UMA FERRAMENTA DE VISUALIZAÇÃO DOS PRINCIPAIS TÓPICOS DE ARTIGOS CIENTÍFICOS BASEADA EM REPRESENTAÇÕES VETORIAIS DE PALAVRAS GERADAS POR MEIO DE DEEP LEARNING

Lucas Cerqueira Figueiredo (IC) e Orlando Bisacchi Coelho (Orientador)

Apoio: PIBIC Mackpesquisa

RESUMO

Neste trabalho exploratório buscamos apresentar elementos, técnicas e conceitos das áreas de *Deep Learning*, Representação Textual, Agrupamento de Artigos Científicos, Visualização de Documentos e Identificação de Tópicos em Artigos Científicos, que podem vir a ser utilizados na elaboração de uma ferramenta de visualização de tópicos de um conjunto de artigos científicos. Tal ferramenta facilitaria o trabalho de exploração da literatura e construção da base bibliográfica de uma pesquisa científica – tarefas inerentes a todo trabalho científico – tornando tal processo mais eficiente para o pesquisador. Dentro do campo de representação textual, exploramos as principais técnicas de Processamento de Linguagem Natural, com foco nas técnicas baseadas em *Deep Learning*, apresentando a evolução das arquiteturas dos modelos de linguagem até o presente estado da arte, representado pelo BERT (*Bidirectional Encoder Representations from Transformers*) e seus derivados. Apresentamos diversos trabalhos que mostram como tais representações textuais obtidas por modelos de linguagem podem ser agrupadas através de algoritmos de agrupamento e como tais representações e agrupamentos podem ser visualizados de forma a auxiliar o pesquisador na sua exploração de um dado conjunto de artigos científicos, facilitando o seu entendimento do contexto dos artigos componentes de tal conjunto e promovendo insights a respeito da relação entre os artigos do conjunto, por meio de técnicas de visualização.

Palavras-chave: *Deep Learning*. Identificação de Tópico em Artigos Científicos. Processamento de Linguagem Natural.

ABSTRACT

In this exploratory work we seek to present elements, techniques and concepts in the areas of Deep Learning, Textual Representation, Clustering of Scientific Articles, Document Visualization and Topic Identification in Scientific Articles that may be used in the elaboration of a topic visualization tool for scientific articles. Such a tool would help in exploring the literature and building the bibliographic basis for any given scientific research – steps that occur in any scientific work – making this process more efficient for the researcher. Within the

realm of textual representation, we explore the main techniques of Natural Language Processing, focusing on techniques based on Deep Learning, presenting the evolution of the language model architectures up to the present state-of-the-art, represented by BERT (Bidirectional Encoder Representations from Transformers) and its derivatives. We present several works that show how such textual representations obtained by language models can be grouped by using clustering algorithms and how such representations and groupings can be visualized in order to assist the researcher in her exploration of a given set of scientific articles, facilitating their understanding of the context of the articles in the set and promoting insights about the relationship between articles in the set, through visualization techniques.

Keywords: Deep Learning. Topic Identification in Scientific Articles. Natural Language Processing.

1. INTRODUÇÃO

A análise bibliográfica trata-se de uma tarefa inerente a todo e qualquer trabalho científico. O esforço e o tempo utilizados para coletar, organizar e examinar uma coleção de artigos científicos – que representarão as bases teóricas e metodológicas de uma pesquisa – são recursos que têm sido alvo de trabalhos que visam otimizar a tarefa de revisão da literatura, seja ela integrativa ou sistemática. Ferramentas de gerenciamento de referências, como o Mendeley (2020) e o Zotero (2020), são muito utilizadas por pesquisadores para organizar as referências já estabelecidas de seus trabalhos. Mas a composição de tal base bibliográfica resulta da busca e revisão de trabalhos que se enquadrem dentro do escopo da pesquisa, escopo que por sua vez é composto de variados tópicos. Cada um dos tópicos a serem abordados por um trabalho científico apresenta também um conjunto de diferentes trabalhos relacionados que podem vir a ser relevantes para embasá-lo teórica e metodologicamente. Porém, partindo de referências já estabelecidas do trabalho sendo desenvolvido, a atividade de identificar na bibliografia do artigo referenciado novos trabalhos relevantes e que tratam de tópicos similares não é uma tarefa objetiva. Muitas vezes, tópicos tratados por um artigo não são explicitamente indexados pelas ferramentas de gerenciamento de referências, algumas palavras-chave importantes para a descrição do artigo acabam sendo omitidas; ou mesmo as palavras-chave atribuídas acabam sendo generalistas demais e não permitem que o leitor possa distinguir o trabalho dentro do campo descrito. Isso acaba requerendo do pesquisador uma análise individual de cada uma das referências que compõem a bibliografia de um artigo, tarefa que pode se tornar pouco produtiva e efetiva na construção de sua própria base bibliográfica, dependendo da dimensão da coleção de referências e dos tópicos a serem explorados.

