Análise de documentos científicos utilizando Mapas Auto-Organizáveis

Mary A. Casara[1], **Pollyana C. S. Notargiacomo**[1], **Leandro A. Silva**[1]

[1]Instituto Presbiteriano Mackenzie (IPM)
Rua da Consolação, 930 – Consolação, São Paulo – SP – Brazil

adriana.casara@hotmail.com, pollynot@gmail.com, leandroaugusto.silva@mackenzie.br

Abstract. *The quality of the scientific production of the Brazilian Graduate Programs, represented by the theses and dissertations produced by its student body, is one of the parameters used by government entities and research funding agencies to allocate financial resources to institutions involved with relevant research for the scientific development of the country. This paper aims to analyze these documents using text mining techniques and artificial neural networks to identify patterns, correlations and predominant themes, results that may help in the understanding of such production and, consequently, in the decisions of the aforementioned agencies.*

Resumo. *A qualidade da produção científica dos Programas de Pós-Graduação do Brasil, representadas pelas teses e dissertações produzidas pelo seu corpo discente, é um dos parâmetros utilizados por entidades governamentais e agências de fomento à pesquisa para destinar recursos financeiros a instituições envolvidas com pesquisas consideradas relevantes para o desenvolvimento científico dos país. Esse trabalho tem como objetivo analisar esses documentos utilizando técnicas de mineração de texto e redes neurais artificiais para identificar padrões, correlações e temas predominantes, resultados que podem auxiliar no entendimento de tal produção e, consequentemente, nas decisões das agências mencionadas.*

1. Introdução

A produção científica dos Programas de Pós-Graduação do Brasil (PPGs) se manifesta principalmente na forma de teses e dissertações geradas pelo seu corpo discente. No último quadriênio (2013 a 2016) os PPGs produziram, de acordo com a CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior), 294.364 teses e dissertações, sendo que a maioria está disponível para consulta no Catálogo de Teses e Dissertações (CT), base de dados pública e online mantida pela CAPES desde 2002.

A qualidade da produção acadêmico-científica tem sido foco de interesse de entidades governamentais e agências de fomento à pesquisa, que buscam destinar recursos financeiros a instituições ou indivíduos que demonstram capacidade de produzir resultados que contribuem para o desenvolvimento científico e impactem positivamente a comunidade regional.

A análise dos documentos acima citados também pode revelar padrões, tendências e correlações entre os temas abordados, elementos que podem ser utilizados para direcionar verbas de fomento por parte de entidades interessadas em linhas de pesquisa es-

DOI: 10.5753/cbie.wcbie.2019.99

pecíficas, como podem também revelar possíveis grupos de colaboração entre Instituições e pesquisadores.

A quantidade de documentos acumulados inviabiliza a leitura convencional de cada um deles. Por isso, o uso de técnicas computacionais, tais como mineração de textos e redes neurais artificiais tornaram-se primordiais para a extração de conhecimento a partir de tais fontes [Delen and Crossland 2008, Bakus et al. 2002, Blei and A.Y. Ng 2003].

A partir dessa contextualização, o objetivo desse trabalho é analisar teses e dissertações utilizando as técnicas computacionais acima citadas, para identificar padrões, correlações e temas predominantes, ou seja, para produzir um panorama sobre a produção científica dos PPGs do Brasil no período de 2013 a 2016.

Por limitações de escopo do trabalho as técnicas mencionadas acima foram aplicadas à uma única Grande Área de Conhecimento, denominada "Ciências Exatas e da Terra". Porém, acredita-se que o escopo da análise pode ser expandido ou substituído para abranger outras áreas.

Tendo em vista o que foi colocado anteriormente, esse artigo está organizado como segue: esta introdução explica o contexto da pesquisa e seu objetivo. A seção dois descreve os conceitos e tecnologias que embasaram o projeto. A seção três detalha os procedimentos, materiais e métodos utilizados na execução da pesquisa com o objetivo de permitir sua reprodutibilidade. A seção quatro descreve os experimentos realizados e resultados obtidos e, por fim, a seção cinco apresenta as conclusões parciais obtidas até a presente data.

