Representação semântica vetorial para análise de similaridade de documentos textuais

Kátia Kelvis Cassiano[1], **Douglas Farias Cordeiro**[1]

[1]Faculdade de Informação e Comunicação – Universidade Federal de Goiás (UFG)
74.690-900 – Goiânia – GO – Brazil

{katiakelvis, cordeiro}@ufg.br

Abstract. *This paper is based on a Natural Language Processing tool called* ***Doc2Vec****, for the semantic representation of textual documents. The database of interest is composed of 44 (forty-four) undergraduate course final papers. Text mining techniques were used to process the digital archives of the monographs and generate the text. Each document is represented by word vectors and the model performs term inferences for semantic analysis. As a result, the similarity of the documents is in the form of a weighted graph, closeness between each element of the data sample.*

Resumo. *Este artigo descreve um modelo baseado em uma ferramenta de Processamento de Linguagem Natural denominada* ***Doc2Vec****, para representação semântica de documentos textuais. A base de dados de interesse é composta por 44 (quarenta e quatro) monografias de trabalhos de conclusão do curso Gestão da Informação da Universidade Federal de Goiás. Técnicas de mineração de texto foram utilizadas para processamento dos arquivos digitais das monografias e geração do corpus. Cada documento é representado por vetores de palavras e o modelo realiza inferência de termos para análise semântica. Como resultado, a similaridade dos documentos é apresentada na forma de um grafo ponderado, realçando a proximidade entre cada elemento da amostra de dados.*

1. Introdução

O registro de informações em documentos textuais é uma prática comum e cotidiana, sendo observada nos mais distintos cenários, desde o ambiente acadêmico, por meio dos registros das produções científicas em artigos, teses e dissertações, como também em ambientes organizacionais, em documentos, projetos, planejamentos, entre outros. O crescente volume deste tipo de registro de dados acaba por emergir desafios e obstáculos no que se refere à sua análise, demandando soluções que possam automatizar e fornecer informações de forma mais eficiente e eficaz, proporcionando a descoberta de conhecimento para o desenvolvimento de estratégias e tomadas de decisão mais assertivas.

Neste cenário, com os avanços alcançados no âmbito da mineração de dados, e neste caso mais específico, da mineração de textos, a aplicação de modelos provenientes desta área pode ser considerada como uma interessante alternativa, com resultados consideravelmente satisfatórios. Entretanto, para que seja possível alcançar soluções que sejam adequadas aos problemas a serem tratados, é necessário a realização de uma análise minuciosa acerca dos dados a serem considerados no problemas, assim como dos resultados desejados [Loh 2001]. Diante disso, através da

mineração de textos é possível explorar e desenvolver soluções voltadas à diferentes aspectos, tais como: classificação, segmentação, clusterização, associação, entre outros [Andrade 2015],[Castro and Ferrari 2016],[Hussein et al. 2015][Jurafsky and Martin 2009].

Diante deste contexto, a modelagem da similaridade semântica de documentos textuais pode ser destacada como uma área extremamente significativa para a ciência da informação, por possibilitar resultados nos campos da análise de sentimentos, recuperação e a geração de conhecimento para apoio à tomada de decisão, classificação de documentos, entre outros [Silva et al. 2016]. Entretanto, embora existam diversos modelos [Morais and Ambrósio 2007], tais como *bag-of-words*, *n-grama*, *skip-grama*, a subjetividade inerente ao conceito de similaridade acaba por tornar este um problema de complexa modelagem computacional. Neste sentido, é importante observar que a análise de documentos textuais requer uma modelagem consistente que considere aspectos relevantes como: ordenação, semântica e composicionalidade das palavras em uma sentença.

Neste sentido, este artigo traz um estudo de caso da aplicação de mineração de textos e representação semântica para o levantamento de similaridades em documentos textuais acadêmicos, como o propósito de fomentar dados e informações que possam ser posteriormente utilizadas em processos subsequentes de clusterização, agrupamento ou classificação. É importante ressaltar que o trabalho tem como propósito demonstrar a aplicação de uma das técnicas disponíveis para mineração de dados não estruturados, sem contudo desenvolver discussão comparativa com outras técnicas do estado da arte do PLN.

Primeiramente, são abordados os conceitos de Mineração de Texto e Processamento de Linguagem Natural. Em seguida, as características do algoritmo escolhido e a metodologia adotada para o desenvolvimento do trabalho, sendo apresentados detalhes da aplicação da técnica de Processamento de Linguagem Natural (PLN) denominada **Doc2Vec**, introduzido por [Le and Mikolov 2014], sobre uma base de dados textuais composta por um conjunto de monografias de trabalhos de conclusão do curso Gestão da Informação da Universidade Federal de Goiás. Por fim, para fins de análise, a matriz de similaridade dos documentos é apresentada na forma de um grafo ponderado, realçando a proximidade entre cada elemento da amostra de dados.

