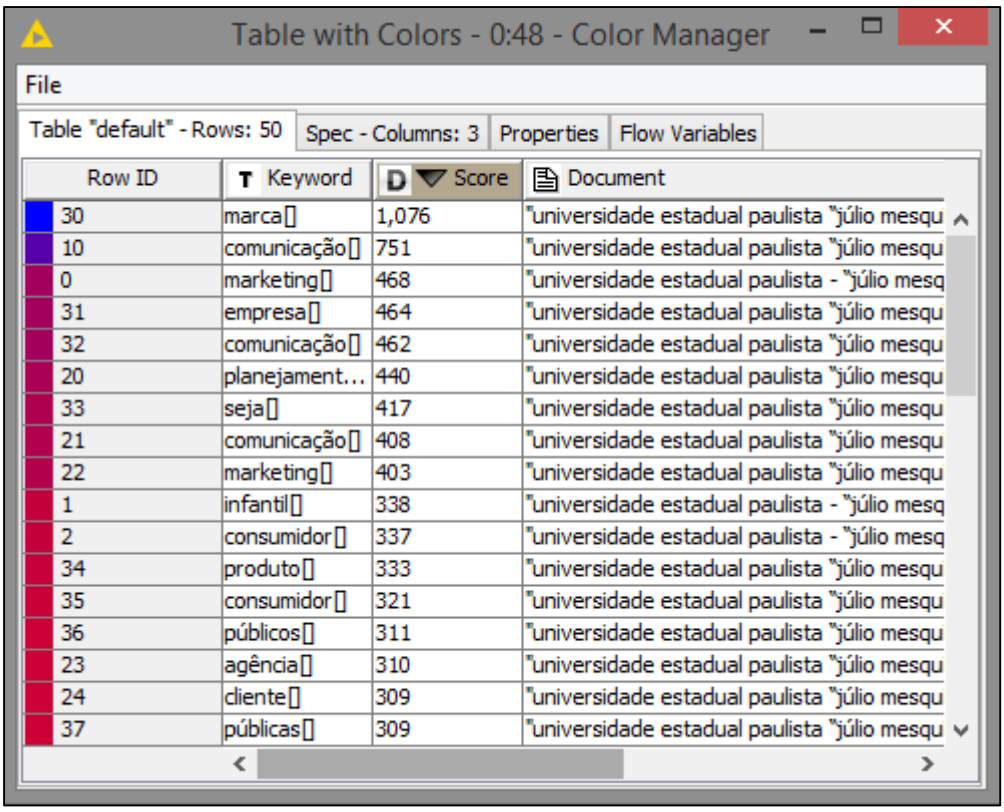
Row ID	Keyword	Score	Document
40	idoso[]	489	"universidade estadual "julio mesquita filho"
41	idosos[]	455	"universidade estadual "julio mesquita filho"
42	idade[]	402	"universidade estadual "julio mesquita filho"
43	cultural[]	350	"universidade estadual "julio mesquita filho"
44	folha[]	304	"universidade estadual "julio mesquita filho"
10	p[]	270	"universidade estadual paulista "júlio mesqui
45	jornalismo[]	262	"universidade estadual "julio mesquita filho"
46	paulo[]	261	"universidade estadual "julio mesquita filho"
47	ilustrada[]	247	"universidade estadual "julio mesquita filho"
11	rolling[]	246	"universidade estadual paulista "júlio mesqui
30	cassandra[]	230	"universidade estadual paulista "júlio mesqu
48	análise[]	225	"universidade estadual "julio mesquita filho"
49	–[]	222	"universidade estadual "julio mesquita filho"
12	contracultura[]	208	"universidade estadual paulista "júlio mesqui
0	the[]	183	"unesp – universidade estadual paulista julic
13	imprensa[]	180	"universidade estadual paulista "júlio mesqui
14	jornal[]	170	"universidade estadual paulista "júlio mesqui

Fonte: Captura de tela feita pelo autor (2017)

Os termos com valores mais altos no fluxo de Jornalismo são as palavras “idoso”, “idosos” e “idade”, podendo indicar a prevalência de um tema específico dentro dos textos abordados. Já em Relações Públicas, os termos chave são “marca”, “comunicação” e “marketing”, podendo indicar a prevalência de temas mais abrangentes.