2. Referencial teórico

2.1. Mineração de textos

Para realizar a análise que esse trabalho se propôs, ou seja, a análise do conteúdo das teses e dissertações escritas no período de 2013 a 2016, foi necessária a utilização de técnicas de mineração de textos para a extração de conhecimento a partir de dados não-estruturados do tipo texto [Souza et al. 2018].

Mineração de textos é o processo automatizado de descoberta de novo conhecimento por meio da extração de informação de fontes não estruturadas, como é o caso de documentos. Mineração de textos é uma variação da mineração de dados e é também conhecida como *Intelligent Text Analysis*, *Text Data Mining* ou *Knowledge-Discovery in Text*. A mineração de texto se baseia em técnicas emprestadas de outras áreas, tais como *Information Retrieval*, *Machine Learning*, estatística e linguística computacional [Gupta and Lehal 2009].

Para serem submetidos a algoritmos de mineração de texto os documentos, que são basicamente um conjunto de caracteres, devem ser transformados em uma representação estruturada compatível com tais algoritmos. Um tipo de representação comumente utilizada é o Modelo Espaço Vetorial (*Vector-Space Model*). Neste modelo, um texto é representado por um vetor cujos elementos representam a ocorrência de palavras dentro do texto [Miner et al. 2012].

A transformação de textos em representações estruturadas é obtida em uma etapa de pré-processamento que, de acordo com [Miner et al. 2012], é composta por atividades

de organização e limpeza que iniciam com a escolha do escopo de textos a serem processados e passam por etapas que promovem uma uniformização dos termos que serão analisados.

A escolha do escopo pode não ser relevante quando o texto inteiro necessita ser analisado, como é o caso de *e-mails* e mensagens de texto. Porém, para documentos extensos, pode-se pensar em utilizar apenas algumas seções ou parágrafos. A definição do escopo depende da aplicação que se pretende desenvolver. O conjunto de documentos selecionados para análise é denominado *corpus*.

A partir do escopo selecionado, realiza-se o processo de *tokenization* que consiste na separação do texto em palavras, também denominadas *tokens*. A forma mais comum de obter essa separação é utilizando as pontuações e espaços em branco como separadores.

A seguir realiza-se uma das atividades mais comuns da etapa de pré-processamento dos dados, que é a eliminação de *stopwords*. *Stopwords* são, em geral, artigos, preposições, pronomes e outros que dificilmente contribuem para diferenciação e análise de um texto. A eliminação de tais palavras pode contribuir para reduzir o espaço de armazenamento e para aumentar a velocidade de processamento das etapas posteriores.

Outra atividade importante de pré-processamento é o *stemming*, que consiste na redução das palavras ao seu radical (*stem*). Isso ocorre pela eliminação de sufixos, prefixos e plural. Como resultado, as palavras que possuem o mesmo radical são normalmente tratadas como sendo um único termo, o que ajuda na redução de dimensionalidade e, consequentemente, na melhora de performance dos algoritmos de classificação, agrupamento, entre outros.

A normalização ortográfica também é bastante utilizada pois os erros de ortografia podem aumentar significativamente o vetor de representação de texto, já que uma palavra erradamente grafada será entendida como um termo independente. A correção automática de ortografia, quer seja com o auxílio de dicionários ou pelo uso de outras técnicas, é fundamental quando a fonte de dados é informal, como é o caso de mensagens obtidas a partir de redes sociais.

A detecção de sentenças, que implica na segmentação de textos em frases, normalmente realizada com base na pontuação e a conversão em maiúsculas ou minúsculas também são atividades normalmente realizadas para a uniformização de termos.

