Mineração de textos para agrupamento de teses e dissertações por meio de análise de similaridade

Renata Moreira Limiro
Mestranda em Engenharia de Produção pela Universidade Federal de Goiás (UFG). Supervisora de Inteligência de Dados na Implanta.
E-mail: renatamlimiro@gmail.com

Núbia Rosa da Silva
Doutora em Ciências da Computação e Matemática Computacional pela Universidade de São Paulo (USP). Docente do Departamento de Ciência da Computação da Universidade Federal de Goiás (UFG).
E-mail: nubia@ufcat.edu.br

Douglas Farias Cordeiro
Doutor em Ciências da Computação e Matemática Computacional pela Universidade de São Paulo (USP). Docente da Faculdade de Comunicação e Informação da Universidade Federal de Goiás (UFG).
E-mail: cordeiro@ufg.br

RESUMO

A construção de redes de conhecimento é um dos grandes desafios da pesquisa científica e inovação, no tocante ao uso e processamento da informação. Produzir conhecimento em redes de pesquisa e colaboração é importante para o entendimento e internacionalização das investigações. A utilização de métodos que possam revelar áreas de conhecimento implicitamente relacionadas é uma interessante alternativa. Entretanto, a realização deste tipo de construção requer a aplicação de métodos e técnicas específicos, os quais possam, a partir de amostras de dados, gerar informação que auxilie tal tarefa. Este artigo tem como objetivo apresentar a aplicação dos métodos de mineração de dados Doc2Vec e classificação de Reinert para a inferência de redes de conhecimento com base na similaridade e agrupamento de tópicos de documentos científicos textuais. O desenvolvimento do trabalho é baseado na metodologia KDT (do inglês, *Knowledge Discovery from Texts*). Diante disso, foram obtidas, utilizando técnicas de Web Scraping, dados de teses e dissertações disponibilizados na Biblioteca Digital Brasileira de Teses e Dissertações do IBICT, os quais foram tratados e submetidos à rotinas de processamento de mineração textual. São apresentados resultados quanto à classificação das amostras em grupos por similaridade semântica e grafos que representam a relação entre tais grupos.

Palavras–chave: Redes de conhecimento. Mineração de texto. Doc2Vec.

ABSTRACT

The construction of knowledge networks is one of the great challenges in the scope of the advances in research and innovation. The use of methods that can help in understanding implicitly related areas is an interesting alternative. However, the realization of this type of construction requires the application of specific methods and techniques, which can, from data samples, generate information to assist this task. This article presents the application of Doc2Vec data mining methods and Reinert classification for the construction of similarity and grouping networks based on topics. The development of this work is based on the KDT (Knowledge Discovery from Texts), which predicts a sequence of activities for the process of generating knowledge from textual data. Therefore, data from theses and dissertations available at the Brazilian Digital Library of Theses and Dissertations of IBICT were obtained using Web Scraping techniques, which were treated and submitted to textual mining processing routines. Through the application of techniques of text mining, results are presented regarding the classification of samples into groups by semantic similarity and graphs that represent

the relationship between these groups. The results allow the development of analyzes aimed at surveying knowledge networks, providing new possibilities for research and innovation.

Keywords: Knowledge networks. Text mining. Doc2Vec.

1 INTRODUÇÃO

A sociedade está cada vez mais envolvida com o termo redes, e é comum encontrar no discurso da mídia citações variadas de redes de negócios, redes de relacionamentos sociais, rede neural, rede digital, rede de pesquisadores e, até mesmo, rede de terroristas. Este cenário acaba emergindo-se de uma sociedade em rede, ainda que não haja consciência disso. Neste contexto, uma rede social pode ser considerada uma estrutura social, composta por pessoas ou organizações, conectadas por uma ou vários tipos de relações que compartilham valores e objetivos comuns. Redes sociais interconectam os seres humanos desde o início da espécie humana, sempre agindo como agente da circulação de informação. Por outro lado, as redes de conhecimento são redes com propósito de criar e disseminar conhecimento, geralmente constituída por instituições de pesquisa, ONG's e agências governamentais. O conceito de redes pode ser um instrumento importante para auxiliar na compreensão dos processos de interação institucional e de geração do conhecimento.

Medieta e Schmidt (2002) conceituam que uma rede é composta de atores e relações entre eles, onde atores são indivíduos, grupos ou entidades naturais ou sociais, e relações podem ser analisadas quanto a sua quantidade e qualidade, permitindo identificar padrões de vinculação, podendo ser simétricas ou assimétricas; diretas ou indiretas; horizontais ou hierárquicas; recíprocas, intensas, duráveis entre outras. De um modo geral, para Figueiredo (2011) uma rede pode ser descrita como uma representação entre pares de elementos de um mesmo conjunto.

Diferentes situações do convívio humano podem ser representadas por meio de redes, o que pode acabar gerando diversos questionamentos acerca do estudo de redes. A importância está em saber o quão próximos os atores estão entre si, que interações têm cada ator, saber se todos os atores que estão na rede têm o mesmo papel para realização de determinado fenômeno, e que relações possuem entre si (VIEIRA; CORDEIRO, 2019). Questões essas que podem ser moldadas e respondidas através do estudo das redes utilizando a modelagem de redes complexas (CORDEIRO et al., 2022).

As redes podem representar diversas situações reais, desde relações de amizade até mesmo propagação de doenças. Um exemplo de rede são as comunidades virtuais, tais como Facebook, LinkedIn, Twitter, entre outras, onde cada usuário pode ser definido como um vértice e a relação com os demais usuários, as arestas. Outro ponto de considerável importância no que se refere à representação através de redes são os relacionamentos entre autores no âmbito de colaboração e coautoria. Neste sentido, através da exploração de redes, podem ser levantados dados e informações relacionados, por exemplo, às áreas de interesse de pesquisa (SILVA; CASSIANO; CORDEIRO, 2019).

É interessante pontuar que, conforme descrito por Silva (2002), o desenvolvimento científico é algo colaborativo, que requer o estabelecimento de mecanismos e estratégias que se baseiem no compartilhamento e associação entre pares, visando uma potencialização da produção de conhecimento. Além disso, também é notável a valorização por parte de agências de fomento à pesquisa científica, principalmente no âmbito do incentivo à trabalhos compartilhados buscando uma economia de recursos, materiais e tempo, o que está diretamente relacionado à tecnologia (MAIA; CAREGNATO, 2008).

> O desenvolvimento tecnológico dinamizou a interação entre os cientistas e contribuiu por sua vez, para o alto número de publicações com colaborações científicas, possibilitando uma comunicação mais dinâmica e interativa, como exemplo destacam-se as redes sociais para pesquisadores *researchgate* e *academia.edu*. Essas por sua vez, em muitos casos exercem o primeiro contato entre os pesquisadores, em que a divulgação de artigos de periódicos, resumos expandidos, discussões sem o aval do rigor científico, circulam nas redes sociais, cita-se ainda os Blogs; lista de discussão; grupos de Facebook e Twitter. Dessa forma, as mídias sociais configuram-se como um importante canal de comunicação. (FIRME; MIRANDA; SILVA, 2017).