2. Mineração de Textos

A Mineração de Textos sempre se apresentou como um grande desafio no contexto da Ciência da Computação e da Ciência da Informação. Neste sentido, a constante evolução das técnicas de mineração de dados acabaram por proporcionar avanços consideráveis nesta área, permitindo a automatização de tarefas e processos que anteriormente demandavam um grande esforço humano. De forma geral, a mineração de documentos textuais pode ser vista como uma subárea da Recuperação de Informação (RI) [Salton and 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. Neste sentido, é interessante observar que os fundamentos que regem esta área estão intrinsecamente ligados às definições de dado, informação e conhecimento.

Embora os conceitos de dado, informação e conhecimento possam ser considerados relativamente triviais, é comum encontrar interpretações que acabam por mesclar suas

definições. Diante disso, de modo a qualificar os conceitos relacionados à classificação de documentos, é importante fazer um parênteses e pontuar tais definições. De acordo com [Silva et al. 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. É importante destacar que nem toda informação gerada é necessariamente útil e utilizada, e que nem todo dado processado é garantia de informação. Finalmente, conhecimento pode ser descrito como a informação explorada com algum propósito específico, ou seja, utilizada para, por exemplo, tomada de decisão, produção de cenários, entre outros.

Neste contexto, é possível concluir que a matéria prima essencial para o processo de análise são os dados. De forma geral, de acordo com [Silva et al. 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. Neste sentido, os dados estruturados podem ser descritos como aqueles que se referem ao resultado de processos transacionais, ou ainda de medição ou observação, sendo normalmente armazenados em uma tabela, ou em um formato que obedece um padrão pré-definido, facilmente compreensível por máquina. Por outro lado, os dados não-estruturados referem àqueles que não apresentam padrões pré-definidos, sendo necessárias rotinas adicionais para seu tratamento e processamento, tais como ocorre com textos, sons, vídeos e imagens.

A partir disso, no âmbito dos dados estruturados, a obtenção de informação e a geração de conhecimento podem ser alcançadas através do modelo KDD (do inglês, *Knowledge Discovery in Databases*) [Fayyad et al. 1996]. De forma 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 tais atividades, a Mineração de Dados se destaca como a etapa mais importante na obtenção dos resultados, a qual pode ser definida como uma área multidisciplinar que proporciona, através de rotinas automatizadas, o reconhecimento de padrões, o levantamento de estatísticas, a visualização, e a extração de informações em grandes conjuntos de dados. Entretanto, devido às particularidades inerentes às bases de dados textuais, é necessário o emprego de técnicas e modelos especializados, proveniente da subárea denominada Mineração de Textos.

A Mineração de Textos pode ser descrita como um processo baseado na utilização de rotinas computacionais, para extração de padrões e conhecimento sobre conjuntos de dados textuais não-estruturados [Loh 2001]. É interessante observar que alguns autores consideram que a Mineração de Textos pode ser definida como a aplicação do modelo KDD sobre dados textuais [Morais and Ambrósio 2007], porém é importante destacar que a mineração sobre este tipo de dado demanda técnicas que extrapolam os processos tradicionais do KDD, compondo o que é denominado de KDT (do inglês, *Knowledge Discovery from Text*). Os avanços relacionados ao KDT incluem, entre outras coisas, 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, de relacionamentos entre os termos mais relevan-

tes de um documento, análise de conteúdo, etc. Um exemplo deste tipo de aplicação é a análise de sentimentos em textos curtos aplicada no âmbito das mídias sociais [Silva 2016], ou ainda a exploração da mineração de textos para classificação de documentos [Hussein et al. 2015].

Neste sentido, é possível concluir que as técnicas provenientes do KDT podem trazer vantagens e benefícios tanto para problemas relacionados à volumes de dados extraídos em documentos eletrônicos da Internet, assim como em quaisquer outros cenários, como em documentos gerados por sistemas de informações gerenciais, ou similares. De acordo com [Beppler and Fernandes 2005], a maior parte dos dados de empresas estão contidos em documentos textuais, o que também acaba por potencializar a importância do KDT na geração de informação e conhecimento. Em termos práticos, pode-se descrever o KDT, por meio do conceito de 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 e contextualizadas 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 and 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). A Análise Semântica é usada na proposta apresentada no presente trabalho, sendo descrita com maiores detalhes na Seção 3.