Podemos observar ainda que na tabela de Jornalismo, as linhas de número 10 (termo “p”), número 49 (termo “–”) e número 0 (termo “the”) podem assinalar que os dados precisam ser filtrados com mais minúcia, pois não são termos relevantes para a análise. O primeiro não tem significado; o segundo é apenas um símbolo e o terceiro um termo da língua inglesa. Retomando os conceitos das características do *Big Data*, são “ruídos” que se mantiveram até o final do fluxo. Sendo assim, outros nódulos podem ser adicionados ou configurações alteradas para solucionar esses problemas.

Figura 22 – Palavras-chave (Relações Públicas)

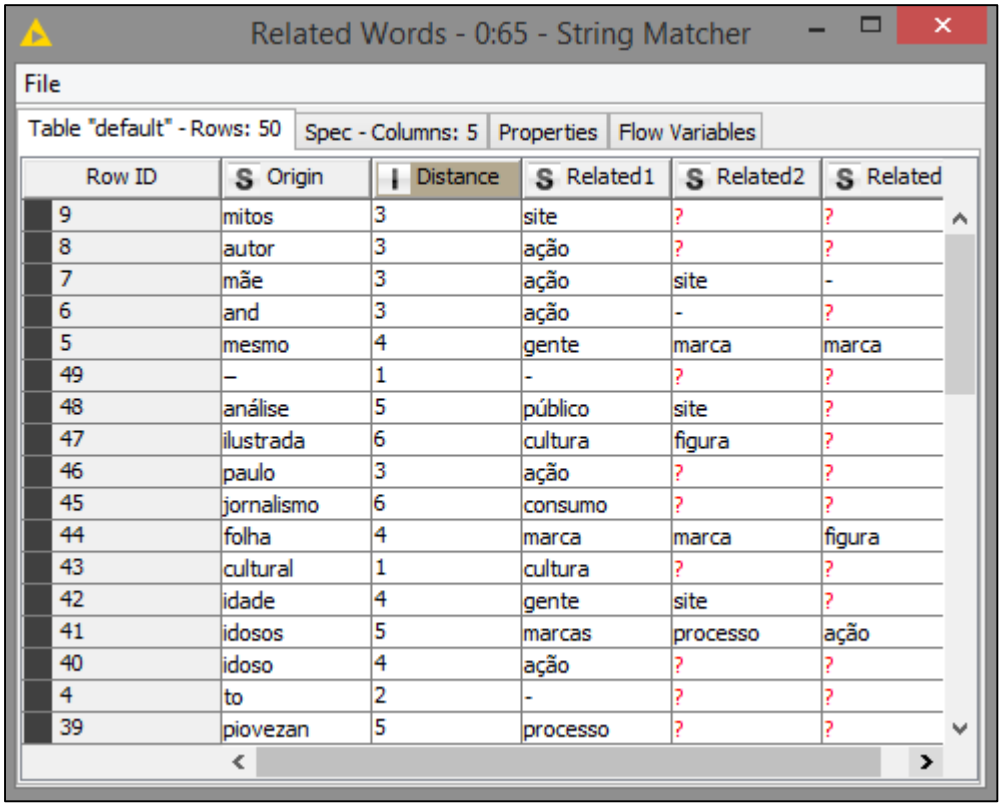


Row ID	Keyword	Score	Document
30	marca[]	1,076	"universidade estadual paulista "júlio mesqu
10	comunicação[]	751	"universidade estadual paulista "júlio mesqu
0	marketing[]	468	"universidade estadual paulista - "júlio mesq
31	empresa[]	464	"universidade estadual paulista "júlio mesqu
32	comunicação[]	462	"universidade estadual paulista "júlio mesqu
20	planejament...	440	"universidade estadual paulista "júlio mesqu
33	seja[]	417	"universidade estadual paulista "júlio mesqu
21	comunicação[]	408	"universidade estadual paulista "júlio mesqu
22	marketing[]	403	"universidade estadual paulista "júlio mesqu
1	infantil[]	338	"universidade estadual paulista - "júlio mesq
2	consumidor[]	337	"universidade estadual paulista - "júlio mesq
34	produto[]	333	"universidade estadual paulista "júlio mesqu
35	consumidor[]	321	"universidade estadual paulista "júlio mesqu
36	públicos[]	311	"universidade estadual paulista "júlio mesqu
23	agência[]	310	"universidade estadual paulista "júlio mesqu
24	cliente[]	309	"universidade estadual paulista "júlio mesqu
37	públicas[]	309	"universidade estadual paulista "júlio mesqu

Fonte: Captura de tela feita pelo autor (2017)