As influências e interconexões alcançadas através dos inúmeros avanços tecnológicos acabam por simplificar a construção de intersecções entre pesquisadores e grupos de pesquisa de diferentes áreas, e geograficamente distantes, o que é um fator de fundamental relevância no fortalecimento da coautoria científica. De acordo com Santin, Vanz e Caregnato (2018), as últimas décadas apresentaram um crescimento consideravelmente significativo na colaboração científica, juntamente com um aumento quantitativo de pesquisadores e instituições de pesquisas. Os autores apresentam um estudo analítico sobre redes de colaboração científica a partir de indicadores

bibliométricos, onde destacam a realização de uma série de estudos empíricos, baseados em coautoria, produtividade e impacto, que demonstram o crescimento de colaborações (PRICE, 1976; PERSSON; GLÄNZEL; DANELL, 2004; ADAMS, 2013; LARIVIÈRE et al., 2015), assim como no que se refere ao aumento da produtividade em face da colaboração (BEAVER; ROSEN, 1979; PERSSON; GLÄNZEL; DANELL, 2004).

Esse cenário se reflete em uma crescente discussão sobre as abordagens metodológicas a serem aplicadas em termos da mensuração da colaboração científica. Uma estratégia é a aplicação de indicadores bibliométricos, os quais, conforme descrito por Price (1976) e Glänzel (2002), proporcionam resultados quantitativos em termos do desempenho científico, inclusive levando-se em conta séries temporais. Santin, Vanz e Caregnato (2018) destacam que através da exploração de indicadores bibliométricos, é possível se mensurar a produção científica colaborativa, assim como realizar o mapeamento de cooperações existentes.

Para além do uso da bibliometria como mecanismo de mensuração de redes de cooperação, a mineração de dados, por meio da mineração de textos, e a inteligência artificial podem ser utilizadas como estratégias para geração de informações implícitas em dados de pesquisadores, neste caso, não volta para a mensuração de redes de colaboração ou indicadores estatísticos, mas sob um viés de descoberta de conhecimento, com intuito de inferir possibilidades de cooperação a partir de padrões implícitos nos dados.

Para tanto, um dos desafios se refere a como coletar dados que sejam passíveis de processamento e permitam a realização de análises computacionais que proporcionem resultados relativos às redes de cooperação implícitas em pesquisas. Neste sentido, uma alternativa é a utilização de documentos textuais acadêmicos, tais como artigos, teses e dissertações, os quais, por estarem vinculados aos seus autores, através da realização de processos de mineração de dados, podem servir de base para a geração de redes de conhecimento.

Este artigo tem como objetivo apresentar uma proposta de construção de redes de similaridade e de agrupamento por tópicos a partir de um conjunto amostral de documentos textuais compostos por teses e dissertações de uma Instituição Federal de Ensino Superior (IFES), disponibilizadas através da Biblioteca Digital de Teses e Dissertações (BDTD) do Instituto Brasileiro de Informação em Ciência e Tecnologia

(Ibict). Na proposta é utilizada a técnica Doc2Vec para construção da matriz de similaridades, e do método de classificação de Reinert para determinação das classes.

2 MINERAÇÃO DE TEXTOS

A crescente evolução das técnicas de mineração de textos trouxe melhorias e avanços relevantes nesta área do conhecimento, os quais proporcionaram a automatização de rotinas e processos que antes eram realizados com maior esforço humano, o que consequentemente demandava maior tempo. De maneira geral, a mineração de textos pode ser entendida como uma subárea da Recuperação da Informação (RI) (SALTON; MCGILL, 1983), na qual, através de um conjunto de rotinas de processamento e análise de padrões, a informação é recuperada a partir de dados textuais, gerando, consequentemente conhecimento. Assim, destaca-se que a fundamentação dessa área está ligada às definições de dado, informação e conhecimento.

Buscando um melhor alinhamento sobre os conceitos relacionados à mineração de textos, é interessante pontuar as suas definições. De acordo com Silva, Peres e Boscarioli (2016), dado pode ser descrito como algo bruto, sem contexto, ou seja, um símbolo ou um conjunto de símbolos quantificados ou quantificáveis. Por outro lado, a informação pode ser descrita como dados tratados, os quais possuem significado. Deve-se observar que nem toda informação gerada é necessariamente útil e utilizada, e que nem todo dado processado é garantia de informação. Por fim, conhecimento pode ser definido como a informação explorada com algum propósito específico, ou seja, utilizada para, por exemplo, tomada de decisão, construção de cenários, entre outros.

Nesse sentido, pode-se entender dados como matéria prima indispensável para análise. Ainda de acordo com Silva, Peres e Boscarioli (2016), os dados podem ser classificados de duas formas: estruturados e não-estruturados. A identificação do tipo de dado é essencial para que o processo de mineração possa ser aplicado, uma vez que as peculiaridades de cada tipo de dado demandam rotinas específicas para seu processamento. De forma geral, os dados estruturados são aqueles que se referem ao resultado de transações, ou ainda, de medição ou observação, podendo ser armazenados em uma tabela, ou em um formato que siga um padrão pré-definido, facilmente compreensível por máquina. Enquanto os dados não-estruturados referem-se àqueles

que não apresentam padrões pré-definidos, sendo necessária a aplicação de rotinas para o seu tratamento e processamento.

Na perspectiva de dados estruturados, a obtenção de informação e geração de conhecimento podem ser alcançadas através do modelo KDD (do inglês, *Knowledge Discovery in Databases*) (FAYYAD; SHAPIRO-PIATETSKY; SCHMIDT, 1996). De modo geral, o KDD se refere a um conjunto de processos que vão desde a definição do problema até a geração dos resultados em si, ou seja, a geração de informação e conhecimento. Estes processos podem ser descritos como: Seleção, Pré-processamento, Transformação, Mineração de Dados, e Avaliação. Dentre estes, a Mineração de Dados destaca-se como a etapa mais relevante na obtenção dos resultados e pode ser definida como uma área multidisciplinar que, através de rotinas automatizadas, proporciona o reconhecimento de padrões, o levantamento de estatísticas, a extração, e a visualização de informações em grandes conjuntos de dados (OLIVEIRA et al., 2021). Entretanto, devido às particularidades inerentes às bases de dados textuais, é necessário o emprego de técnicas e modelos específicos, proveniente da subárea denominada Mineração de Textos.

A Mineração de Textos pode ser definida como um processo baseado na utilização de processos computacionais, para extração de padrões e conhecimento sobre conjuntos de dados textuais não-estruturados (LOH, 2001). Observa-se que alguns autores consideram que a Mineração de Textos pode ser definida como a aplicação do modelo KDD sobre dados textuais (PROVOST; FAWCETT, 2016), porém é importante pontuar que a mineração sobre este tipo de dado demanda técnicas que vão além dos processos tradicionais do KDD, compondo o que é denominado de KDT (do inglês, *Knowledge Discovery from Text*).