Entre as aplicações possíveis através da Mineração de Textos usando Análise Semântica, está a Classificação de Textos. De acordo com [Jurafsky and Martin 2009], a classificação de documentos textuais trata da identificação de um determinado documento d com relação a um conjunto de classes C, sendo $C = c_1, c_2, c_3, ..., c_n$, 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 forma geral, de acordo com [Andrade 2015], os modelos de classificação de documentos textuais podem ser classificados de duas formas: single-label, no qual cada documento está associado a apenas uma classe, ou multi-label, no qual um documento pode estar associado a uma ou mais classes. Neste sentido, uma das soluções 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*).

Entretanto, em problemas em que não há disponibilidade de dados pré-rotulados, uma alternativa é a utilização de modelos não-supervisionados, como o caso do presente trabalho. Neste caso, pode ser realizada a clusterização de textos através de modelos de aprendizado não-supervisionado. De acordo com [Castro and Ferrari 2016], a

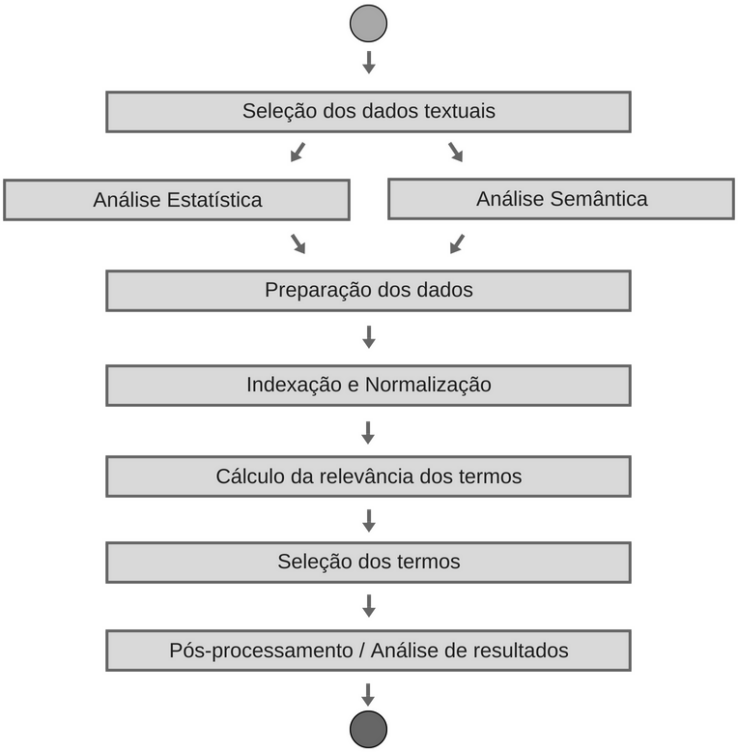

Figura 1. Processo de Mineração de Textos.

clusterização se refere a uma técnica de segmentação de dados com base na proximidade de padrões e tendências.

A clusterização é referenciada como um dos estudos iniciais realizados em processos de mineração, uma vez que através dela é possível realizar análise exploratória com a identificação de grupos ou classes que compartilham padrões e similaridades, 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 clusterização.

3. Processamento de Linguagem Natural

O Processamento de Linguagem Natural (PLN) constitui um conjunto de técnicas de inteligência artificial que possibilitam a comunicação entre seres humanos e computadores. No contexto do processamento de documentos textuais, PLN pode ser utilizado para diversos fins, tais como recuperação de informação, tradução, sumarização e classificação [Norvig and Russel 2011].

Em geral, um sistema PLN processa a linguagem em três níveis: morfológico, sintático e semântico. O sistema interpreta uma dada sentença por meio da análise de seu conteúdo léxico, das regras gramaticais e do significado dos termos, inicialmente armazenados em um dicionário. No contexto da inteligência artificial, o objetivo do PLN é expressar o conhecimento de forma tratável para os sistemas computacionais. Portanto, faz-se necessário e pertinente estudar formas de representação de sentenças e/ ou documentos em linguagem natural.

Segundo [Specia and Rino 2002], o processamento semântico é considerado um dos maiores desafios do PLN, dado a variabilidade morfológica e sintática das unidades lexicais e ainda, para o caso da língua portuguesa, ambiguidade de significados dos termos ou palavras. Neste sentido, um dos principais problemas sobre a interpretação semântica é referente ao processo composicional: as formas semânticas das entradas lexicais podem resultar em sentidos diversos para os termos ou palavras, dado que a interpretação depende de regras gramaticais.

A partir disso, o significado de um determinado termo de uma sentença depende dos significados dos demais termos e das regras gramaticais relacionadas. Tudo isso permite concluir que representação do conhecimento para o processamento de linguagem natural não é trivial, demando a aplicação de técnicas e adequações específicas para cada tipo de problema tratado, levando-se em conta fatores como a representação semântica envolvida.