A execução das atividades de pré-processamento acima citadas resulta em um conjunto de dados estruturados, consistentes e que contém os termos mais relevantes dos documentos selecionados. Porém, como nem todo termo é igualmente importante dentro de um documento, faz-se necessário executar mais uma etapa de preparação que é o cálculo do peso dos termos.

As formas mais comuns de cálculo de peso são baseadas em cálculos simples de frequência: frequência absoluta, frequência relativa e frequência inversa de documentos [Morais and Ambrósio 2007].

A frequência absoluta é também conhecida como TF (*Term Frequency*). $TF(w_i, x)$ é o número de vezes que o termo w_i aparece no documento x.

A frequência relativa considera o tamanho do documento para normalizar o peso

de um determinado termo. Assim, a frequência relativa de um termo é obtida dividindo-se a sua frequência absoluta pela número total de palavras desse documento N.

$$F_{rel} = \frac{TF}{N} \tag{1}$$

Quando aplicada a uma coleção de documentos, a métrica TF atribui maior valor aos termos que aparecem frequentemente em vários documentos. Estudos mostraram que o uso da métrica TF/IDF (*Term Frequency/Inverse Term Frequency*) pode melhorar a performance dos algoritmos de classificação e os resultados encontrados [Salton and Buckley 1988].

O IDF é calculado como o logaritmo da razão do número de documentos de um corpus (n) pelo número de documentos que contêm determinado termo $(DF(w_i))$.

$$IDF(w_i) = log\Big(\frac{n}{DF(w_i)}\Big) \tag{2}$$

A combinação TF/IDF atribui maior importância a palavras que são frequentes em um documento, mas que são raras dentro do *corpus*.

A representação de textos gerada a partir das etapas de pré-processamento e cálculo de pesos mencionadas acima pode ser utilização em diversas aplicações. De acordo com [Aggarwal and Zhai 2012], são exemplos de aplicações e algoritmos de mineração de textos a sumarização, o agrupamento ou *clustering* e a categorização.

A sumarização consiste na geração de resumos pela redução do tamanho e dos detalhes de um documento, ao mesmo tempo em que mantém seu significado geral e pontos principais.

O agrupamento ou *clustering* é um processo de classificação não-supervisionada. Os métodos de aprendizagem não-supervisionados são aqueles que não fazem uso de bases de dados previamente rotuladas ou "treinadas" e, portanto, não necessitam de esforço manual de preparação. A clusterização tem como objetivo agrupar documentos que são semelhantes entres si sem que haja um conjunto pré-definido de categorias [Delen and Crossland 2008].

Já a categorização é um processo de classificação supervisionada. Pela categorização, os temas principais de um documento são identificados e, como resultado, o documento é assinalado a uma das categorias de um conjunto de categorias previamente definido [Delen and Crossland 2008]. Os métodos da aprendizagem supervisionados baseiam-se em bases de dados previamente treinadas, ou seja, onde cada registro é marcado com um rótulo, para gerar um "classificador"ou função de regressão que podem ser utilizadas para fazer previsões sobre novos registros.

Esse trabalho utilizou-se mapas auto-organizáveis para realizar o agrupamento, ou clusterização, das teses e dissertações que são o foco da análise proposta.

2.2. Redes neurais artificiais e mapas auto-organizáveis

Redes neurais artificiais (RNAs) são algoritmos computacionais projetados para modelar a forma como o cérebro humano processa informações ou realiza uma tarefa particular.

As RNAs se assemelham a um cérebro humano no sentido de que adquirem conhecimento através de um processo de aprendizagem e utilizam conexões (sinapses) para armazenar o conhecimento adquirido [Haykin 1999].

Nesse trabalho utilizou-se mapas auto-organizáveis (*Self-Organizing Maps*, ou SOM) que é uma RNA com capacidade de organizar, normalmente de forma bidimensional, dados complexos em grupos (*clusters*), de acordo com as suas relações, de forma que os objetos similares sejam posicionados próximos uns aos outros [Kohonen 1990].