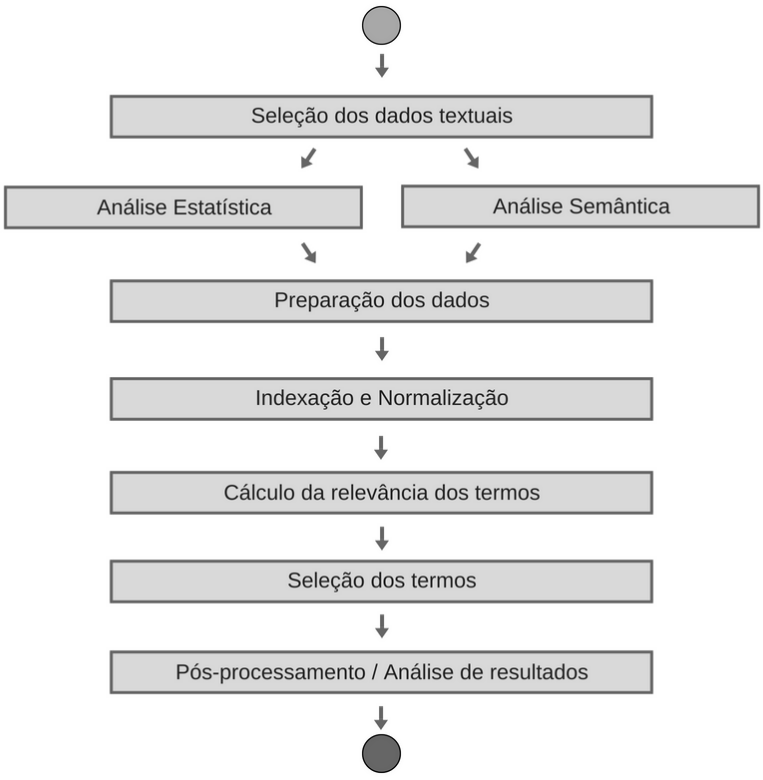
Os avanços relacionados ao KDT incluem contribuições que vão desde à exploração analítica em grandes bases de dados textuais, quantitativamente e qualitativamente, até a busca de informações em documentos, representadas através, por exemplo, da identificação de quais são os termos mais relevantes de um documento e como estes se relacionam, análises de conteúdo considerando o contexto, etc. Alguns exemplos de aplicação de KDT são, a análise de sentimentos em textos curtos aplicada no âmbito das mídias sociais (SILVA, 2016), e ainda a exploração da mineração de textos para classificação de documentos (HUSSEIN; ALAAELDIN; HASSAN, 2015).

Neste sentido, é possível inferir que as técnicas advindas do KDT podem trazer vantagens e benefícios tanto para problemas relacionados a volumes de dados extraídos

em documentos eletrônicos da Internet, quanto em outros variados tipos de cenários, como, por exemplo, em documentos gerados por sistemas de informações gerenciais, ou similares. De acordo com Beppler e Fernandes (2005), a maior parte dos dados de empresas estão contidos em documentos textuais, o que também potencializa a importância do KDT na geração de informação e conhecimento. De maneira prática, pode-se descrever o KDT, através da Mineração de Textos, como um conjunto de processos que auxiliam na descoberta de conhecimento, ou seja, a realização de análises que transformam dados em informação, as quais devem então ser verificadas, analisadas, levando em conta o contexto, dentro de seus propósitos.

A Figura 1 apresenta as etapas do processo de Mineração de Textos. É importante destacar que o desenvolvimento da análise textual pode seguir, de forma geral, duas abordagens distintas: a Análise Estatística e Análise Semântica. A Análise Estatística trata aspectos mais relacionados à quantificação dos termos na base de dados, incluindo, por exemplo, estimativas, codificação e modelos de representação (MORAIS; AMBRÓSIO, 2007). A Análise Semântica, por outro lado, explora aspectos mais ligados à representatividade de um termo em relação a outros, tendo sua base em PLN (Processamento de Linguagem Natural).

Figura 1 - Processo de mineração textual.



Fonte: baseado em Hussein, Alaaeldin e Hassan (2015).

2.1 CLASSIFICAÇÃO E AGRUPAMENTO DE DOCUMENTOS TEXTUAIS

Entre as aplicações possíveis por meio da Mineração de Textos usando Análise Semântica, está a Classificação de Textos. De acordo com Jurafsky e Martin (2009), a classificação de documentos textuais refere-se a identificação de um determinado documento d com relação a um conjunto de classes *C*, onde $C = c_1, c_2, c_3, ..., c_n$, em outras palavras, trata da determinação de qual classe c_i que o documento d pertence. Este tipo de aplicação pode ser utilizado em diversos contextos, tais como, definição de categorias, classificação de projetos, identificação de áreas de pesquisa, entre outros.

De acordo com Andrade (2015), os modelos de classificação de documentos textuais podem ser classificados de duas maneiras: *single-label*, onde cada documento está associado a apenas uma classe, ou *multi-label*, onde um documento pode estar associado a uma ou mais classes. Assim sendo, uma das alternativas para o processo de classificação é a utilização de algoritmos de aprendizagem baseados em dados previamente rotulados, o que é denominado de aprendizado supervisionado. Alguns exemplos de classificadores são: Naives Bayes, *Random Forest*, Árvores de Decisão e SVM (do inglês, *Support Vector Machine*).

No entanto, em problemas onde os dados não são pré-rotulados, uma solução é a utilização de modelos não-supervisionados, como o caso do presente trabalho. De acordo com Castro e Ferrari (2016), o agrupamento se refere a uma técnica de segmentação de dados com base na proximidade de padrões e tendências.

O agrupamento é referenciado como um dos estudos iniciais realizados em processos de mineração, através dela é possível realizar análises exploratórias com identificações de padrões e similaridades que possibilitam a identificação grupos ou classes, os quais podem ser utilizados como entrada para outros métodos. No presente trabalho será explorada a técnica *Paragraph Vector* para levantamento de características relacionadas à similaridade de documentos textuais para propósitos de agrupamento.

2.2 PARAGRAPH VECTOR (DOC2VEC)

A técnica conhecida como *Paragraph Vector*, referenciada pela sigla Doc2Vec, proposta por Mikolov et al. (2013) é uma generalização do método Word2Vec (MIKOLOV et al., 2013). O Doc2Vec se refere a uma ferramenta de Processamento de Linguagem

Natural, a qual tem por objetivo principal a representação de documentos, sendo descrita como um modelo de aprendizado não-supervisionado, baseado em representações vetoriais distribuídas dos termos ou palavras de um texto. A técnica permite que os textos referentes ao corpus textual posam ser de tamanho variável, de sentenças a documentos completos. A partir disso, os vetores são treinados para predizer palavras ou termos em um parágrafo e assim atribuir uma representação semântica.