Neste contexto, no âmbito da linguística, o termo semântica caracteriza-se como um dos componentes do conhecimento, cuja função é representar o significado de uma sentença. No contexto do PLN, a extração do significado de expressões e a representação do mesmo por meio de estruturas semânticas constituem estágios do processamento [Specia and Rino 2002]. Quando se trata de documentos textuais, é importante observar que o significado depende não somente das informações semânticas dos termos isolados mas também da forma como essas informações são dispostas em um contexto. É importante destacar que existem uma série de técnicas e métodos baseados nestes conceitos e definições, entre as quais está a técnica **Doc2Vec**, utilizada no presente trabalho.

3.1. Paragraph Vector (Doc2Vec)

Paragraph Vector, também descrito como **Doc2Vec** e proposto por [Le and Mikolov 2014], é uma ferramenta de Processamento de Linguagem Natural para representar documentos e é uma generalização do método **Word2Vec**, introduzido por [Mikolov et al. 2013a]. De forma geral, esta técnica consiste de um modelo de aprendizado não-supervisionado que utiliza de representações vetoriais distribuídas dos termos ou palavras de um texto.

Neste sentido, os textos referentes à base de dados considerada podem ser de tamanho variável, de sentenças a documentos completos e, de uma forma geral, os vetores são treinados para predizer palavras ou termos em um parágrafo e assim atribuir uma representação semântica. A partir disso, inicialmente, é realizado um mapeamento baseado em probabilidades, de forma que as palavras que possuem o mesmo sentido são distribuídas em um mesmo espaço vetorial, possibilitando realizar a distinção semântica entre as palavras de um parágrafo. Por exemplo, sejam os termos “poderoso”, “forte” e “Paris” em um segmento textual, esta representação vetorial pode descrever que o termo “poderoso” é semanticamente mais próximo de “forte” que de “Paris”.

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 et al. 1986]. É

importante destacar 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 et al. 1986] propuseram dois algoritmos para a geração de vetores de parágrafo:

- **PV-DM** (do inglês, *Distributed Memory Model of Paragraph Vectors*): neste modelo, 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"*.

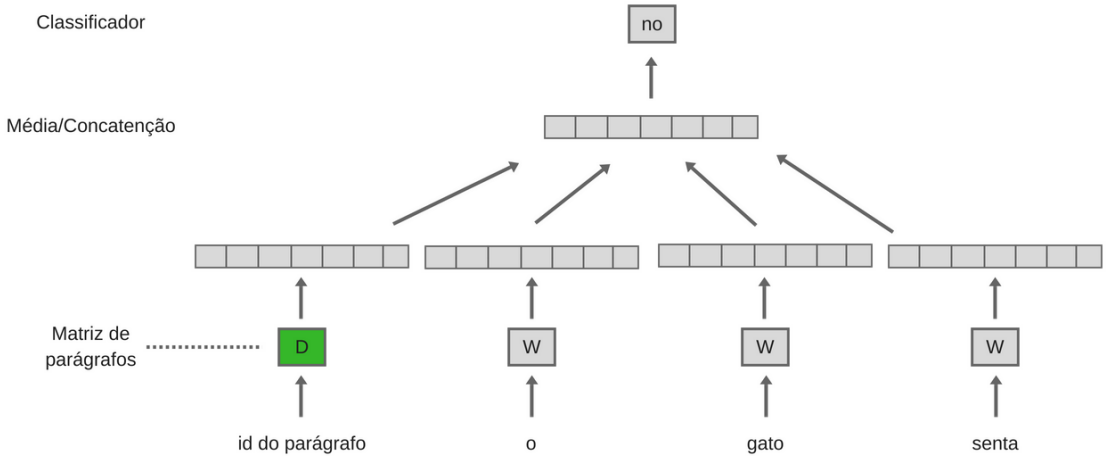

Figura 2. Estrutura do modelo PV-DM.

- **PV-DBOW** (do inglês, *Distributed Bag of Words version of Paragraph Vector*): neste modelo, 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.