A tarefa de busca e assimilação de trabalhos relevantes normalmente envolve leituras e interpretações do que está sendo tratado no artigo ou na coleção de artigos candidatos. Num primeiro momento, essa leitura é feita de maneira com que se possa detectar os tópicos mais explícitos de cada documento, fornecendo um panorama superficial do que está sendo tratado. Caso um artigo apresente tópicos relacionados ao trabalho em desenvolvimento, ele então é analisado com maior profundidade e, caso tenha sua relevância comprovada, passa a compor a base bibliográfica do trabalho, podendo ser integrado também à seção de trabalhos relacionados.

O principal objetivo deste trabalho é apresentar técnicas e práticas que possam vir a compor uma ferramenta de visualização de conjuntos de artigos científicos, agrupados a partir dos seus principais tópicos, que combine conceitos dos campos de Representação Textual, Agrupamento de Artigos Científicos, Visualização de Texto e Identificação de Tópicos em Artigos Científicos, de forma a auxiliar pesquisadores em seus esforços de revisão da

literatura e busca por trabalhos relacionados dentro de determinado tópico. Para isso, foram estudadas as principais técnicas de visualização, especificamente visualização de documentos, e selecionados os elementos visuais mais relevantes e descritivos para uma revisão de literatura. Em seguida, foram exploradas as técnicas mais modernas para codificar palavras, cuja importância está na sua capacidade de extrair uma representação das palavras de um *corpus*, de modo a conter na sua representação forte dose de informação semântica e de uso na língua da palavra. Nesse sentido, focou-se nas representações vetoriais contínuas de palavras desenvolvidas na área de *Deep Learning*. Por fim, foram analisados trabalhos que utilizavam algoritmos de agrupamento para, operando sobre essas codificações de palavras, agrupar os artigos científicos, de modo a refletir a similaridade entre os tópicos abordados por diferentes trabalhos.

Na seção 2, a seguir, serão apresentados os principais trabalhos relevantes, distribuídos nos três campos pesquisados: Representação Textual, Agrupamento e Visualização de Texto. Na Seção 3 apresenta-se a metodologia da pesquisa. Na Seção 4 os resultados dos experimentos são descritos e analisados. Conclui-se o trabalho com as discussões finais e as a respeito de possíveis trabalhos futuros.

2. REFERENCIAL TEÓRICO

2.1. Deep Learning

Aprendizagem de Máquina (Mitchell, 1997) é uma área da Inteligência Artificial focada no desenvolvimento de algoritmos que permitem a computadores aprenderem a realizar uma determinada tarefa sem serem explicitamente programados para tal (Samuel, 1959). Algoritmos de Aprendizagem de Máquina (Raschka, 2015) são capazes de automaticamente criar modelos preditivos a partir de exemplos (dados) que lhes são fornecidos.

Dentre as várias técnicas de Aprendizagem de Máquina, uma que que tem se destacado nos últimos anos, e é a mais impactante, atualmente, chama-se *Deep Learning* (Aprendizado Profundo) (Goodfellow; Bengio; Courville, 2016). Uma Rede Neural Profunda (*Deep Neural Network*) é um modelo extremamente simplificado do Sistema Nervoso Central, consistindo em um tipo particular de grafo orientado e rotulado, com os vértices dispostos em camadas. Os neurônios artificiais – os vértices da rede – são unidades computacionalmente simples, que atuam como integradores não lineares dos sinais recebidos de outros neurônios (modulados pelas sinapses que eventualmente ligam esses neurônios) e, a partir disso, tornam-se mais ou menos ativos. As arestas do grafo correspondem às sinapses existentes entre os neurônios. Seus rótulos – números reais com sinal, chamados de pesos – indicam a força relativa de cada sinapse e se ela influencia o neurônio que recebe a sinapse a ficar mais

ou menos ativo. Um algoritmo de aprendizagem em *Deep Learning* adapta os pesos de modo a fazer a rede desempenhar uma determinada tarefa computacional, transformando as ativações das unidades da camada de entrada nas ativações desejadas para as unidades da camada de saída.

Com a aplicação de técnicas baseadas em Redes Neurais na área de Processamento de Linguagem Natural, podemos destacar resultados inovadores em termos de tradução automática de texto (Bahdanau; Cho; Bengio, 2015), bem como grandes avanços na resolução de uma série de tarefas, tais como: atribuição de papel semântico (Collobert et al, 2011), geração de sequências textuais (Sutskever; Martens; Hinton, 2011), representação da sintaxe e semântica lexical (Mikolov et al., 2013) e reconhecimento de fala (Graves; Mohamed; Hinton, 2013).

Em particular, à medida que os resultados obtidos com *Deep Learning* foram se consolidando, especialmente a partir do resultado obtido por Krizhevsky, Sutskever e Hinton (2012) na tarefa de classificação de imagens correspondente à competição ImageNet LSVRC-2010, notou-se que redes neurais treinadas a partir de muitas imagens (1,2 milhão no caso) aprendiam a desenvolver em suas primeiras camadas representações que eram úteis para qualquer tarefa envolvendo processamento visual. A partir de então, ao se desenvolver uma rede *Deep Learning* para realizar uma nova tarefa de processamento visual, tornou-se comum estender essa arquitetura (ou variantes mais poderosas dela posteriormente desenvolvidas) com outras camadas de computação, as quais seriam responsáveis por aprender essa nova tarefa; mas mantendo nas primeiras camadas a arquitetura original – bem como seus pesos. Dessa forma, o aprendizado extraído pela rede original, que também seria útil para a nova tarefa, seria transferido para a nova rede. Esse paradigma passou a ser conhecido como transferência de aprendizagem (*Transfer Learning*).