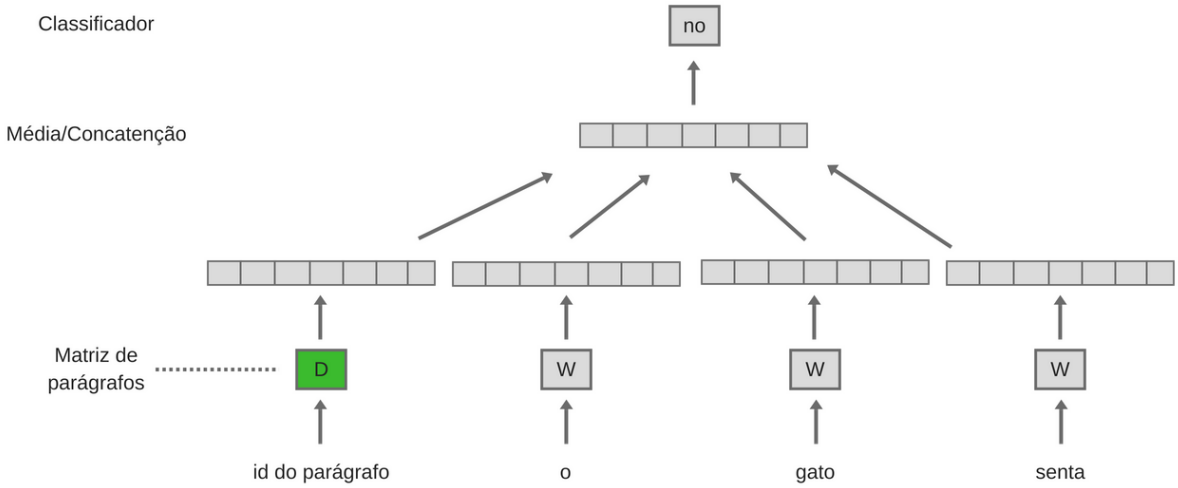
O método realiza um mapeamento baseado em probabilidades, de modo com que as palavras que possuem o mesmo sentido sejam distribuídas em um mesmo espaço vetorial, possibilitando realizar a distinção semântica entre as palavras de um parágrafo. Sequencialmente, o método realiza o mapeamento dos parágrafos para vetores distintos aos de palavras, concatenando o vetor do parágrafo com vários vetores de palavras presentes no parágrafo, com o objetivo de predizer a próxima palavra no contexto considerado. Dessa forma, são levados em conta o tamanho variável das sentenças, a ordem das palavras e a semântica. Tanto os vetores de palavras, quanto os de parágrafo são treinados pela descida de gradiente estocástica e pós-propagação (RUMELHART; HINTON; WILLIAMS, 1986).

Um ponto importante a ser destacado é o fato de que enquanto os vetores de parágrafo são únicos entre os parágrafos, os vetores de palavras são compartilhados (o vetor de uma palavra é o mesmo para todos os parágrafos que possuem aquela palavra). No momento da predição, os vetores de parágrafo são inferidos corrigindo os vetores de palavra e treinando o novo vetor de parágrafo até a convergência. Rumelhart, Hinton e Williams (1986) propuseram dois algoritmos para a geração de vetores de parágrafo:

- PV-DM (do inglês, *Distributed Memory Model of Paragraph Vectors*): cada parágrafo é mapeado para um vetor exclusivo, representado por uma coluna em uma matriz D. Cada palavra também é mapeada para um vetor exclusivo, representado por uma coluna em uma matriz W. A concatenação ou média do vetor de parágrafo com os vetores de palavras são utilizados para prever a próxima palavra em um contexto. O vetor de parágrafo pode ser considerado uma pseudo-palavra e representa as informações que faltam no contexto atual, atuando como uma memória do tópico do parágrafo em questão. A Figura 2 apresenta a estrutura do modelo PV-DM, considerando a sentença “o gato senta no sofá”.

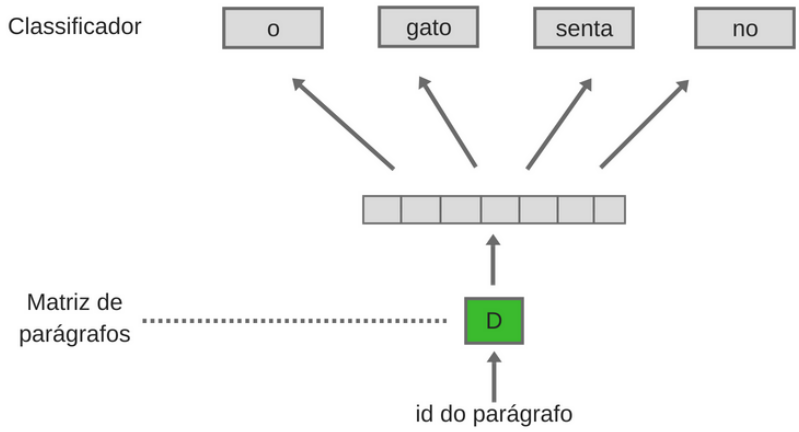
- PV-DBOW (do inglês, *Distributed Bag of Words version of Paragraph Vector*): as palavras de contexto são ignoradas na entrada e previstas aleatoriamente a partir do vetor de parágrafo. Na Figura 3 é apresentada a estrutura do modelo PV-DBOW.

Figura 2 - Estrutura do modelo PV-DM.



Fonte: autores.

Figura 3 - Estrutura do modelo PVDBOW.



Fonte: autores.

3 METODOLOGIA

De acordo com Schwartz (2002 *apud* TEIXEIRA, 2011), as redes de conhecimento podem ser definidas como um ambiente de troca de informações e experiências entre profissionais de áreas distintas. Atualmente, com a multiplicação dos meios de produção

de conteúdo e comunicação aliada à necessidade de conhecimento para a competitividade entre as empresas, houve uma constante transformação da informação do estado físico para o estado digital. Essa informação digital passou a ser transmitida, utilizada, multiplicada e alterada de maneira simples com custo relativamente baixo (TEIXEIRA, 2011). Este cenário permite a aplicação de uma série de abordagens voltadas à análise e geração de novas informações, e consequentemente, conhecimento. Neste contexto, no âmbito do objetivo do presente trabalho, pretende-se realizar o levantamento e análise de redes de conhecimento partindo-se da aplicação de conceitos, técnicas e ferramentas de mineração de dados e redes complexas.