4. Metodologia

O objetivo do presente trabalho é aplicar o método **Doc2Vec** com o propósito de construir uma representação da similaridade entre documentos textuais dentro uma base de dados textuais composta por documentos acadêmicos. Para tanto, a base de dados tratada neste trabalho é composta por 44 (quarenta e quatro) monografias do curso Gestão da Informação da Universidade Federal de Goiás (UFG), disponibilizadas em formato pdf (*Portable Document File*). No escopo deste trabalho, o termo documento pode fazer

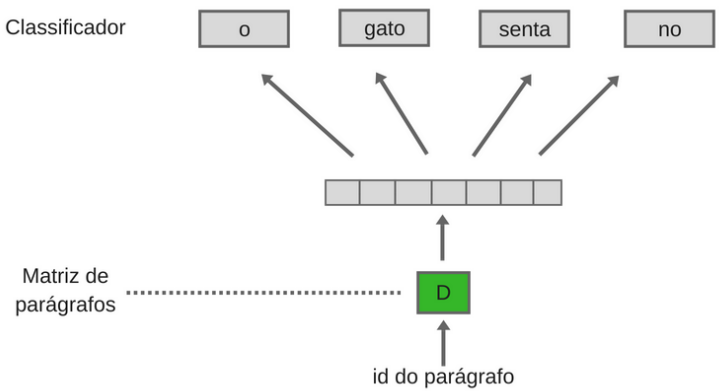

Figura 3. Estrutura do modelo PV-DBOW

referência a uma sentença, um artigo científico, uma monografia, um livro ou ainda um resumo destes. Em Processamento de Linguagem Natural (PLN) uma coleção ou conjunto de documentos é denominada *corpus*.

A partir disso, primeiramente, foi realizado o pré-processamento da base de dados para gerar, então, o *corpus* textual. Neste, cada documento é representado em uma linha e descrito por 5 (cinco) atributos, a saber:

- **doc-id**: identificador da monografia
- **autor**: nome do autor da monografia
- **orientador**: nome do professor orientador da monografia
- **titulo**: título da monografia
- **resumo**: resumo da monografia

Para reduzir a influência de termos ou palavras muito frequentes e irrelevantes no que tange à semântica, tais como preposições, pronomes e artigos, foram retiradas as *stopwords* de todo o *corpus* textual. Para processamento do *corpus* textual foi ainda adicionado um *label* para cada resumo, de forma a identificá-lo posteriormente (**doc-id**).

O algoritmo **Doc2Vec** foi utilizado para criação de um modelo de aprendizado de máquina do tipo não supervisionado. Neste sentido, inicialmente, uma função foi definida para ler e pré-processar o *corpus*. Este pré-processamento consiste, basicamente, em *tokenizar* o texto em termos ou palavras, gerando um vetor para cada documento. Ressalta-se que cada linha do *corpus* define um documento e o seu comprimento pode variar. Como resultado, tem-se a base de treinamento, denominada aqui de **train-corpus**, e a base de testes, denominada de **test-corpus**.

Para treinamento do modelo, a cada documento de **train-corpus** foi associada uma **tag**, sendo utilizado o número da linha que contém o documento, com base em zero. A parametrização do modelo foi realizada com base em trabalhos similares [Lee and Welsh 2005, Mikolov et al. 2013b, Le and Mikolov 2014]. A seguir são apresentados os valores para os principais parâmetros do modelo final validado:

- **size**: dimensionalidade dos vetores de palavras. Foi utilizado $size = 15$;
- **window**: quantidade de palavras anteriores e posteriores à palavra alvo. Este parâmetro é utilizado para a predição da palavra no contexto. Foi utilizado $window = 5$.

- **mincount**: define o 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. Portanto, foi definido empiricamente $mincount = 1$.
- **hs**: se 1, a função *softmax* hierárquico será utilizada para o treinamento do modelo. Foi utilizado $hs = 1$.
- **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.

Para validar o modelo, a partir de **train-corpus** foram gerados vetores de termos para os documentos por meio de inferência. O algoritmo de inferência realiza a predição dos termos a partir dos vetores de palavras e então 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, foi realizada análise de similaridade semântica dos documentos. Um grafo ponderado foi gerado para ilustrar a relação entre os documentos. Para tanto, 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

5. Resultados, discussões e conclusão

A representação semântica tratada no presente trabalho tem como objetivo a análise de similaridade dos documentos. Para melhor visualização dos resultados obtidos, é apresentado na Figura 4 o grafo ponderado da matriz de similaridade dos documentos. No referido grafo, cada nó se refere a um documento da base de dados, o qual por sua vez representa uma monografia de trabalho de conclusão do curso Gestão da Informação da UFG. Tendo como critério o tamanho do nó, pode-se ter uma ideia da distribuição dos documentos: nós de mesmo tamanho representam documentos similares entre si; quanto maior a diferença de tamanho entre um nó e outro, menor a similaridade entre os documentos representados pelos respectivos nós, ou seja, abordam conceitos de áreas de conhecimento distintas.

