VII SEMANA UNIVERSITÁRIA DA URCA – XXV Semana de Iniciação Científica da URCA e VIII Semana de Extensão da URCA

12 a 16 de dezembro de 2022

Tema: "DIVULGAÇÃO CIENTÍFICA, INDEPENDÊNCIA E SOBERANIA NACIONAL"

ISSN 1983-8174

APLICAÇÃO DE MINERAÇÃO DE TEXTO EM ARTIGOS PUBLICADOS PELOS CURSOS DE ECONOMIA DA UNICAMP, FGV, URCA PARA CLASSIFICAÇÃO METODOLÓGICA

Julie Evelyn Alves Moreira[1], Yure Emanuel de Melo Feitosa Araújo[2]

Resumo: Este trabalho, fruto de bolsa pibic-fecop, tem por objetivo classificar metodologicamente artigos publicados pelos cursos de economia da Unicamp, FGV, URCA, através da aplicação de mineração de texto. Para tanto foi desenvolvido um banco de dados que consta as metodologias de 91 artigos das instituições abordadas no estudo, sendo 10 da FGV, 51 da Unicamp e 30 da URCA, o corpus foi processado no programa Iramuteq, com a finalidade de se obter uma classificação hierárquica descendente que apresentasse procedimentos metodológicos que foram aplicados nos artigos. Os resultados obtidos revelaram que apenas as instituições Unicamp e URCA apresentaram condições em seus artigos para uma classificação, consideramos ainda a possibilidade de sugestão de serem agregados nos futuros artigos a descrição dos procedimentos metodológicos.

Palavras-chave: Metodologia. Mineração de texto. Iramuteq.

1. Introdução

Com o avanço da tecnologia se tornou possível o armazenamento de grande volume de dados em formato digital, como livros, artigos, revistas, entre outros conteúdos. A partir disso surge o questionamento de como lidar com essa grande quantidade de dados, a fim de extrair informações que realmente seja de interesse de quem está consumido o conteúdo.

As três instituições de interesse deste estudo contam com grandes acervos de artigos, assim a realização de uma classificação metodológica em seus acervos surge como alternativa para melhorar a experiencia do usuário no processo de recuperação da informação.

Para Souza (2005, p.46), "quando o volume de documentos no acervo atinge certo patamar, ou quando os assuntos dos documentos adquirem certo grau de especificidade, não é mais possível ordená-los por meio de grandes classes de assuntos, pois essas classes não são mais suficientemente informativas para representar adequadamente o conteúdo dos documentos e discriminá-lo em relação a outros documentos. Torna-se necessário então utilizar processos de catalogação e de indexação eficazes, de forma que a recuperação das informações que contêm, de acordo com as necessidades dos usuários, seja a mais eficaz possível."

Assim, neste trabalho será realizada a mineração de texto para se alcançar o objetivo proposto. Para Andrade (2015, p.4), "a mineração de texto é o processo de descoberta do conhecimento que utiliza técnicas de análise de

1 Universidade Regional do Cariri, email: julie.evelyn@urca.br
2 Universidade Regional do Cariri, email: yure.araujo@urca.br

ISSN 1983-8174

extração de dados a partir de textos, frases ou palavras. É o processo de extrair padrões interessantes e não triviais ou conhecimento a partir de documentos em textos não estruturados. Essa descoberta de conhecimento envolve diversas aplicações tais como análise de textos, extração de informações, sumarização, classificação, agrupamento, linguística computacional, dentre outros."

Como um primeiro objetivo, foi realizada a obtenção de material para a composição do banco de dados, através do conteúdo encontrado em ambiente digital, para realização da extração do corpus que será analisado. A construção do corpus se deu com a retirada das partes que apresentavam a metodologia dos artigos e a aplicação das técnicas já citadas. Nele foram utilizados 91 artigos, dentre eles 10 da FGV, 51 da Unicamp e 30 da URCA. Para tanto foi realizada a análise de conteúdo que se deu por meio de uma leitura flutuante e em seguida foi realizada a separação do material que foi considerado pertinente para a pesquisa. O corpus foi processado no programa Iramuteq a fim de gerar uma classificação hierárquica descendente que represente o resultado da classificação metodológica.

2. **Objetivo**

O objetivo deste trabalho será classificar metodologicamente artigos publicados pelos cursos de economia da Unicamp, FGV, URCA através da aplicação de mineração de texto.

3. **Metodologia**

Este trabalho visa classificar metodologicamente artigos publicados pelos cursos de economia da Unicamp, FGV, URCA, através da aplicação de mineração de texto. Se trata de uma pesquisa exploratória com base em literatura e análise de dados, de natureza qualitativa.