Durante a etapa inicial, foi considerado uma amostra de dados textuais, provenientes de bases de teses e dissertações obtidas a partir Biblioteca Digital de Teses e Dissertações (BDTD) do Ibict. Neste contexto, para a amostra considerada foi extraído o conjunto de teses e dissertações disponibilizadas no portal entre os anos de 2000 e 2018, para uma Instituição Federal de Ensino Superior em específico, totalizando 2.644 documentos. Considerando que o portal não disponibiliza um serviço de API (do inglês, *Application Programming Interface*) para extração dos documentos textuais, foi desenvolvido um mecanismo baseado em *Web Scraping*. De acordo com Malik e Rizvi (2011), *Web Scraping* pode ser descrito como um método de extração automática de dados da web a partir de documentos HTML, com base em análise estrutural das páginas. O mecanismo construído, obteve de forma automática todos os documentos disponíveis em formato PDF. A partir disso, esse conjunto de textos foi então convertido para um formato de leitura compreensível por sistemas computacionais, gerando uma nova base de textos seguindo um padrão ASCII, de acordo com padrões de dados abertos processáveis por computador (FERNANDES; CORDEIRO, 2016).

O segundo passo foi a extração dos campos textuais de interesse. Para tanto, no âmbito do presente trabalho, foram considerados apenas os dados referentes aos resumos de cada documento. A partir disso, os dados foram então extraídos compondo um único corpus textual, onde cada resumo foi sinalizado com um identificador único, associado de forma crescente no intervalo [1, 2.644]. O corpus textual foi então submetido ao método Doc2Vec para detecção de similaridade e ao método de classificação proposto por Reinert (1990).

Para classificação através do método de Reinert (1990) foi utilizado o software de análise textual IRaMuTeQ1. Para a classificação foi selecionado o método de classificação

simples baseada em segmentos de textos, com número inicial de classes igual a dez. É importante destacar que este valor trata apenas de uma estimativa inicial, que é ajustada através do processamento do algoritmo.

No âmbito do método Doc2Vec, foi realizado um pré-processamento de modo a reduzir a influência de termos ou palavras irrelevantes à semântica (preposições, pronomes e artigos). Neste sentido, o algoritmo Doc2Vec foi então aplicado para a concepção de um modelo de aprendizado de máquina não-supervisionado. A parametrização do modelo, para propósitos de treinamento sobre a base denominada aqui de *train-corpus*, foi realizada com base em trabalhos correlatos (LEE, 2005; MIKOLOV et al., 2013; MIKOLOV et al., 2013a), dispondo dos seguintes dados:

- Dimensionalidade dos vetores de palavras (*size = 15*);
- *window*: quantidade de palavras anteriores e posteriores à palavra alvo. Este parâmetro é utilizado para a predição da palavra no contexto (*window = 5*);
- *mincount*: valor mínimo de frequência, a partir do qual palavras ou termos serão consideradas, ou seja, atribui uma noção de relevância, descartando palavras com poucas ocorrências. Segundo revisão bibliográfica, utilizada para referência na definição dos parâmetros, a faixa de valores [10,20] é utilizada para corpus contendo dezenas de milhares a milhões de documentos. Empiricamente, tais trabalhos demonstraram que sem uma variedade de exemplos representativos de documentos (como é o caso do presente trabalho), a retenção de muitas palavras raras pode tornar o modelo pior. O valor oi definido empiricamente como *mincount = 1*;
- *hs*: se 1, a função *softmax* hierárquica será utilizada para o treinamento do modelo (*hs=1*);
- *dm*: define o algoritmo de treinamento. Por padrão, o DBOW é usado (*dm = 0*). O outro é o DMPV (*dm = 1*). Foi utilizado *dm = 1*;
- *dm-concat*: se 1, usa a concatenação dos vetores de palavras e vetores de parágrafo para atribuir o contexto da representação. Portanto, foi utilizado *dm-concat = 1;*
- *iter*: número de iterações (épocas) de treinamento sobre *train-corpus*. Foi utilizado *iter = 100*. Tal valor foi definido empiricamente e considerando a dimensionalidade do conjunto de treinamento.

Com o propósito de validação do modelo, a partir de *train-corpus* foram gerados vetores de termos para os documentos por meio de inferência. Neste sentido, o algoritmo

de inferência realiza a predição dos termos com base nos vetores de palavras, sendo que estes novos vetores podem ser comparados com os vetores do modelo treinado. Basicamente, nesta abordagem, *train-corpus* é tratado como um dado desconhecido pelo modelo e, uma vez identificada semelhança entre os vetores (inferidos e modelados) obtém-se uma noção da consistência do modelo. Embora não seja um valor real de precisão, é uma forma de validar quão representativo é o modelo para as características dos documentos da base de dados. Neste sentido, a partir do modelo treinado e validado, se realizou a análise de similaridade semântica dos documentos. Para critérios de representação, um grafo ponderado foi gerado, ilustrando a relação entre os documentos. Neste sentido, a matriz de similaridade foi convertida em um grafo direcional ponderado, no qual, para cada nó é associado um peso referente ao somatório dos valores de similaridade com os outros documentos.

4 RESULTADOS

Através do método de classificação de Reineirt (1990), foram obtidos dados referentes às classes e aos termos representativos de cada uma destas. Conforme pode ser observado na Figura 4, é possível notar que o método identificou seis classes principais. O Quadro 1 apresenta, para cada uma das classes, os dez termos mais relevantes. Essa análise indica a possibilidade de seis áreas principais de pesquisa no âmbito das teses e dissertações da IFES considerada. A Figura 5 apresenta a distribuição de cada um dos documentos em cada classe, enquanto a Figura 6 exibe os termos mais relevantes de todas as classes e uma representação do peso dos mesmo em relação à amostra geral.

Quadro 1 - Dez termos mais relevantes de cada classe.

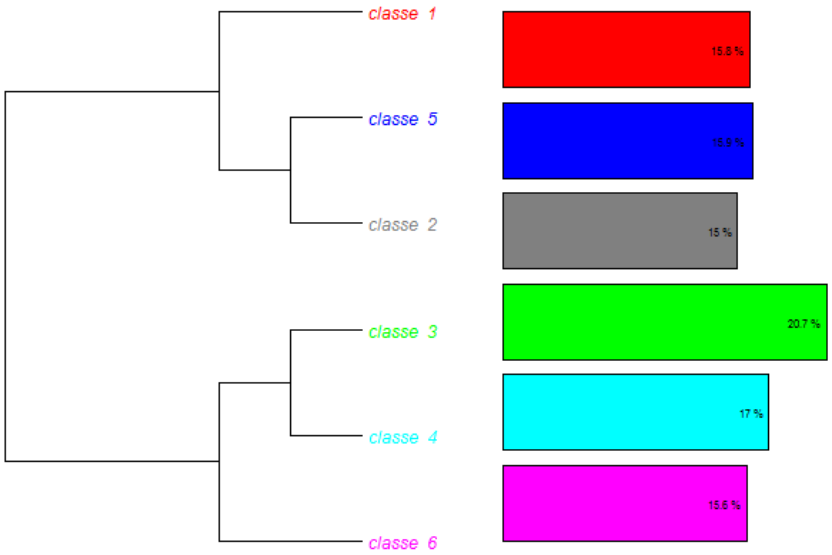
Classe	**Termos**
Classe 1	paciente; clínico; doença; idade; prevalência; diagnóstico; exame; sexo; grupo; risco.
Classe 2	concreto; método; resistência; nano-partículas; ensaio; modelo; linear; adsorção; superfície; magnético.
Classe 3	obra; autor; discurso; teórico; história; arte; cultural; identidade; pensamento; literário.
Classe 4	econômico; político; social; setor; urbano; ambiental; processo; gestão; organização; mercado.