Por último, o nódulo "*String Matcher*" recebe as informações de ambos os nódulos "*Keygraph Keyword Extractor*" e utiliza um método para calcular a distância entre os valores *string* presentes nas duas tabelas e apresentar uma lista dos valores mais similares uns aos outros. É relevante notar que a similaridade das palavras relacionadas, nesse caso, diz respeito apenas aos caracteres individuais que cada palavra possui, não contendo nenhuma relação semântica.

Figura 23 – Palavras Relacionadas



Row ID	Origin	Distance	Related1	Related2	Related
9	mitos	3	site	?	?
8	autor	3	ação	?	?
7	mãe	3	ação	site	-
6	and	3	ação	-	?
5	mesmo	4	gente	marca	marca
49	–	1	-	?	?
48	análise	5	público	site	?
47	ilustrada	6	cultura	figura	?
46	paulo	3	ação	?	?
45	jornalismo	6	consumo	?	?
44	folha	4	marca	marca	figura
43	cultural	1	cultura	?	?
42	idade	4	gente	site	?
41	idosos	5	marcas	processo	ação
40	idoso	4	ação	?	?
4	to	2	-	?	?
39	piovezan	5	processo	?	?

Fonte: Captura de tela feita pelo autor (2017)

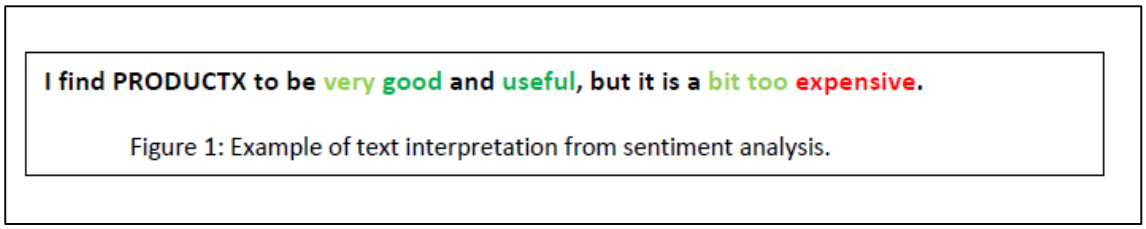
O *workflow* demonstrado foi capaz de apresentar alguns *insights* sobre o *corpus* de textos estudados, destacando algumas diferenças entre as duas categorias (Jornalismo e Relações Públicas). Por se tratar apenas de um exemplo prático de uso do *software Knime* em um computador pessoal, alguns adendos devem ser comentados.

O número total de arquivos (10) foi determinado arbitrariamente, e apesar de ser um número relativamente pequeno, os processos demoraram um tempo considerável para serem concluídos (cerca de 4 minutos para o processamento e resolução completa de todos os nódulos). Naturalmente, espera-se que o tempo para a conclusão do fluxo seja maior em computadores com *hardware* de menor capacidade ou mais antigos. Entretanto, como apontado brevemente na introdução deste trabalho, as capacidades e infraestruturas de *hardware* necessárias para processamentos computacionais estão cada vez mais acessíveis devido aos serviços e tecnologias em "nuvem". Dessa maneira, o foco técnico nas capacidades do programa se mantém, visto que as análises e resultados

podem ser reproduzidos de maneira idêntica em um ambiente de *Big Data* e com um número muito maior de documentos e dados.