Os agrupamentos (*clusters*) produzidos por um SOM possuem algumas características desejáveis para o agrupamento de documentos. Uma delas é a de ser adaptativa, ou seja, não é necessário fornecer à rede parâmetros de tamanho e sobreposição, pois ela consegue se adaptar à base de dados em questão, o que significa que ela aprende a partir dos documentos com os quais ela tem que lidar. Os documentos submetidos a uma RNA do tipo SOM estão sujeitos a uma comparação geral que permite a inclusão de documentos que não possuem exatamente os mesmos termos em um mesmo *cluster* [Bote et al. 2002].

Espera-se, como resultado da aplicação de uma rede do tipo SOM na coleção de documentos composta pelas teses e dissertações, que sejam gerados agrupamentos pela similaridade de conteúdos, e não somente pela ocorrência de termos em comum. Com isso, deseja-se avaliar se as características dos documentos de entrada, tais como ano da defesa da tese/dissertação ou Região, são relevantes para os agrupamentos gerados.

3. Materiais e métodos

O trabalho foi implementado com a linguagem de programação `R`, que é voltada para implementação de modelos estatísticos avançados e que possui diversas funções relacionadas à ciência de dados [Prajapati 2013]. O *script* `R` foi gerado com a utilização do IDE (*Integrated Development Environment*) do `R-Studio`. A metodologia utilizada neste projeto é composta de 4 etapas, conforme ilustra a Figura 1

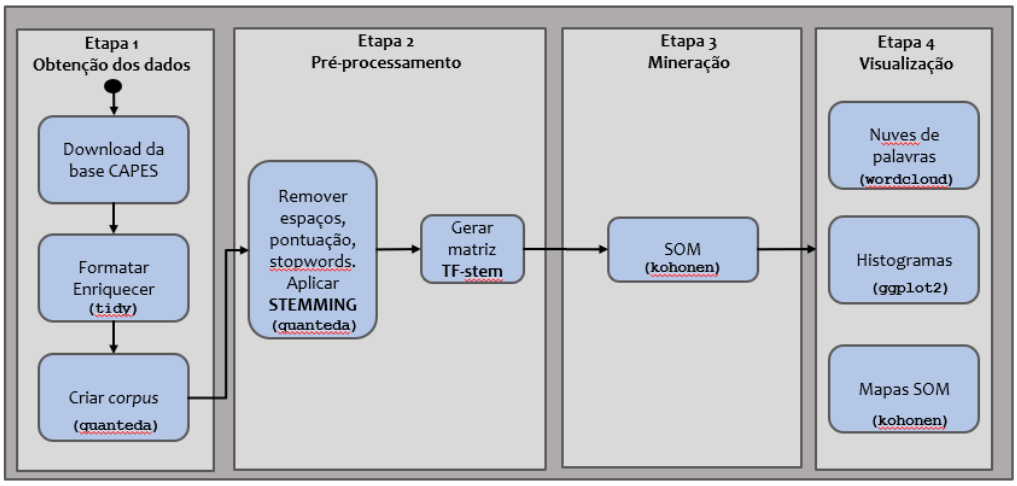

Figura 1. Metodologia de implementação

Etapa 1 – Obtenção do dados

Nessa etapa foram obtidos, a partir do Catálogo de Teses da CAPES, arquivos no formato `csv` contendo dados sobre as teses e dissertações dos programas da Grande Área de Conhecimento "Ciências Exatas e da Terra" (código 10000003) geradas entre janeiro de 2013 e dezembro de 2016.

A cada registro obtido do CT foi agregada a nota do Programa de Pós-Graduação, tal como divulgada na última avaliação quadrienal da CAPES de 2017. A notas foram obtidas a partir do site dos Resultados da Avaliação Quadrienal, também em formato `csv`.

Os pacotes `R tidyverse` (*Easily Install and Load the "Tidyverse"*) [Wickham 2017] e `tidytext` (*Text Mining using "dplyr", "ggplot2", and Other Tidy Tools*) [Silge and Robinson 2016] foram utilizados para a manipulação dos dados e dos textos.