Seguindo o processo necessário para realização da mineração de texto, a coleta de dados se deu por meio da seleção de artigos dos professores do departamento de economia das instituições alvo deste estudo, que são elas UNICAMP, FGV e URCA. Os artigos foram obtidos por meio de uma busca no site das instituições em que estava disponível a lista de professores atuantes e seus respectivos currículos lates, através desses foi possível identificar as obras e selecioná-las, seguindo o critério de que deveriam possuir a estrutura de um artigo e estarem disponíveis em língua portuguesa. Ao todo foram utilizados 91 artigos para compor o banco de dados da pesquisa, sendo eles 10 da FGV, 51 da Unicamp e 30 da URCA. As abordagens utilizadas foram a abordagem semântica e a abordagem estatística.

Na etapa de pré-processamento é realizada a normalização dos dados, para isso foi aplicada a técnica stemming através do Método de Poter, que consiste em reduzir um termo ao seu radical, eliminando assim variações morfológicas. O processo de indexação se deu na separação dos textos no desenvolvimento do corpus, onde as obras foram dispostas por seus respectivos títulos.

ISSN 1983-8174

O programa utilizado para processar os dados foi o Iramuteq, que se trata de um software ancorado no ambiente estatístico do software R e na linguagem Python. Para o processamento dos dados o corpus foi disposto em um bloco de notas e separado com uma identificação (**** *artigo_ seguido do número de correspondência do artigo), além de que foi elaborado um corpus para cada uma das instituições seguindo os procedimentos descritos anteriormente, ao passo de que também foram processados separadamente. Assim, o corpus foi inserido no programa e foram ajustadas as propriedades de acordo com a necessidade da pesquisa. Os resultados obtidos serão apresentados na seção seguinte.

4. **Resultados**

Após aplicação de mesmo procedimento nos artigos das três instituições, não foi possível obter os resultados desejados, como será apresentado a seguir. Optamos por apresentar os resultados em classificação hierárquica descendente, em que os segmentos de textos são divididos em função de seus respectivos vocabulários, e o conjunto deles é distribuído em função da frequência das formas reduzidas. A classificação não será apresentada em referência a cada artigo, mas fará referência a cada instituição, sendo disposta de maneira descrente, no que diz respeito a frequência com as quais aparecem no corpus. Assim, a maior frequência de um termo não implica que este aparece em todos os artigos, sendo possível uma análise e classificação geral dos artigos a partir do corpus.

Os resultados obtidos a partir dos artigos da Unicamp foram relativamente satisfatórios, pois apresentam na disposição das classes obtidas termos que não se referem aos procedimentos metodológicos seguidos, embora apresente grande quantidade de palavras não relacionadas ao tema. Como é possível observar após aplicação do método de classificação hierárquica descendente a classe 3 é a que apresenta maior quantidade de termos que se referem aos procedimentos metodológicos utilizados, sendo eles "exploratório", "bibliográfico" e "explicativo".