2.2. Representação Textual

Em Processamento de Linguagem Natural (PLN), a tarefa de representação textual é definida como o processo de extrair de elementos linguísticos apresentados textualmente uma representação passível de ser utilizada por modelos computacionais, de forma a solucionar a incapacidade dos computadores em entender e processar adequadamente a linguagem humana. O esforço de medir quantitativamente a semântica ou significado das palavras iniciou-se com um conceito da Psicologia denominado Diferencial Semântico (*Semantic Differential*) proposto por Osgood el al. (1957), que definia-se como uma escala que media as atitudes, opiniões e valores de um indivíduo em relação ao objeto, evento ou conceito representado por uma dada palavra. Algoritmos computacionais que executam tal tarefa tentam capturar a semântica de um termo, frase ou documento por meio de variadas técnicas

e abordagens, e com aplicações igualmente variadas. A princípio, dois tipos de modelos de representação textual são os mais comuns na literatura: modelos baseados em frequência e modelos baseados em predição.

2.2.1. Modelos Baseados em Frequência

Modelos baseados em frequência seguem uma premissa relativamente simples: documentos recebem sua representação vetorial como resultado de uma contagem. Os elementos dessa contagem variam de algoritmo para algoritmo, com alguns sendo descritos a seguir.

Proposta por Harris (1954), uma das abordagens mais tradicionais da literatura é conhecida como Saco de Palavras (*Bag-of-Words*), onde o valor de cada dimensão da representação vetorial de determinada sentença ou documento indica o número de vezes que um termo ocorreu em cada documento. Ao aplicar tal técnica a um *corpus* que apresenta n palavras únicas, a representação de cada documento apresentaria n elementos (ou dimensões), cada um referindo-se a um termo do corpus, e o valor de determinado elemento do vetor equivale à frequência no documento do termo ao qual se refere.

A ideia de que a frequência de um termo evidencia a sua significância é quase axiomática dentro do campo de Recuperação de Informação, porém foi proposta inicialmente por Luhn (1958). O autor desenvolveu um processo de geração automática de resumos de artigos científicos que utilizava o conceito de frequência de termos (TF - *Term Frequency*) para atribuir pesos às palavras – onde palavras mais frequentes recebiam pesos mais altos – e, em seguida, à sentenças, passando então a selecionar as frases do documento que melhor o descrevessem, compondo assim um resumo automático de tal artigo.

Porém, por basear-se em apenas em uma contagem, o conceito de TF pretere o fato de que termos muito frequentes em um documento nem sempre eram relevantes para sua definição. Palavras pertencentes a classes gramaticais como artigos, pronomes, preposições e conjunções normalmente são muito comuns em um texto, porém são termos que possuem pouco significado semântico e não contribuem significativamente para a definição de seu conteúdo. Essas palavras são conhecidas na linguística como palavras funcionais (contrastando com palavras lexicais, como substantivos e verbos), e na literatura de PLN como *stop words* (palavras de parada – em tradução livre) (BIRD; KLEIN; LOPER, 2019**)**.

Apesar de listas de *stop words* para diferentes idiomas serem utilizadas para filtrar tais termos em etapas de pré-processamento de texto em variadas tarefas de PLN, Sparck (1972) propôs uma medida estatística que iria atuar como uma métrica para a importância de um termo dentro de um corpus. Introduzida como *Term Specificity* (Especificidade de Termo) e disseminada na literatura como *Inverse Document Frequency* (*IDF* - Inverso da Frequência

nos Documentos), tal métrica pode ser entendida como uma medida da informação que uma palavra contém, sendo amplamente utilizada em motores de busca para resgatar os documentos que melhor correspondem a um dado conjunto de termos. Considerando $tf_{i,j}$ como a frequência do termo i no documento j, df_i como a quantidade de documentos que contêm i e N como o total de documentos no corpus, pode-se ponderar o fator de importância w de i contido em j ao combinar as medidas de frequência de termo (TF) e inverso da frequência em documentos (IDF):

$$w_{i,j} = tf_{i,j} \times log\left(\frac{N}{df_i}\right) \quad (1)$$

Apesar de o fator de importância w extraído pelo TF-IDF não ser uma representação vetorial, sua utilidade está em atribuir pesos diferentes a palavras em um texto de acordo com sua relevância, sendo possível então obter uma representação ponderada para tal documento, em função do fator de seus termos componentes.

2.2.2. Modelos Baseados em Predição

Nos modelos baseados em predição, uma representação vetorial para uma palavra é obtida através dos pesos das redes neurais utilizadas para realizar a predição de palavras dado um contexto, ou do contexto de uma dada palavra.