O último passo dessa etapa foi a geração do *corpus*, que é um conjunto estruturado de textos em formato eletrônico, para ser utilizado nas etapas posteriores do trabalho. Esse passo foi realizado com o pacote `R` denominado `quanteda` (*Quantitative Analysis of Textual Data*) [Benoit 2018], que também viabilizou a limpeza e geração das representações (matrizes) necessárias para alimentar a rede SOM, realizadas na etapa seguinte de pré-processamento.

Etapa 2 – Pré-processamento

Nessa etapa ocorreu o tratamento ou "limpeza"dos textos para obtenção dos termos mais relevantes com potencial de influenciar os resultados das análises posteriores. Foi realizada a remoção de espaços, pontuação e *stopwords* e aplicação de *stemming*.

A seguir foi gerada uma matriz de representação da coleção de documentos, utilizando a frequência absoluto (TF) como peso.

O processamento descrito acima resultou em uma matriz esparsa, pois muitos dos termos aparecem em apenas um ou em poucos documentos. Matrizes com essa característica impactam a performance das etapas posteriores ou, às vezes, as inviabilizam. Por isso, como passo final da etapa de pré-processamento, foram excluídos os termos com frequência absoluta menor do que 50.

Etapa 3 – Mineração de Textos

Foi gerado um mapa SOM com base na matriz obtida na etapa anterior. Foram utilizados 5 x 5, 10 x 10 e 15 x 15, como parâmetros para o tamanho do *grid*, sendo que o último resultou em agrupamentos menos adensados, o que facilitou as análises subsequentes. Também utilizou-se o formato hexagonal para o *grid* para possibilitar uma análise mais abrangente da vizinhança dos neurônios selecionados.

A geração do SOM foi realizada com o auxílio do pacote `R kohonen` (*Supervised and Unsupervised Self-Organising Maps*) [Wehrens and Buydens 2007].

Etapa 4 – Visualização

Por fim, na etapa 4, foram gerados diferentes tipos de visualização com o intuito de apoiar as análises dos resultados obtidos nas etapas anteriores.

Com o auxílio do pacote `R ggplot2` (*Create Elegant Data Visualisations Using*

the Grammar of Graphics) [Wickham 2016] foram gerados histogramas que suportaram análises quantitativas da coleção de documentos como, por exemplo, a distribuição de documentos por ano, Região e nota.

Nuvens de palavras foram utilizadas para exibir os termos mais relevantes de todas a coleção de documentos e de neurônios selecionados de uma vizinhança do SOM. Essa visualização foi gerada com o pacote `R wordcloud` (*Word Cloud*) [Fellows 2018].

De acordo com [Vesanto 1999], a visualização de dados complexos multidimensionais é uma das principais aplicações de SOM, pois este implementa um mapa ordenado de dimensionalidade reduzida dos dados de treinamento. Diversos tipos de plotagem de SOM foram utilizados para análise dos agrupamentos gerados como, por exemplo, o mapa de contagem de elementos por neurônio (*counts plot*) e o mapa de distâncias entre vizinhos (*U-matrix*).

4. Resultados obtidos

A seguir serão apresentados os resultados dos experimentos realizados para cada uma das etapas da metodologia utilizada.

A etapa 1 resultou em uma coleção de 27.123 documentos. Após a eliminação de documentos escritos em idiomas diferentes de Português (1.461 documentos) e de documentos com identificação duplicada (2 documentos), a coleção final resumiu-se em 25.660 documentos.

A cada documento foi adicionada a nota recebida pelo seu PPG na avaliação quadrienal da CAPES. Para efeito de simplificação, as notas de 1 a 4 foram agrupadas com o rótulo “até 4”.

Para explorar possíveis similaridades de temas ou outros padrões subjacentes, prosseguiu-se com a geração do *corpus*, que serviu de entrada para as próximas etapas de remoção da pontuação, conversão dos caracteres em minúsculos, remoção de caracteres numéricos e *stemming*.