Classe 5	extrato; experimento; proteína; dose; planta; ração; milho; folha; casca; cultivar.
Classe 6	professor; pesquisa; ensino; educação; curso; entrevista; universidade; aluno; aula; graduação.

Fonte: autores.

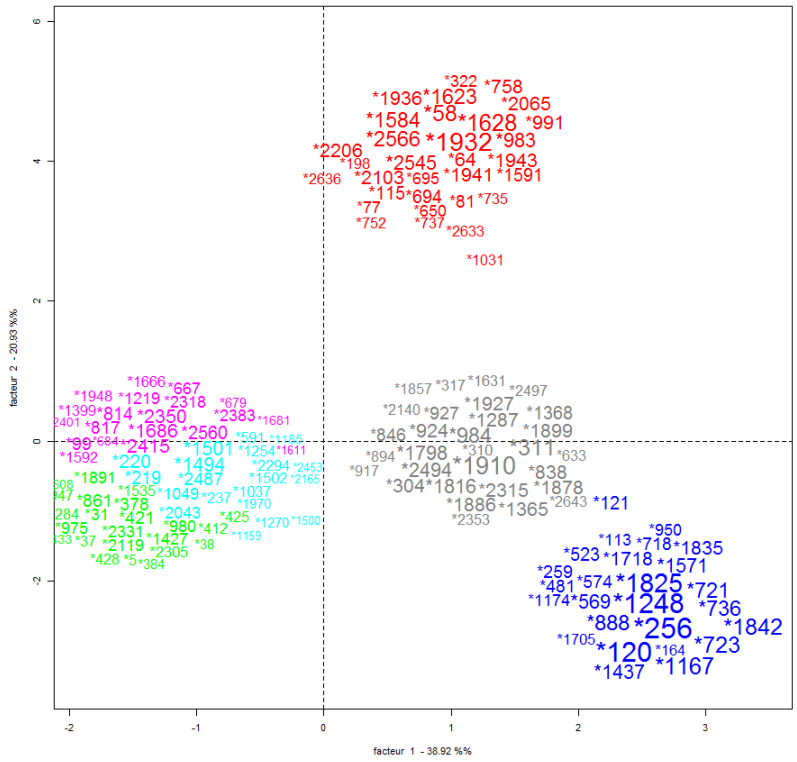
A partir disso, foi construído também o grafo de similitude referente ao corpus textual gerado (Figura 7). Esse grafo apresenta a relação entre os termos de maior relevância, apresentado como os mesmos se inter-relacionam dentro do conjunto de dados, ou seja, é possível inferir ligações intrínsecas entre termos característicos de diferentes classes.

Figura 4 - Dendograma de distribuição de classes segundo o método de Reinert (1990).



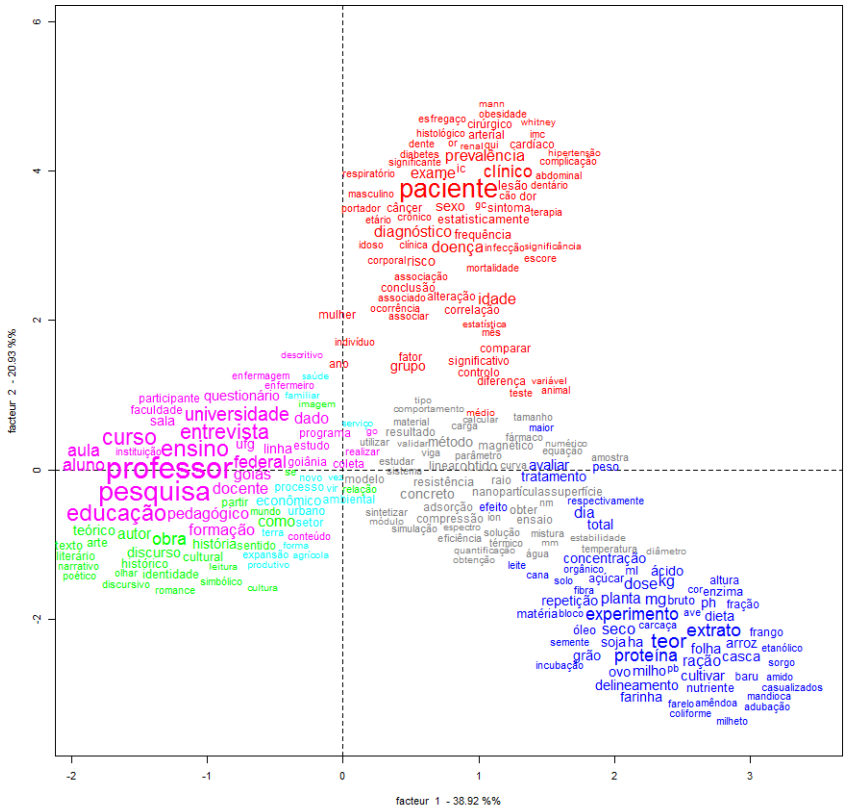
Fonte: autores.

Figura 5 - Rotulação dos documentos em suas respectivas classes.



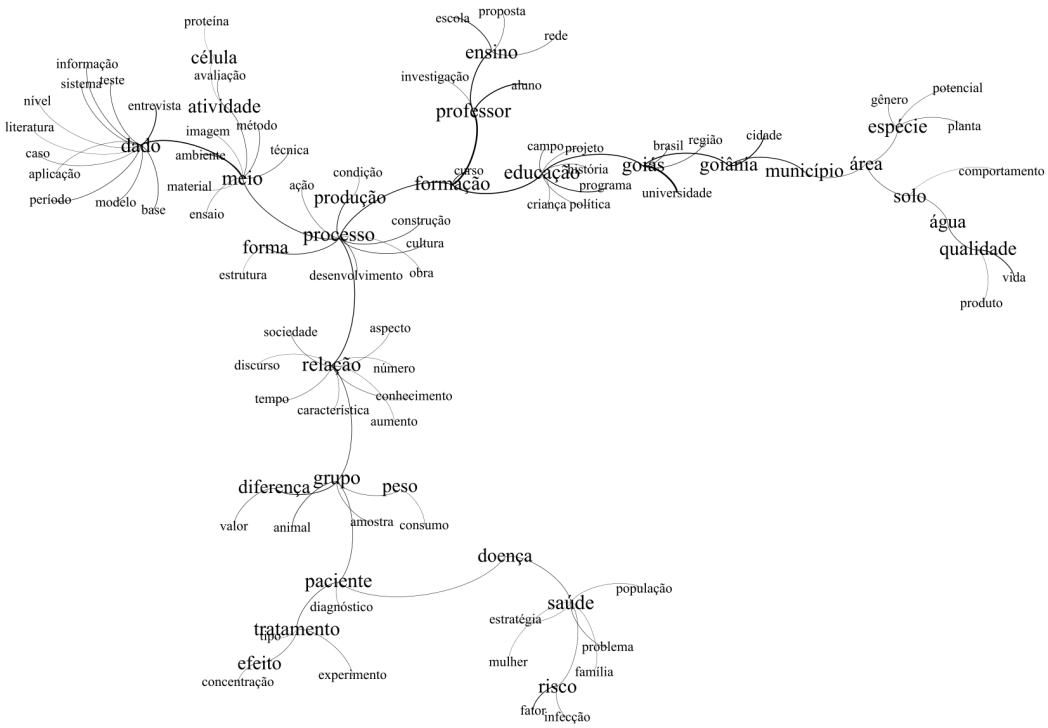
Fonte: autores.