De uma forma geral é pertinente dizer que o grafo transmite uma ideia de agrupamento dos documentos de acordo com o tamanho do nó, e ainda que grande parte dos

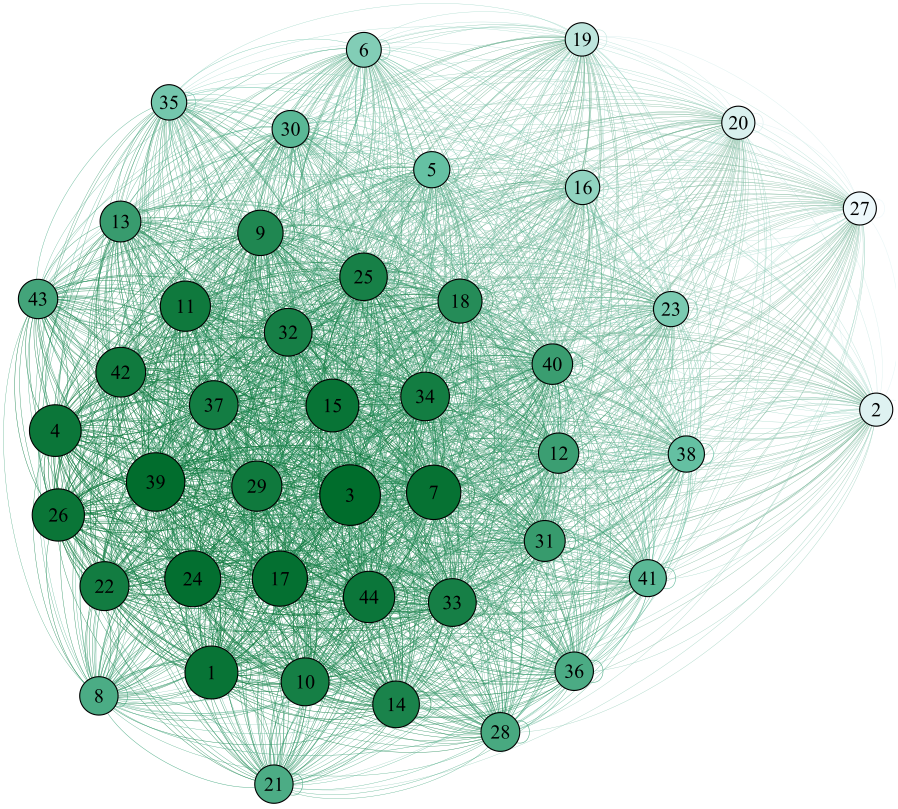

Figura 4. Grafo ponderado da similaridade dos documentos.

documentos estão concentrados em um dos grupos, que se difere dos demais pelos tamanhos dos nós. Ressalta-se ainda que tal observação é apenas uma inferência tomada a partir da análise visual, uma vez que não são objetivos do modelo proposto realizar a clusterização tampouco a classificação de documentos. No entanto, este tipo de hipótese é relevante e coerente quando da exploração de características e extração de padrões para subsidiar, posteriormente, processos de geração de conhecimento.

As Figuras 5 e 6, em Anexo, apresentam os resultados da análise de similaridade, bem como as características (**doc-id**, **título**,**resumo**) dos documentos representados pelos nós 42 e 5, respectivamente. Para cada documento alvo (**TARGET**), são apresentados seu identificador, o título e o resumo (atributos separados por **”**) e, em seguida, estas mesmas características para os documentos de maior (**MOST**) e menor (**LEAST**) grau de similaridade.

As características do documento 42, por exemplo, validam a informação apresentada no grafo da 4. Os conceitos abordados no resumo do respectivo documento (definição de indicador/ gestão estratégica) são, de fato, significativamente distintos dos conceitos abordados no documento 2 (modelagem de processos) e coerentes com os conceitos abordados no documento 11 (análise de risco/ gestão estratégica).

O grau de similaridade é uma característica importante na presente análise. Observa-se na Figura 6 que os graus de similaridade para o documento 5 são em valor

absoluto relativamente menores que os apresentados para o documento 42 - vide grau apresentado para o documento 20, que é o mais similar ao documento 5. Tal resultado permite o seguinte apontamento: o fato de um documento ser similar ao outro não significa, necessariamente, que tratam do mesmo tema ou área de conhecimento do curso, mas pode ser uma característica relevante para identificação de grupos em processos de clusterização de documentos.

No que tange ao tamanho dos nós, observa-se no grafo que os nós referentes aos documentos 5 e 20 tem praticamente o mesmo tamanho e são relativamente menores que o nó referente ao documento 42. Tal fato possibilita definir uma hipótese de que os documentos 5 e 20 pertencem a um mesmo grupo e, ainda, por similaridade, o documento 42 é distinto do documento 20, assim como a análise da 5 demonstrou ser distinto do documento 2. A hipótese definida anteriormente é validada pelas informações apresentadas na Figura 6, conforme descrito anteriormente.