A seguir foi gerada a matriz de documentos e palavras, que na denominação do pacote `quanteda`, é chamado de DFM (*Document-Term Matrix*). Como a DFM inicialmente gerada continha muitos termos de frequência muito baixa, optou-se pela eliminação dos termos com frequência inferior a 50. A DFM resultante passou a ter 25.660 documentos e 5.612 termos. Essas dimensões, tornaram viável a execução das próximas etapas.

Com a DFM gerada, o experimento prosseguiu para a etapa de mineração, com a geração do SOM. Optou-se pelo *grid* de 15 x 15 pois este apresentou uma distribuição com menos neurônios vazios do que o *grid* de 20 x 20, ao mesmo tempo em que apresentou neurônios menos densos do que o *grid* de 10 x 10.

Analisou-se a distribuição das variáveis de interesse – ano de defesa, nota do PPG, Região e área de conhecimento – no SOM gerado. Em particular, a distribuição das áreas de conhecimento demonstrou a formação de *clusters*, como mostra a figura 2 .

Nota-se que existem áreas do mapa onde claramente prevalecem documentos de uma determinada área de conhecimento, como é o caso do canto superior esquerdo, composto em sua maioria por documentos de Química.

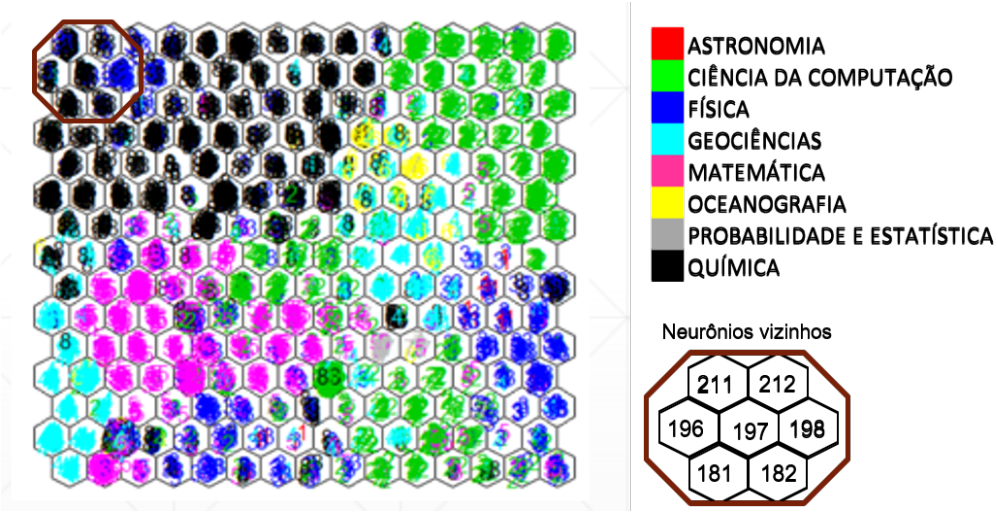

Figura 2. Distribuição das Áreas de Conhecimento no SOM

Com o intuito de explorar a similaridade entre os documentos agrupados pelo SOM, foram examinados os termos mais frequentes utilizados nos documentos mapeados em neurônios vizinhos. A seleção dos neurônios foi feita com base nas distâncias entre eles, ou seja, foram escolhidos neurônios que apresentaram distâncias menores, em comparação aos demais. A figura 2 destaca a vizinhança selecionada e a figura 3 exibe os termos mais frequentes dos neurônios dessa vizinhança. Vale lembrar que os neurônios de um SOM são identificados por um número sequencial, sendo o primeiro localizado na posição mais abaixo e à esquerda e os demais contados da esquerda para a direita e de baixo para cima.

É possível observar na figura 3 que alguns termos, tais como “nanopartícula”e “espectroscopia”, aparecem em mais de um neurônio, o que indica que existe similaridade entre seus conteúdos. Também é possível relacionar esses termos à área predominante na vizinhança, no caso à área de Química.