O número e tipos de nódulos utilizados nessa análise também representam uma parcela pequena do que está disponível no software. É recomendado, então, que o usuário se familiarize com outros tipos de dados e outras transformações possíveis em suas configurações para descobrir e aprender com solidez suas capacidades e funcionalidades. É possível, por exemplo, dar à cada palavra uma cor representando valores semânticos determinados pelo usuário, como "negativo", "neutro" e "positivo", demonstrado no exemplo da figura 24.

Figura 24 – Exemplo de Análise de Sentimento



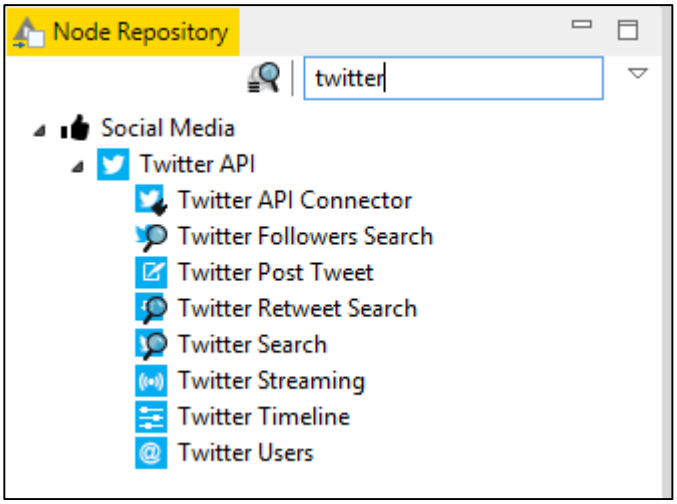
Fonte: THIEL et al., 2012, p.4

No caso, os termos em inglês "very" (muito) e "bit too" (um pouco) estão em verde-claro, apontando que tendem ao positivo. Já "good" (bom) e "useful" (útil) em verde-escuro, indicam que são positivos de fato. O termo "expensive" (caro) está em vermelho, indicando que é negativo. Os outros termos não contém uma coloração específica, e podemos inferir que são neutros. Esse tipo de análise, também conhecida como Análise de Sentimento, não pôde ser feita no *workflow* descrito nesse trabalho, uma vez que o programa *Knime* ainda não possui uma biblioteca de termos que contenha esse tipo de classificação para as palavras de língua portuguesa. Porém, essa capacidade está disponível para textos em inglês, alemão e espanhol.

No caso de análises de texto, é interessante para o profissional de Jornalismo saber que apesar de demandar um conhecimento um pouco mais complexo do *software*, é possível fazer o cruzamento de dados de textos de suas próprias fontes com dados da rede social *Twitter* e outras plataformas. O *Knime* possui um "arsenal" de nódulos especificamente desenvolvidos para tal. Assim, pode-se entender, por exemplo, a

proximidade de assuntos tratados num *corpus* determinado com os principais tópicos tratados na rede social em tempo real, aumentando ainda mais o valor dessa ferramenta.

Figura 25 – Nódulos voltados para o *Twitter*



Fonte: Captura de tela feita pelo autor (2017)

Assim, a variedade de nódulos e a crescente integração dos mesmos com outros serviços, aplicativos e *websites* tornam o *Knime* uma poderosa ferramenta para lidar com um amplo espectro de dados de uma maneira simples, intuitiva e efetiva.