O CBOW (*Continuous Bag of Words*) (Mikolov et. al., 2013) é um algoritmo de predição de termos com base no contexto, que utiliza redes neurais simples (uma camada escondida) que recebem como entrada as palavras presentes no contexto de uma palavra (numa janela definida ao redor da palavra alvo), com cada palavra representada esparsamente (através de uns e zeros). O objetivo do treinamento da rede que gera o CBOW é, a partir do contexto de uma sequência de palavras, predizer qual seria a palavra faltante no contexto. Após o treinamento em diversos pares de contexto e palavra, a matriz de pesos dessa rede é então usada para representar de modo denso e contínuo cada uma das palavras presentes no vocabulário do *corpus*.

De modo inverso, agora representando a predição do contexto a partir de uma palavra, temos o algoritmo *Skip-Gram* (Mikolov et. al., 2013). Essencialmente, seu funcionamento é o inverso do CBOW: para cada exemplo de treinamento, recebe-se apenas a palavra que se busca obter a representação e, então, procura-se predizer as palavras presentes no contexto da entrada. Mikolov e colegas apresentaram um dos primeiros modelos de representação textual: o Word2Vec, contribuindo com algumas otimizações (*Hierarchical Softmax e Negative Sampling*) para que seu desempenho e acurácia melhorassem. Eles demonstraram que a arquitetura *Skip-Gram* se mostrou mais eficiente do que o CBOW para essa tarefa.

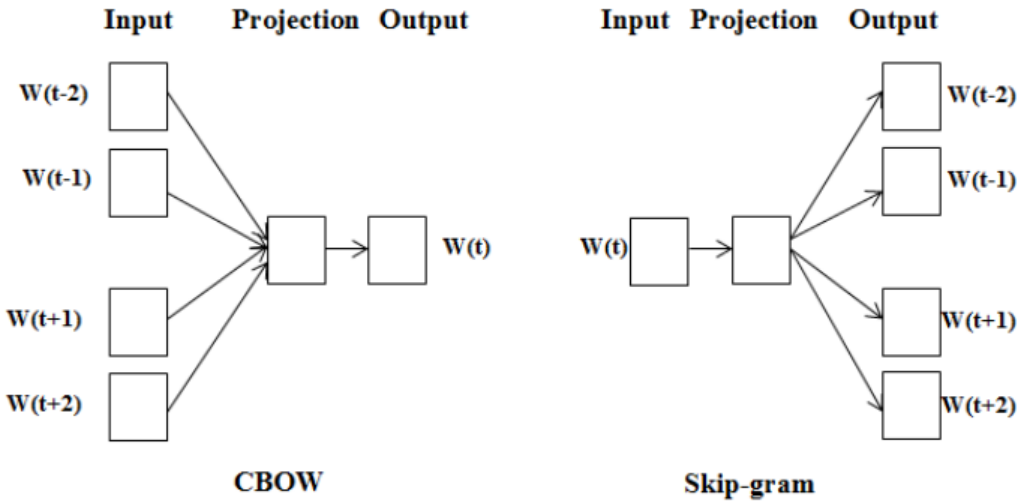


Figura 1 - Uma ilustração demonstrando as diferenças entre os modelos CBOW e Skip-gram. No primeiro modelo, temos como entrada quatro palavras (tamanho da janela = 4) e uma palavra como saída. No segundo, temos apenas uma entrada e quatro palavras como saída.

Partindo do Word2Vec, que utilizou o Skip-Gram clássico, Pennington et. al (2014) propôs o GloVe, que passou a utilizar aspectos estatísticos globais do texto para obter as representações. Em seguida foi proposto o ELMo (*Embeddings from Language Models*) (Peters et al., 2018) que passou a levar em consideração também a posição que a palavra se encontra numa sentença (e seus radicais). Como o representante do estado da arte atual, foi proposto o BERT (Bidirectional Encoding Representations from Transformers) (Devlin et al, 2019) que utiliza uma arquitetura de Transformers – arquitetura Deep Learning do tipo Encoder-Decoder – capaz de capturar um aspecto do processamento chamado de Atenção, que codifica a relação das palavras dentro de uma sentença (Vaswani et al, 2017).

O atual estado da arte em representação de palavras está dominado pelos modelos inspirados no BERT (DistillBERT, RoBERTa, SciBERT, etc). Isso mostra que a arquitetura *Transformer* é a que se mostra mais popular no processo de representação, apesar de ser custosa para ser trabalhada (uma grande massa de dados é necessária, como na maioria dos problemas de *Deep Learning*). Tais modelos apresentam também versões pré-treinadas que podem ser utilizadas, de modo que o trabalho de treiná-las seja minimizado. Os modelos pré-treinados podem então ser calibrados em bases de dados específicas do projeto, para que então obtenha-se representações consistentes com o domínio do problema, prática conhecida na literatura como *fine-tuning*.