Por outro lado, alguns dos termos exibidos na figura 3 são comuns a qualquer área de conhecimento e não colaboram para o entendimento dos temas tratados pelo conjunto de documentos. Esse é o caso dos termos “amostra”, “complexo”e “trabalho”, para citar alguns exemplos.

5. Conclusões e Trabalhos Futuros

Este trabalho empregou técnicas de mineração de textos e redes neurais artificiais do tipo SOM para descobrir padrões na produção científica gerada pelos Programas de Pós-Graduação do Brasil, mais especificamente da Grande Área de Conhecimento “Ciências Exatas e da Terra”, no período de 2013 a 2016.

Os experimentos realizados até o momento demonstraram a aplicabilidade dessas técnicas ao problema em questão, pois foi possível identificar, por meio do SOM, agrupamentos por Área de Conhecimento e por tema, o que indica que é possível explorar outras similaridades entre os documentos mapeados em um mesmo *cluster* do mapa.

Os resultados obtidos com o SOM e análise dos termos mais frequentes de uma vizinhança (etapa 3) podem ser utilizados para aprimorar a etapa de pré-processamento

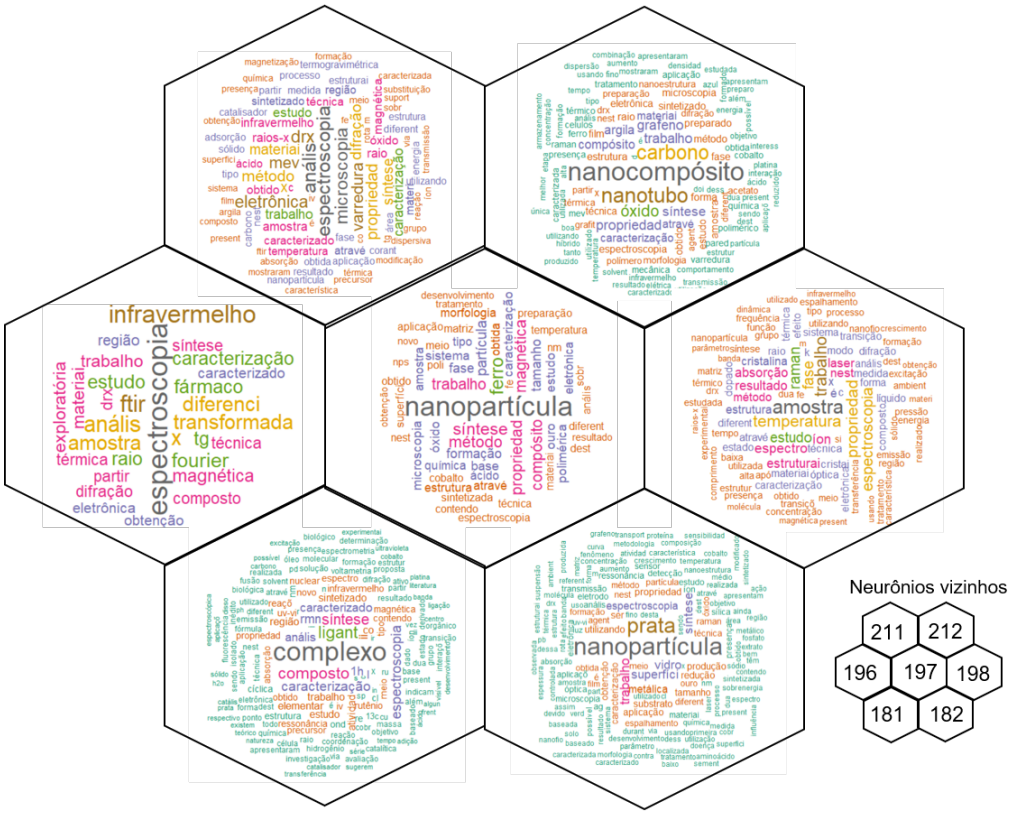

Figura 3. Termos mais frequentes dos neurônios vizinhos selecionados

(etapa 2), por meio da inclusão de termos muito comuns a todas as áreas de conhecimento na lista de *stopwords* a serem excluídas, sendo que essa retro-alimentação pode ocorrer de forma iterativa.