De uma forma geral, o modelo apresentado no presente trabalho é potencialmente relevante do ponto de vista de aplicação de técnicas e ferramentas de Processamento de Linguagem Natural para Recuperação da Informação. Foi realizada análise do conteúdo dos resumos das monografias do curso Gestão da Informação da UFG numa tentativa de representá-las semanticamente, atribuindo assim significado.

Uma abordagem potencial para trabalhos futuros é utilizar os resultados 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. É possível, ainda, utilizar os resultados para classificação e correlação de áreas científicas. Nesse sentido, a solução abre um leque de oportunidades para análise de dados, recuperação da informação e geração de conhecimento no âmbito do Processamento de Linguagem Natural de textos científicos.

Referências

Andrade, P. H. M. A. (2015). Aplicação de técnicas de mineração de textos para classificação de documentos: um estudo da automatização da triagem de denúncias na cgu. Dissertação de mestrado, Instituto de Ciências Exatas - Universidade de Brasília, Brasília.

Beppler, M. D. and Fernandes, A. M. R. (2005). Aplicação de text mining para extração de conhecimento jurisprudencial. In *Anais do I Congresso Sul Catarinense de Computação*.

Castro, L. N. and Ferrari, D. G. (2016). *Introdução à Mineração de Dados: conceitos básicos, algoritimos e aplicações*. Editora Saraiva, São Paulo.

Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). *From Data Mining to Knowledge Discovery in Databases*. American Association for Artificial Intelligence.

Hussein, H., Alaaeldin, H., and Hassan, M. (2015). Selection criteria for text mining approaches. *Computers in Human Behavior*, 51:729–733.

Jurafsky, D. and Martin, J. H. (2009). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice-Hall, Upper Saddle River, NJ, USA.

Le, Q. V. and Mikolov, T. (2014). Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32, Beijing, China.

Lee, M. D. and Welsh, M. (2005). An empirical evaluation of models of text document similiarity. In *CogSci2005*, pages 1254–1259.

Loh, S. (2001). *Abordagem baseada em conceitos para descoberta de conhecimento em textos*. Tese de doutorado, Unviersidade Federal do Rio Grande do Sul, Porto Alegre.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. *Proceedings of Workshop at ICLR*, abs/1301.3781.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems*, volume 2, pages 3111–3119.

Morais, E. A. M. and Ambrósio, A. P. L. (2007). Mineração de textos. Technical report, Universidade Federal de Goiás, Goiânia.

Norvig, P. and Russel, S. (2011). *Inteligência Artificial*. Elsevier, 3 edition.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back propagating errors. *Nature*, 323:533–536.

Salton, G. and McGill, M. J. (1983). *Introduction to Modern Information Retrieval*. John Wiley & Sons, New York.

Silva, L. A., Peres, S. M., and Boscarioli, C. (2016). *Introdução à Mineração de Dados: com aplicações em R*. Elsevier, Rio de Janeiro.

Silva, N. F. F. (2016). *Análise de sentimentos em textos curtos provenientes de redes sociais*. Tese de doutorado, Instituto de Ciências Matemáticas e de Computação, São Carlos.

Specia, L. and Rino, L. H. (2002). Representação semântica: Alguns modelos ilustrativos. Technical report, NILC - ICMC-USP.

TARGET : 42 | Tornando tangível a participação social na Rede Humaniza SUS: proposta de indicador | Considerando que seria de interesse para as políticas públicas a adoção de mecanismos que explicitem níveis de interação horizontal, proporcionando ativação da inteligência coletiva e gerando propostas e soluções mais elaboradas, a investigação tem como objetivo propor um indicador que torne visível a participação social dos usuários de uma rede social online. Método: Baseia-se em análise de literatura pertinente relativa ao histórico de participação social no Brasil, no Sistema Único de Saúde e, particularmente à Rede Humaniza Sus (RHS). Apresentam-se características de indicadores que respaldam a discussão dos elementos que compõem o indicador proposto. Resultados: Foram identificadas as variáveis, considerando-se os elementos de postagem, comentários e votação e, com base nestes, criou-se um indicador ponderado de "participação social". Conclusão: O indicador proposto apresentou quais são os usuários mais importantes da rede e também pode auxiliar em traçar o tipo de perfil que esse usuário possui

SIMILAR/DISSIMILAR DOCS PER MODEL Doc2Vec(dm/c,d15,n5,w5,s1e-05):