2.3. Agrupamento de Artigos Científicos

Um exemplo típico e contemporâneo de agrupamento de documentos em bases científicas baseado em palavras – a utilização de aprendizado de máquina na tarefa de avaliação de grandes coleções de publicações científicas – é o desenvolvido por Zhang et. al (2018). Utilizando como base artigos disponíveis na MEDLINE contendo ambas as palavras "groundwater" e "disease", os autores desenvolveram ferramentas que realizam o pré-processamento dos artigos selecionados e usam algoritmos de agrupamento que analisam termos comuns nos títulos, resumos e palavras-chave dos artigos e os agrupam de acordo o tópico inferido.

Afonso e Duque (2014) desenvolveram um estudo a respeito de técnicas de agrupamento automático. Eles estudaram a eficiência de três algoritmos de agrupamento (Simple k-means, sIB e EM) na tarefa de agrupamento de artigos científicos e textos jornalísticos em português. Dividiram os experimentos em quatro estágios: (i) seleção dos textos e divisão em conjuntos de treinamento, teste e validação; (ii) categorização de cada palavra (*token*) de acordo com sua classe gramatical); (iii) algoritmos de filtragem para identificar as palavras mais significativas de cada documento; (iv) algoritmos de agrupamento, agrupando os documentos de acordo com sua similaridade, a partir das palavras obtidas no passo anterior.

2.4. Visualização de Texto

Como a identificação dos tópicos de artigos científicos é uma tarefa linguística, é razoável fazer a hipótese de trabalho de que uma melhor identificação dos tópicos será obtida se essas representações forem usadas na codificação dos textos dos artigos. Como as representações em si geram vetores de grande dimensionalidade, o correspondente a uma nuvem de palavras obtida de um texto científico exige o recurso a representações gráficas de vetores de alta dimensão.

Beck, Koch e Weiskopf (2016) desenvolveram e apresentaram uma ferramenta, chamada SurVis, que permite que autores de artigos de revisão (*surveys*), possam estruturar suas extensas listas de referências de forma visual, interativa e de fácil disseminação. O autor de um artigo de revisão, por exemplo, pode utilizar a ferramenta para permitir que outros pesquisadores e leitores de tal artigo interajam com uma representação visual da coleção de documentos citadas no trabalho. Tal visualização é composta de gráficos como a linha do tempo das publicações (por ano de publicação) e *sparklines* (pequenos gráficos do tamanho de palavras), permitindo também que o usuário interaja com os gráficos e visualizações ao aplicar filtros a palavras-chave, ano, autores, citações, entre outros. A ferramenta também permite que os artigos da coleção sejam agrupados (por meio do algoritmo *k-means*) em um

número de grupos definido pelo usuário, onde artigos serão agrupados com base em termos como palavras-chave ou autores, sendo então apresentados os agrupamentos por meio de pequenas nuvens de palavras, onde as palavras que melhor definem um grupo são visualmente maiores. Para indicar similaridade entre documentos na coleção são utilizados os vínculos entre citações, ou seja: se o documento A cita o documento B, os dois são ditos semanticamente similares.

Kucher e Kerren (2015) argumentaram a respeito da importância que o subcampo de visualização de texto vem obtendo, devido à constante e ubíqua necessidade de busca por trabalhos relacionados durante uma pesquisa, reforçando a relevância do trabalho aqui proposto. Os autores apresentam uma revisão da literatura que aborda as técnicas de visualização textual existentes; tal revisão é apresentada através de uma ferramenta interativa de desenvolvimento dos próprios autores, chamada de TextVis, que utiliza grafos (redes de co-autoria) para relacionar e exibir publicações de tópicos similares. Para desenvolver a ferramenta, os autores também propõem uma taxonomia, definindo métricas e elementos essenciais para ferramentas de visualização textual.

2.5. Identificação de Tópicos em Artigos Científicos

Tian e Zhuo (2017) apresentaram uma maneira de representar a similaridade entre artigos científicos através de uma análise da rede de citações de um dado conjunto de artigos, como forma de desenvolver um sistema de recomendação de artigos. Os autores apresentam as técnicas correntes utilizadas em sistemas de recomendação e propõem uma alternativa que desfruta também da representação densa (vetores contínuos) obtida através da técnica de fatoração de matrizes, comum em sistemas de recomendação. Para cada artigo em um dado conjunto é atribuída uma representação vetorial obtida através de sua rede de citação, que então pode ser utilizada para comparar a semelhança dos tópicos tratados por diferentes documentos do conjunto. Como inovação, nota-se que é levada em consideração apenas a rede de citações e como elas se estruturam nos documentos, extrapolando a dinâmica da representação semântica do nível de palavra ao nível de documento. A desvantagem dessa abordagem está ligada principalmente ao fato de que todo o processamento, treinamento e atribuição de representações vetoriais são realizados no escopo local de cada conjunto de artigos, tornando a comparação de artigos individuais fora do conjunto inviável.