Em trabalhos posteriores, pretende-se explorar se as iterações mencionadas acima melhoram os agrupamentos subsequentes e ajudam a revelar similaridades entre os temas abordados.

Adicionalmente, pretende-se explorar a distribuição de todos documentos de um único ano e todos os documentos de uma única Região, sem restringir-se à Grande Área de Conhecimento “Ciências Exatas e da Terra”, para verificar se essa abordagem revela outros padrões relevantes para os objetivos desse trabalho.

Embora a baixa dimensionalidade das representações gráficas geradas pelos SOMs permitam a direta interpretação dos resultados, entende-se que a inspeção visual não é suficiente para determinar a qualidade dos mapas e, portanto, também serão aplicadas, em trabalhos futuros, técnicas de validação dos agrupamentos gerados. Com isso, pretende-se aumentar a confiabilidade das análises visuais subsequentes.

Referências

Aggarwal, C. C. and Zhai, C. X. (2012). *Mining Text Data*. Springer, 233 Spring Street, New York, NY 10013, USA.

Bakus, J., Hussin, M. F., and Kamel, M. (2002). A som-based documesaltonnt clustering using phrases. *Proceedings of the 9th International Conference on Neural Information*

Processing, 2002. ICONIP '02., 5:2212–2216.

Benoit, K. (2018). quanteda: Quantitative analysis of textual data. Online.

Blei, D. M. and A.Y. Ng, M. I. J. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Bote, V. P. G., Anegón, F. M., and Solana, V. H. (2002). Document organization using kohonen's algorithm. *Information Processing Management*, 38(1):79 – 89.

Delen, D. and Crossland, M. D. (2008). Seeding the survey and analysis of research literature with text mining. *Expert Systems with Applications*, 34(3):1707 – 1720.

Fellows, I. (2018). *wordcloud: Word Clouds*. R package version 2.6.

Gupta, V. and Lehal, G. S. (2009). A survey of text mining techniques and applications. *Journal of Emerging Technologies in Web intelligence*, 1:60–76.

Haykin, S. (1999). *Neural Networks and Learning Machines*. Pearson - Prentice Hall.

Kohonen, T. (1990). The self-organizing map. *Proceedings of the IEEE*, 78(9):1464–1480.

Miner, G., Elder, J., Fast, A., Hill, T., Nisbet, R., and Delen, D. (2012). *Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications*. Academic Press, 225 Wyman Street, Waltham, MA 02451, USA.

Morais, E. A. M. and Ambrósio, A. P. L. (2007). Mineração de textos. *Relatório Técnico do Instituto de Informática (UFG).*

Prajapati, V. (2013). *Big Data Analytics with R and Hadoop*. Packt Publishing Ltd, Birmingham, UK.

Salton, G. and Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing Management*, 24:513–523.

Silge, J. and Robinson, D. (2016). *tidytext: Text Mining and Analysis Using Tidy Data Principles in R*.

Souza, E., Costa, D., Castro, D. W., Vitório, D., Teles, I., Almeida, R., Alves, T., Oliveira, A. L. I., and Gusmão, C. (2018). Characterising text mining: a systematic mapping review of the portuguese language. *IET Software*, 12(2):49–75.

Vesanto, J. (1999). Som-based data visualization methods. *Intelligent Data Analysis*, 3(2):111–126.

Wehrens, R. and Buydens, L. (2007). *Self- and Super-organising Maps in R: the kohonen package*.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*.

Wickham, H. (2017). *tidyverse: Easily Install and Load the 'Tidyverse'*. R package version 1.2.1.