Figura 6 - Termos de maior relevância em suas respectivas classes.



Fonte: autores.

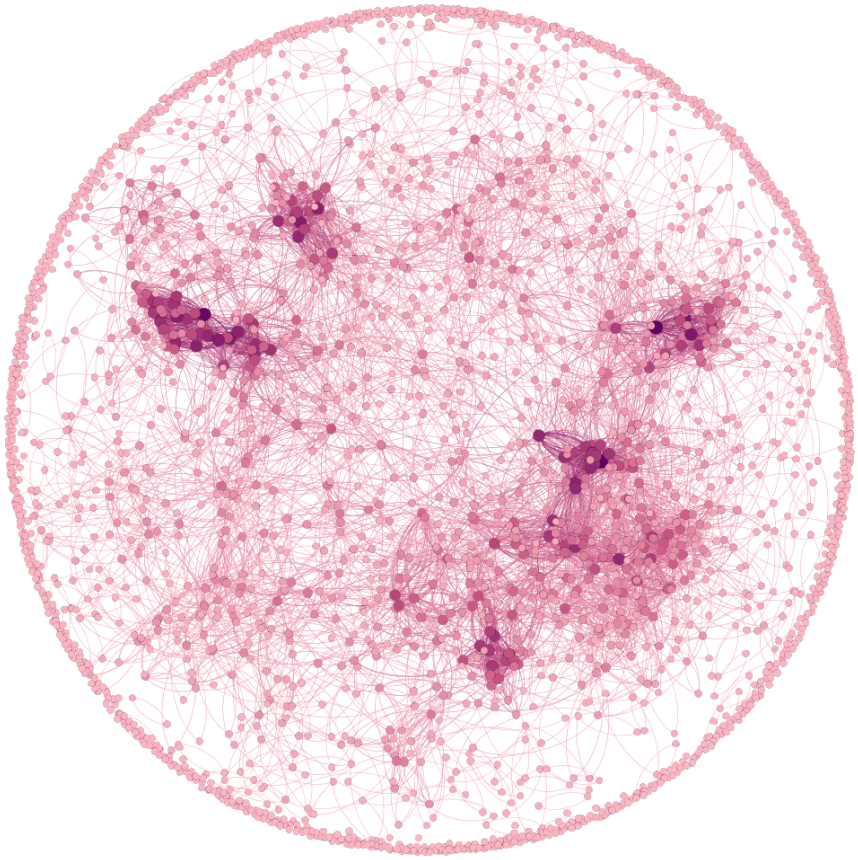
Figura 7 - Grafo de similitude da base de dados textual.



Fonte: autores.

Finalmente, a Figura 8 apresenta o grafo de similaridades entre os documentos, obtido através da aplicação da técnica Doc2Vec. Neste grafo, devido à questão de visualização dos dados, foram apresentadas as relações apenas entre os documentos com valor de similaridade maior ou igual a 0.9, sendo que os valores de similaridade estão no intervalo [0, 1]. Neste sentido, é possível ver a formação de zonas de agrupamentos, as quais validam as classes identificadas com método de Reinert (1990). Os nós referentes às fronteiras do grafo referem-se aos documentos que apresentaram similaridade menor que 0.9 para com todos os outros da base amostral.

Figura 8 - Grafo de similaridade entre documentos.



Fonte: autores.

5 CONCLUSÃO

De modo geral, é possível concluir que o estudo proposto é potencialmente relevante do ponto de vista de aplicação de técnicas e ferramentas de Processamento de Linguagem Natural para Recuperação da Informação. Ainda neste sentido, através dos resultados obtidos é possível identificar tendências e prováveis redes de conhecimento, determinadas através de tópicos e termos de relevância. Tal abordagem é interessante enquanto mecanismo de apoio em pesquisas bibliométricas. Para trabalhos futuros sugere-se utilizar os procedimentos metodológicos explorados (instrumento e técnicas) para a criação de coleções de dados científicos e, dessa forma, obter não apenas um estoque de dados, mas uma organização analítica de dados focada na extração de características e geração de conhecimento. Além disso, análises mais profundas e específicas podem ser aplicadas para detecção das conexões entre pesquisadores e

programas de pós-graduação com base nos resultados obtidos, assim como a classificação automática de novos documentos textuais.

REFERÊNCIAS

ADAMS, J. Collaborations: the fourth age of research. **Nature**, v. 497, n. 7.451, p. 557-560, 2013.

ANDRADE, P. H. M. A. **Aplicação de Técnicas de Mineração de Textos para Classificação de Documentos: um Estudo da Automatização da Triagem de Denúncias na CGU**. 2015. 65f. Dissertação de Mestrado - Instituto de Ciências Exatas - Universidade de Brasília, Brasília, 2015.

BEAVER, D. B.; ROSEN, R. Studies in scientific collaboration. Part II. Scientific co-authorship, research productivity and visibility in the French scientific elite, 1799-1830. **Scientometrics**, v. 1, n. 2, p. 133-149, 1979.

BEPPLER, M. D.; FERNANDES, A. M. R. Aplicação de text mining para extração de conhecimento jurisprudencial. In: Congresso Sul Catarinense de Computação, I, Criciúma. **Anais**..., Porto Alegre:SBC, 2005.

CASTRO, L. N.; FERRARI, D. G. **Introdução à Mineração de Dados: conceitos básicos, algoritmos e aplicações**. São Paulo: Editora Saraiva, 2016.

CORDEIRO, D. F.; LEAL, M. R. C.; VIEIRA, L. M.; DA SILVA, N. R. Cartografando comentários e sentimentos no perfil de Jair Bolsonaro no Instagram acerca da Covid-19. Galáxia, v. 47, **e56929**, 2022.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From Data Mining to Knowledge Discovery in Databases. **American Association for Artificial Intelligence**, v. 17, n. 3, 37-54, 1996.

FAWCETT, T.; PROVOST, F. **Data Science para Negócios**, Rio de Janeiro: Alta Books, 2016.

FERNANDES, J. L. F.; CORDEIRO, D. F. Avaliação de formatos de publicação de dados abertos governamentais através de indicadores de usabilidade. **Tendências da Pesquisa Brasileira em Ciência da Informação**, v. 9, n. 1, p. 65-84, 2016.