Baseando-se em arquiteturas já existentes para representação vetorial contínua de palavras, Beltagy, Lo e Cohan (2019) apresentam o SciBERT, um modelo pré-treinado de linguagem baseado em *Transformer* (Vaswani et al., 2017) e derivado do BERT (Devlin et al., 2018) que é especializado na representação semântica do vocabulário comum a artigos científicos. Essa especialização foi obtida através do treinamento do modelo em *corpora* de

documentos de cunho científico. Os autores argumentam que os modelos de linguagem no estado da arte são mais comumente treinados em *corpora* de domínio geral, tornando tarefas envolvendo documentos do domínio científico menos efetivas, pois, segundo análises dos autores, há apenas 42% de sobreposição entre o vocabulário geral e o vocabulário científico.

3. METODOLOGIA

Primeiramente, será descrito o processo de construção da base de dados que foi utilizada nos experimentos. Em seguida, apresentamos a construção de um breve experimento na extração de palavras chave de textos através de um modelo pré-treinado do BERT.

Para que base de dados a ser utilizada fosse consistente com os objetivos do projeto, foram elaboradas três características que precisariam ser providas pelo critério de seleção de artigos, sendo elas:

1. Além dos critérios de busca (presente em praticamente todas as bases de artigos), a base deve permitir que seja realizado um filtro em relação às áreas do conhecimento consideradas na busca e ao intervalo de anos de publicação, de forma a limitar a quantidade de artigos retornados, assim como sua esparsidade temática.

2. A base deve fornecer para cada artigo, no mínimo: título e *abstract*. Esses são os elementos chave para o desenvolvimento do experimento.

3. A base de dados deve permitir que uma seleção de artigos seja exportada para um formato de fácil uso no experimento, com valores devidamente rotulados.

Após algumas buscas e interações com diversas ferramentas, o serviço que satisfez todas as propriedades foi o Web of Science. Tratando-se de um serviço de indexação de referências, há a possibilidade de filtragem das buscas em diversas facetas; cada artigo em sua base apresenta elementos essenciais como título, *abstract* e lista de citações (dentre diversos outros); e, por fim, a seleção realizada pode ser exportada para um formato acessível e bem definido.

Como critério de busca, foi utilizada o seguinte termo: "scientific paper representation", e os resultados foram filtrados de forma a limitar apenas artigos referentes à área COMPUTER SCIENCE ARTIFICIAL INTELLIGENCE. O resultado da busca retornou 74 artigos, que foram utilizados como a base dos experimentos.

Desenvolvemos um breve experimento cuja finalidade era extrair de documentos suas palavras-chave, utilizando para tal um modelo baseado na arquitetura de *Transformers,* o BERT. O experimento foi implementado em Python e seguiu os seguintes passos:

1. Divisão do conjunto – Uma tarefa comum em experimentos de Aprendizado de Máquina é dividir o conjunto de dados do experimento em diferentes subconjuntos, que serão utilizados para a avaliação de desempenho do modelo. Dos 74 artigos utilizados, 50 foram utilizados como conjunto de treinamento e os 14 restantes como conjunto de teste.

2. Pré-processamento – Os artigos passaram por um processo de *tokenização* **(**BIRD; KLEIN; LOPER, 2019), que consiste em separar os documentos e frases em função de seus termos. Cada documento passou a ser representado como uma lista de *tokens*. Em seguida, todos os elementos das listas foram normalizados em letras minúsculas. Por fim, o tamanho de todas as listas foi igualado (*padded*) para corresponder ao tamanho do maior documento, consequentemente, da lista mais longa (75 palavras), o que é essencial quando se trabalha com modelos de *Deep Learning*.

3. Calibragem e treinamento do modelo – Nesse passo o modelo pré-treinado BERT-Base foi utilizado. Por se tratar de um modelo que permite a transferência de aprendizado (*transfer learning*), realizou-se o procedimento de calibragem do modelo em nosso conjunto de dados de treinamento. Esse passo adequa o modelo ao domínio do problema, sem perder a informação obtida num prévio treinamento em larga escala. Os hiper-parâmetros usados podem ser encontrados na Tabela 1. Por se tratar de um processo de calibragem (*fine-tuning*), foi também utilizado um hiper-parâmetro específico para essa tarefa: uma taxa de declínio dos pesos (*weight decay rate*) de 0.01.

Otimizador	Batch-size	Taxa de aprendizado	Epochs
Adam	32	0.00003	4

Tabela 1 - Hiperparâmetros utilizados na calibragem do modelo.

4. Avaliação – Finalizado o treinamento, avaliamos o desempenho do modelo tanto quantitativamente – através das métricas de custo (*loss*), acurácia e F1-Score (GOODFELLOW; BENGIO; COURVILLE, 2016) – quanto qualitativamente, por meio da extração das palavras-chave de textos do conjunto de treinamento.