4 CONSIDERAÇÕES FINAIS

O conhecimento de ferramentas de *Big Data* é indubitavelmente valioso, sobretudo voltado para objetivos mercadológicos. Stone (2014) aponta que em termos de estratégia de mercado, empresas que utilizam o *Big Data* tomam decisões cinco vezes mais rápido que seus concorrentes, são três vezes mais estimuladas à executar decisões como previsto, possuem duas vezes mais chances ser uma das líderes em seu ramo em termos financeiros e são duas vezes mais inclinadas a usar dados frequentemente antes de tomar decisões (p.2). Por esse motivo, segundo a firma de consultoria *New Vantage Partners*, a parcela das empresas mais valiosas do mundo que possuem em sua hierarquia o cargo de CDO ou *chief data officer* (diretor de dados) subiu de 12% em 2012 para 54% em 2016, ressaltando a importância do gerenciamento de aplicações de análises de dados.

No caso das empresas de mídia, para alcançar um estágio de maior aproveitamento desse contexto é preciso abraçar essas oportunidades, especialmente para "igualar" as porções tradicionais dessa indústria às novas empresas (STONE, 2014, p.2) que já nasceram em um ambiente altamente tecnológico, como as redes sociais. O Jornalismo fora das redes sempre foi um grande gerador de dados, compilando as informações do mundo e mantendo registros valiosos da história. Dentro da era da *Internet*, *Internet* das coisas, redes sociais e *Big Data*, não é diferente. O Jornalismo consome e cria informações, que cada vez mais podem ser organizadas, compreendidas e utilizadas em forma de dados.

Cientes disso, podemos sugerir que é fundamental para o jornalista profissional ou estudante que conheça, no mínimo, como funciona sua ferramenta mais básica de trabalho – o texto – dentro do mundo dos dados. Também, é igualmente ideal que explore novas possibilidades e descubra como fazer uso de outros tipos de dados para ampliar seus conhecimentos nesse campo. Profissionais que não mais só recebem relatórios e resultados de análises de empresas especializadas, mas sabem navegar de maneira independente nos variados fluxos de dados a que temos acesso, manipulando-os e criando suas próprias análises tornam-se indivíduos muito mais preparados para crescer profissionalmente dentro da realidade atual.

À medida que a indústria de *Big Data* avança, cresce também a necessidade de profissionais capacitados. O software *Knime* contém várias facilidades e recursos para que essa capacitação possa ser iniciada do "marco zero", principalmente tendo em vista que a área de Jornalismo tradicionalmente não possui ligações diretas com os campos das Ciências da Informação ou Sistemas de Informação.

Portanto, o uso dos conhecimentos adquiridos com o *Big Data* destaca o indivíduo e, de maneira geral, acentua o valor da própria profissão ao público, pois esse recebe em contrapartida conteúdos e produtos mais estimulantes, complexos e especializados, dignos de um profissional de Jornalismo. Seja na esfera profissional ou individual, saber entender o mundo de maneira objetiva e eficiente é um conhecimento seguramente enriquecedor.

5 REFERÊNCIAS BIBLIOGRÁFICAS

ASSOCIATED PRESS. **Data Journalism**. 2017. Disponível em: <https://www.apstylebook.com/ap_stylebook/data-journalism> Acesso em: 22 jul. 2017.

CISCO. **Cisco Global Cloud Index: Forecast and Methodology, 2015-2020**. Cisco Public, 2016. Disponível em: <www.cisco.com/c/dam/en/us/solutions/collateral/.../white-paper-c11-738085.pdf> Acesso em: 14 jul. 2017.

CONOVER, Michael; GONÇALVES, Bruno; RATKIEWICZ, Jacob; FLAMMINI, Alessandro; MENCZER, Filippo. **Political Polarization on Twitter**. Association for the Advancement of Artificial Intelligence, 2011. Disponível em: <https://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/viewFile/2847/3275> Acesso em: 19 jul. 2017.

EMC. **Rich Data & the Increasing Value of the Internet of Things**. EMC Digital Universe Infobrief, 2014. Disponível em: <https://www.emc.com/collateral/analyst-reports/idc-digital-universe-2014.pdf> Acesso em: 11 jul. 2017.