FIGUEIREDO, D. R. Introdução a Redes Complexas. In: SOUZA, A. F.; MEIRA, W. (Orgs.), **Atualizações em informática 2011**, Rio de Janeiro: PUC-RIO, 2011. Introdução às Redes Complexas, p. 303–358.

FIRME, S. M.; MIRANDA, A. C. D.; SILVA, J. A. Produção do conhecimento científico: um estudo das redes colaborativas. **Biblos**: Revista do Instituto de Ciências Humanas e da Informação, v. 31, n. 2, p. 45-61, jun./dez. 2017.

GLÄNZEL, W. Coauthorship patterns and trends in the sciences (1980-1998): a bibliometric study with implications for database indexing and search strategies. **Library Trends**, v. 50, n. 3, p. 461-473, 2002.

HUSSEIN, H.; ALAAELDIN, H.; HASSAN, M. Selection criteria for text mining approaches. **Computers in Human Behavior**, v. 51, p. 729–733, 2015.

JURAFSKY, D.; MARTIN, J. H. **Speech and Language Processing**: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition. Upper Saddle River, NJ, USA: Prentice-Hall, 2009.

LARIVIÈRE V.; GINGRAS Y.; SUGIMOTO C. R.; TSOU, A. Team size matters: collaboration and scientific impact since 1900. **Journal of the Association for Information Science and Technology**, v. 66, n. 7, p. 1323-32, 2015.

LE, Q. V.; MIKOLOV, T. Distributed representations of sentences and documents. In: INTERNATIONAL CONFERENCE ON MACHINE LEARNING, 31, Beijing, China, 2014. **Proceedings** [...] Beijing, China, 2014.

LEE, M. D.; WELSH, M. An empirical evaluation of models of text document similarity. In: ANNUAL MEETING OF COGNITIVE SCIENCE SOCIETY, 27, Stresa, Itália, 2005. **Proceedings** [...] Stresa, Itália, 2005. p. 1254–1259.

LOH, S. **Abordagem baseada em conceitos para descoberta de conhecimento em textos**. 195f. 2001. Tese de Doutorado - Universidade Federal do Rio Grande do Sul, Porto Alegre, 2001.

MAIA, M. F. S.; CAREGNATO, S. E. Co-autoria como indicador de redes de colaboração científica. **Perspectivas em Ciência da Informação**, v. 13, n. 2, p. 18-31, 2008.

MALIK, S. K.; RIZVI, S. H. M. Information extraction using web usage mining, web scrapping and semantic annotation. In: INTERNATIONAL CONFERENCE ON COMPUTATIONAL INTELLIGENCE AND COMMUNICATION NETWORKS, 3, Gwalior – India. **Proceedings... Gwalior – India: IEEE,** 2011, 465-469.

MEDIETA, J. G.; SCHMIDT, S. Análisis de Redes: Aplicaciones en Ciencias Sociales. [S.l.]: Unam, 2002.

MIKOLOV, T. et al. Efficient estimation of word representations in vector space. In: INTERNATIONAL CONFERENCE ON LEARNING REPRESENTATIONS, Arizona, US, 2013. **Proceedings**... Arizona, US: ICLR, 2013.

MIKOLOV, T. et al. Distributed representations of words and phrases and their compositionality. In: INTERNATIONAL CONFERENCE ON NEURAL INFORMATION PROCESSING SYSTEMS, 26, Lake Tahoe, Nevada, US, 2013. **Proceedings**..., 2013. v. 2, p. 3111–3119.

OLIVEIRA, M. B. et al. Lead Time Forecasting with Machine Learning Techniques for a Pharmaceutical Supply Chain. In: INTERNATIONAL CONFERENCE ON ENTERPRISE INFORMATION SYSTEMS, 23, Online, 2021. **Proceedings**..., 2021. v. 1, p. 634-641.

PERSSON, O.; GLÄNZEL, W.; DANELL, R. Inflationary bibliometric values: the role of scientific collaboration and the need for relative indicators in evaluative studies. **Scientometrics**, v. 60, n. 3, p. 421-432, 2004.

PRICE, D. J. S. **O desenvolvimento da ciência: análise histórica, filosófica, sociológica e econômica**. Rio de Janeiro: Livros Técnicos e Científicos, 1976

REINERT, M. Alceste - a methodology of textual data analysis and an application: Aurélia by gérard de nerval. **Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique**, v. 26, n. 1, p. 24–54, 1990.

RUMELHART, D. E.; HINTON, G. E.; WILLIAMS, R. J. Learning representations by back propagating errors. **Nature**, v. 323, p. 533–536, 1986.

SALTON, G.; MCGILL, M. J. **Introduction to Modern Information Retrieval**. New York: John Wiley & Sons, 1983.

SANTIN, D. M.; VANZ, S. A. S.; CAREGNATO, S. E. A Análise de Redes de Colaboração Científica com Base em Indicadores Bibliométricos. In: Franco, S.; Franco, M.; Leite, D. (Org.). **Educação superior e conhecimento no centenário da reforma de Córdoba: novos olhares em contextos emergentes**, Porto Alegre: EDIPUCRS, 2018. p. 189-207.

SCHWARTZ, G. **Redes: vias de acesso às profissões do futuro**. São Paulo: Aprendiz, 2002.

SILVA, C. G.; CASSIANO, K. K.; CORDEIRO, D. F. Mãe solo, feminismo e Instagram: análise descritiva utilizando mineração de dados. In: CONGRESSO DE CIÊNCIAS DA COMUNICAÇÃO NA REGIÃO CENTRO-OESTE, XXI, 2019, Goiânia. **Anais** [...] Goiânia, 2019. p. 1-14.

SILVA, L. A.; PERES, S. M.; BOSCARIOLI, C. **Introdução à Mineração de Dados: com aplicações em R**. Rio de Janeiro: Elsevier, 2016.

SILVA, N. F. F. **Análise de sentimentos em textos curtos provenientes de redes sociais**. 138p. 2016. Tese de Doutorado - Instituto de Ciências Matemáticas e de Computação – Universidade de São Paulo, São Carlos, 2016.

TEIXEIRA, M. R. F. **Redes de Conhecimento em Ciências e o Compartilhamento de Conhecimento**. 142p. 2011. Tese de Doutorado - Universidade Federal do Rio Grande do Sul, Porto Alegre, 2011.

VIEIRA, L. M.; CORDEIRO, D. F. Desinformação e Fakes News nas Redes Sociais: uma análise sob a perspectiva da Escola Canadense de Comunicação. In: ENCONTRO NACIONAL DE PESQUISADORES EM JORNALISMO, 17, 2019, Goiânia. **Anais Eletrônicos** [...] Goiânia: SBPJOR/UFG, 2019. p. 1-16.

Recebido em: 15 de fevereiro de 2022
Aceito em: 22 de dezembro de 2022
Publicado em: 23 de dezembro de 2022