4. RESULTADOS E DISCUSSÃO

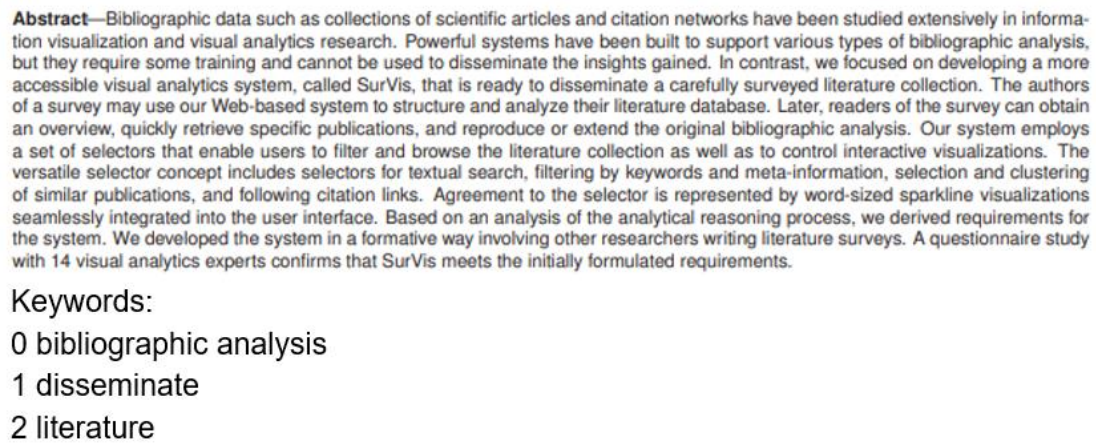
Como resultado do experimento com *Transformers* descrito na metodologia, obtivemos os seguintes resultados no conjunto de teste para as métricas de custo, acurácia e F1-Score:

Loss (custo): 0.102533

Acurácia: 0.975073

F1-Score: 0.466730

Pelo fato de tais valores corresponderem ao desempenho do modelo em predizer palavras faltantes em uma sentença, nota-se que tais métricas não refletem a eficácia do modelo na tarefa de extração de palavras chave dos textos do conjunto. Como solução para tal, realizamos uma análise qualitativa do resultado da classificação dos documentos em função de seus termos componentes. Ao fornecer um abstract de um artigo do conjunto, esperávamos que o experimento fosse capaz de extrair do mesmo os termos que melhor descrevessem seu conteúdo. Na Figura 2 mostramos uma lista de termos extraídos ao utilizar um dos abstracts do conjunto como entrada.



Abstract—Bibliographic data such as collections of scientific articles and citation networks have been studied extensively in information visualization and visual analytics research. Powerful systems have been built to support various types of bibliographic analysis, but they require some training and cannot be used to disseminate the insights gained. In contrast, we focused on developing a more accessible visual analytics system, called SurVis, that is ready to disseminate a carefully surveyed literature collection. The authors of a survey may use our Web-based system to structure and analyze their literature database. Later, readers of the survey can obtain an overview, quickly retrieve specific publications, and reproduce or extend the original bibliographic analysis. Our system employs a set of selectors that enable users to filter and browse the literature collection as well as to control interactive visualizations. The versatile selector concept includes selectors for textual search, filtering by keywords and meta-information, selection and clustering of similar publications, and following citation links. Agreement to the selector is represented by word-sized sparkline visualizations seamlessly integrated into the user interface. Based on an analysis of the analytical reasoning process, we derived requirements for the system. We developed the system in a formative way involving other researchers writing literature surveys. A questionnaire study with 14 visual analytics experts confirms that SurVis meets the initially formulated requirements.

Keywords:
0 bibliographic analysis
1 disseminate
2 literature

Figura 2. Como entrada, o abstract do trabalho de Beck e Weiskopf (2016) foi utilizado.

Ao analisar o artigo de entrada na íntegra, nota-se que as palavras-chave extraídas pelo modelo refletem de modo satisfatório o conteúdo do documento. Tais palavras-chave podem ser então utilizadas como representações do documento, servindo como entrada para algoritmos de agrupamento que poderiam delimitar subconjuntos temáticos dentro de um conjunto de artigos, ou ainda tais representações poderiam ser projetadas em visualizações que permitiriam ao pesquisador, por exemplo, entender a similaridade entre os documentos do conjunto e as palavras-chave mais comuns.

5. CONSIDERAÇÕES FINAIS

Neste trabalho, argumentamos a respeito do esforço que os pesquisadores exercem nas tarefas de análise bibliográfica e propomos o desenvolvimento de uma ferramenta que auxiliaria o pesquisador nesse trabalho. Exploramos e apresentamos trabalhos nos campos de representação textual, agrupamento e visualização, extraindo deles técnicas, conceitos e práticas que poderiam vir a integrar tal ferramenta ou serem de grande utilidade para o desenvolvimento dela.

Dentro do campo de representação textual, focamos nas técnicas de processamento de linguagem natural que são baseadas em redes neurais profundas, e mais especificamente, desenvolvemos um experimento com uma das arquiteturas do estado da arte (BERT) e mostramos que tal modelo poderia ser um componente essencial da ferramenta de visualização ao extrair de documentos as palavras-chave que melhor definissem seu conteúdo.

Apresentamos a literatura que demonstra como o campo de agrupamento poderia ser útil como processamento subsequente à representação textual, como forma de demostrar a relação entre documentos a partir de suas representações vetoriais contínuas, agrupando documentos similares.