MOST (0.8468265533447266): 11 | Gestão da Informação e análise de risco: proposta metodológica baseada em assimetria de informação e rating | O presente trabalho trata da identificação de elementos que contribuem para a ocorrência de riscos no processo de aplicação do Modelo Processual de Administração da Informação proposto por Choo (2006). Mesmo com diversas análises dos processos que compõe o Modelo registradas na literatura, estas análises não são abordadas sob o ponto de vista dos elementos constituintes da Teoria Econômica da Informação (Assimetria de informação e Rating). Diante dessa ausência na literatura, propôs-se um estudo para levantar os elementos que podem gerar riscos para a implementação eficiente do Modelo Processual de Administração da Informação. Num segundo momento este estudo procurou caracterizar instrumentos que possam auxiliar de forma eficaz na implantação do Modelo em estudo. O estudo baseou-se, por um lado, essencialmente em pesquisa bibliográfica e, de outro lado, utilizou-se a chamada "pesquisa básica" que consiste, genericamente, no fato de gerar novos conhecimentos. Ao final pode-se inferir, que, diante da sociedade contemporânea, o Modelo proposto por Choo (2006) é flexível e útil para se tomar decisões quando são criadas novas respostas a situações não esperadas. O Modelo Processual de Administração da Informação é sujeito a vários fatores/elementos dinâmicos (interrupções, ciclos de compreensão, ciclos de fracasso) que podem inviabilizar sua aplicação ou até mesmo impossibilitar que ele tenha êxito ao ser utilizado para tomar uma determinada decisão na organização.

LEAST (-0.674375057220459): 2 | Modelagem de processos aplicada à gestão de organizações públicas: Um estudo de caso em uma IFES | Este trabalho tem como objetivo principal realizar um estudo sobre a aplicação de gestão de processos de negócios em instituições públicas, através de um estudo de caso em um problema específico de uma Instituição Federal de Ensino Superior. Neste sentido, durante o trabalho serão abordados os principais conceitos, ferramentas e metodologias relacionados à informação, gestão de processos e gestão de organizações públicas, abordando questões como a cultura organizacional características destas instituições. A partir disso, é apresentado um estudo de caso sobre um processo específico, a avaliação de estágio probatório de docentes na Universidade Federal de Goiás.

Figura 5. Análise de similaridade do documento 42.

TARGET : 5 | Modelagem de processos e serviços oferecidos em Bibliotecas Universitárias: aplicações em serviços de atendimento | A pesquisa tem por objetivo analisar os processos dos serviços oferecidos pelo Sistema de Bibriotecas/SIBI, da Universidade Federal de Goiás/UFG, especificamente, o serviço de quitação de multas e o de empréstimos entre bibliotecas. Faz-se revisão de literatura contextualizando a pesquisa e abordando os principais conceitos pertinentes. A metodologia adotada é classificada como de natureza aplicada, pois tem como objetivo gerar conhecimentos para auxiliar a solução de problemas específicos. A análise completa foi desenvolvida por meio da identificação, mapeamento e modelagem dos processos citados. Conclui-se, a partir destas considerações, que a gestão por processos é de fundamental importância na gestão de qualquer organização.

SIMILAR/DISSIMILAR DOCS PER MODEL Doc2Vec(dm/c,d15,n5,w5,s1e-05):

MOST (0.2787836194038391): 20 | Processamento de Sinais Digitais Aplicado à Transmissão de Mensagens de Áudio Criptografadas | Este trabalho tem como propósito demonstrar a importância da segurança da informação junto aos métodos de processamento de sinais vinculando a criptografia. Essa demonstração tem como efeito relatar o quanto a segurança da informação é necessária e eficiente para as organizações em suas tomadas de decisões. Sendo assim, este trabalho tem como proposta o envio e recebimento de mensagens criptografadas através de arquivos de áudio com vínculo à segurança informacional.

LEAST (-0.4536975622177124): 40 | Autopoiese do conhecimento em redes sociais de conversação: em busca de evidências das implicações da composição biológica na dinâmica social de relacionamento em rede | O advento da chamada Era do Conhecimento intensificou os estudos acerca da importância da produção cognitiva, sobretudo nos ambientes empresariais. No entanto, para que se possa compreender como ocorrem os processos de realização do real e de como se dá a construção de conhecimento, é necessário analisar a relação entre processos oriundos do modo de ser biológico do ser humano em suas vivências autopoiéticas com a existência do ser em comunidade. Esta produção científica busca investigar a relação entre a composição biológica do ser humano e sua dinâmica relacional mundo contemporâneo e suas implicações na construção de conhecimento do indivíduo. Apontando que o modo de intensificação da comunicação em rede contemporânea tem profunda relação com a maneira de sustentação de visões individuais e coletivas na realização do ser.

Figura 6. Análise de similaridade do documento 5.