FERNEDA, Edberto; LOPES, Tatiane. **O modelo Espaço Vetorial no desenvolvimento de interfaces de busca e recuperação de informação**. Sem data. Disponível em: <https://www.marilia.unesp.br/Home/Eventos/2015/seminariodearquivologiaebibliotecnomia/lopes-t.s.f.-ferneda-e..pdf> Acesso em: 19 jul. 2017.

KNIME. **About Knime**. Sem data. Disponível em: <https://www.knime.org/about> Acesso em: 2 jul. 2017.

LEBOEUF, Kelly. **2016 Update: What happens in one internet minute?** 2016. Disponível em: <http://www.excelacom.com/resources/blog/2016-update-what-happens-in-one-internet-minute> Acesso em: 16 jul. 2017.

LÓSCIO, Bernadette; OLIVEIRA, Hélio de; PONTES, Jonas. **NoSQL no desenvolvimento de aplicações Web colaborativas.** Sem data. Disponível em: <www.addlabs.uff.br/sbsc_site/SBSC2011_NoSQL.pdf> Acesso em: 16 jul. 2017.

MAYIKA, James; CHUI, Michael; BROWN, Brad; Bughin, JACQUES; DOBBS, Richard; ROXBURGH, Charles; BYERS, Angela Hung. **Big Data: The next frontier for innovation, competition, and productivity**. McKinsey Global Institute, 2011. Disponível em: <http://www.mckinsey.com/~/media/McKinsey/Business%20Functions/McKinsey%20Digital/Our%20Insights/Big%20data%20The%20next%20frontier%20for%20innovation/MGI_big_data_full_report.ashx> Acesso em: 3 jul. 2017.

MAYIKA, James; CHUI, Michael; Bughin, JACQUES; DOBBS, Richard; BISSON, Peter; MARRS, Alex. **Disruptive technologies: Advances that will transform life, business, and the global economy**. McKinsey Global Institute, 2013 Disponível em: <http://www.mckinsey.com/~/media/McKinsey/Business%20Functions/McKinsey%20Digital/Our%20Insights/Disruptive%20technologies/MGI_Disruptive_technologies_Full_report_May2013.ashx> Acesso em: 3 jul. 2017.

NEWVANTAGE PARTNERS. **Big Data Executive Survey 2016: An Update on the Adoption of Big Data in the Fortune 1000**. NewVantage Partners LLC, 2016. Disponível em: <http://newvantage.com/wp-content/uploads/2016/01/Big-Data-Executive-Survey-2016-Findings-FINAL.pdf> Acesso em: 22 jul. 2017.

ORACLE. **Media for Long-Term Archiving.** 2014. Disponível em: <storageconference.us/2014/Presentations/Panel1.Lampitt.pdf> Acesso em: 23 jul. 2017.

SINT, Rolf; SCHAFFERT, Sebastian; STROKA; Stephanie; FERSTL, Roland. **Combining Unstructured, Fully Structured and Semi-Structured Information in Semantic Wikis.** Sem data. Disponível em: <ceur-ws.org/Vol-464/paper-14.pdf> Acesso em: 15 jul. 2017.

STONE, Martha. **Big Data for Media**. Reuters Institute for the Study of Journalism, 2014. Disponível em: <https://reutersinstitute.politics.ox.ac.uk/sites/.../Big%20Data%20For%20Media_0.pdf > Acesso em: 20 jun. 2017.

TABLEAU. **Top ten Big Data trends for 2017.** 2017. Disponível em: <https://www.tableau.com/sites/default/files/media/Whitepapers/whitepaper_top_10_big_data_trends_2017.pdf > Acesso em: 06 jul. 2017.

THIEL, Killian; KÖTTER, Tobias; BERTHOLD, Michael; SILIPO, Rosaria; WINTERS, Phil. **Creating Usable Customer Intelligence from Social Media Data: Network Analytics meets Text Mining**. 2012. Disponível em: <https://www.knime.org/files/knime_social_media_white_paper.pdf> Acesso em: 9 jul. 2017.