Pelo fato de o escopo deste trabalho ser fundamentalmente teórico, como seguimento ao mesmo propomos a implementação prática da ferramenta de visualização utilizando os elementos e conceitos aqui explorados.

6. REFERÊNCIAS

AFONSO, Alexandre Ribeiro; DUQUE, Cláudio Gottschalg. Automated Text Clustering of Newspaper and Scientific Texts in Brazilian Portuguese: Analysis and comparison of Methods. Journal of Information Systems and Technology Management (JISTEM), v. 11, n. 2, p. 415-436, 2014.

BAHDANAU, Dzmitry; CHO, Kyunghyun; BENGIO, Yoshua. Neural machine translation by jointly learning to align and translate. 3rd. International Conference on Learning Representations (ICLR 2015), 2015, San Diego. 2015.

BECK, Fabian; KOCH, Sebastian; WEISKOPF, Daniel. Visual Analysis and Dissemination of Scientific Literature Collections with SurVis. IEE Transactions on Visualization and Computer Graphics, v. 22, n. 1, p. 180-189, 2016.

BELTAGY, Iz; LO, Kyle; COHAN, Arman. SCIBERT: A Pretrained Language Model for Scientific Text. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, p. 3615-3620, 2019.

BIRD, Steven; KLEIN, Ewan; LOPER, Edward. Natural Language Processing with Python – Analyzing Text with the Natural Language Toolkit. 2019. Disponível em: https://www.nltk.org/book/. Acesso em 21 de julho de 2020.

COLLOBERT, R. et al. Natural language processing (almost) from scratch. Journal of Machine Learning Research 12, p. 2493–2537, 2011.

DEVLIN et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 2018.

GOODFELLOW, Ian; BENGIO, Yoshua; COURVILLE, Aron. Deep Learning. MIT Press, 2016.

GRAVES, Alex; MOHAMED, Abdel-rahman; HINTON, Geoffrey. Speech recognition with deep recurrent neural networks. International Conference on Acoustics, Speech and Signal Processing (ICASSP 2013), p. 6645–6649, Vancouver, 2013.

HARRIS, Zellig. Distributional Structure. WORD, v. 10 n. 2-3, p. 146-162, 1954.

KRIZHEVSKY, Alex; SUTSKEVER, Ilya; HINTON, Geoffrey. ImageNet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems 25, p. 1106–1114, 2012.

KUCHER, Kostiantyn; KERREN, Andreas. Text Visualization Techniques: Taxonomy, Visual Survey, and Community Insights. Proceedings of the 8th IEEE Pacific Visualization Symposium (PacificVis ‘15), p. 117-121, 2015.

LUHN, Hans Peter. The Automatic Creation of Literature Abstracts. IBM Journal. p.159-165. Abril, 1958.

MENDELEY. Mendeley - Reference Management Software & Researcher Network. Disponível em: https://www.mendeley.com/?interaction_required=true. Acesso em 21 de julho de 2020.

MIKOLOV, Tomas; SUSTKEVER, Ilya; CHEN, Kai; CORRADO, Greg; DEAN, Jeffrey. Distributed representations of words and phrases and their compositionality. Advances in Neural Information Processing Systems (NIPS) 26, p. 3111–3119, Lake Tahoe, 2013.

MITCHELL, Tom. Machine Learning. New York: McGraw-Hill, 1997

OSGOOD, Charles E.; SUCI, George J.; TANNENBAUM, Percy H. The Measurement of Meaning. University of Illinois Press, 1957.

RASCHKA, Sebastian. Python Machine Learning. Mumbai: Packt Publishing, 2015.

SAMUEL, Arthur L. Some Studies in Machine Learning Using the Game of Checkers. IBM Journal of Research and Development, v. 3, n. 3, p. 210-229, 1959.

SPARCK, Karen Jones. A Statistical Interpretation of Term Specificity and its Application in Retrieval. Journal of Documentation, v. 28, n. 1, p. 11-21, 1972.

SUTSKEVER, Ilya; MARTENS, James; HINTON, Geoffrey. Generating text with recurrent neural networks. 28th. International Conference on Machine Learning, p. 1017-1024, Bellevue, 2011.

TIAN, Han; ZHUO, Hankz Hankui; Paper2vec: Citation-Context Based Document Distributed Representation for Scholar Recommendation. 2017. Disponível em: https://arxiv.org/abs/1703.06587. Acessado em 19/06/2020.

VASWANI et al. Attention is All You Need. 31st Conference on Neural Information Processing Systems (NIPS), 2017.

ZHANG, Yiding; JI, Xiaonan; IBARAKI, Motomu; SCHWARTZ, Franklin W. Mining Information from Collections of Papers: Illustrative Analysis of Groundwater and Disease. Groundwater v. 56(6), p. 993-1001, 2018.

ZOTERO. Zotero is a free, easy-to-use tool to help you collect, organize, cite, and share research. Disponível em: https://www.zotero.org/. Acesso em 21 de julho de 2020.

Contatos: lucas_figueiredo0@outlook.com; orlandoc@mackenzie.br