ISSN 1983-8174

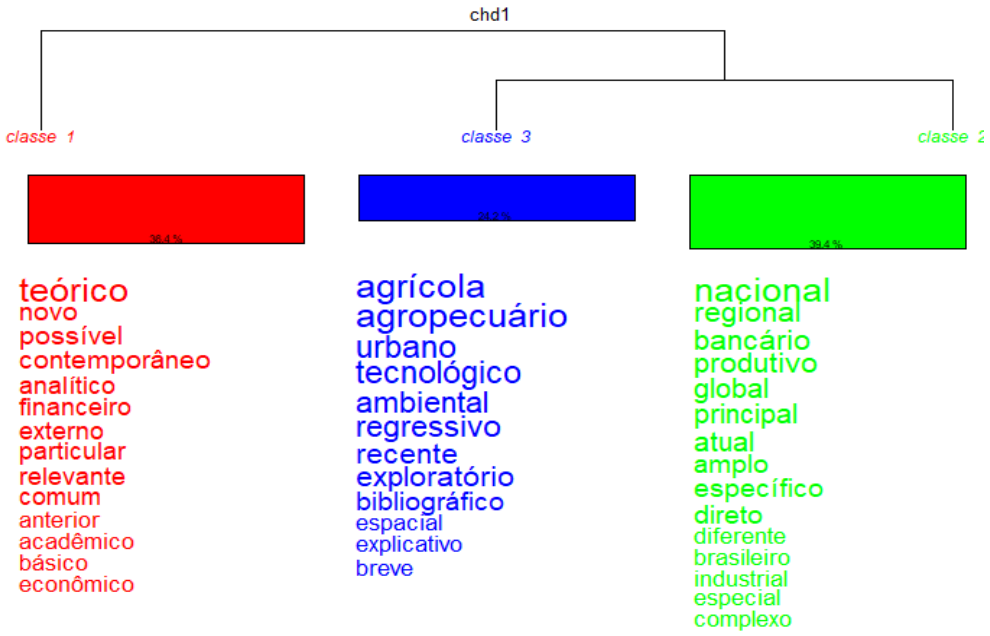

Figura 1 – Classificação Hierárquica Descendente Unicamp

Os resultados apresentados após a análise do corpus, que teve como base os artigos da instituição FGV, não se apresentaram satisfatórios, visto que não foi possível identificar aspectos metodológicos nas classes formadas após aplicação do método de classificação hierárquica descendente. Em parte, isso pode ter ocorrido devido os artigos não apresentarem claramente seus procedimentos metodológicos em sua construção, não podendo se afirmar se este é um padrão ou não da instituição para a publicação dos artigos. E em parte, também pode ter ocorrido devido o programa utilizado levar em consideração na construção de suas classes a frequência com que os termos aparecem, o que demonstra que este é importante no documento, o programa ainda possibilita excluir algumas categorias de palavras para um melhor resultado.

Os resultados mais expressivos foram observados no processamento do corpus da instituição URCA, como é possível observar nas classes apresentadas abaixo, há a presença de um considerável número de palavras de classificação metodológica. Como é possível observar a classe 1 é formada quase que completamente por procedimentos metodológicos, apresentados na seguinte ordem "bibliográfico", "documental", "exploratório", "qualitativo". Embora haja a presença de palavras que não fazem parte da descrição de procedimentos metodológicos, consideramos que estes resultados são significativos. Consideramos, ainda, que esses resultados foram possíveis devido ao grande número de artigos desta instituição possuir uma descrição precisa dos procedimentos metodológicos aplicados em cada artigo, e se reitera que isto facilitou o processamento dos dados por meio do programa utilizado.

ISSN 1983-8174

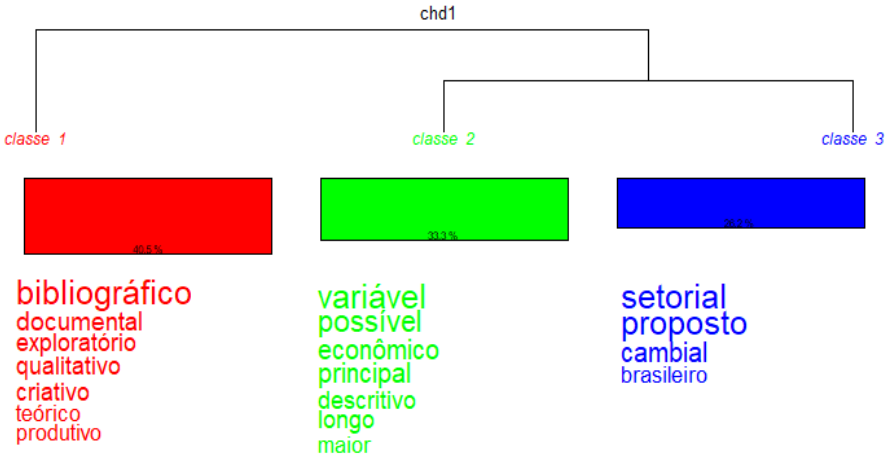

Figura 2 – Classificação Hierárquica Descendente Urca

5. Conclusão

Os resultados obtidos foram considerados parcialmente satisfatórios para instituição Unicamp, pouquíssimos satisfatórios para instituição FGV, e satisfatórios para instituição Urca. Por esta razão, consideramos que os objetivos foram apenas parcialmente alcançados, pois com a utilização deste programa foi possível gerar apenas uma classificação geral sobre as instituições, não sendo possível catalogar os artigos em suas determinadas classes.

Assim, destacamos dentre os desafios para este trabalho as limitações que encontramos sobre a disponibilidade de artigos em língua portuguesa pela FGV e a falta da descrição por parte dos autores dos artigos de como se realizaram os procedimentos metodológicos desta mesma instituição o que prejudicou a construção do corpus e dificultou a classificação por meio do programa utilizado.

Acreditamos haver muitas possibilidades de melhoria, especialmente com a inclusão da descrição dos procedimentos metodológicos utilizados para o desenvolvimento dos artigos, assim como a possibilidade de utilização de outros programas e algoritmos.

6. Referências

ANDRADE, P. H. M. A. Aplicação de técnicas de mineração de textos para classificação de documentos: um estudo da automatização da triagem de denúncias na CGU. Brasília, 2015. 65 p.

SOUZA, R. R. Uma proposta de metodologia para escolha automática de descritores utilizando sintagmas nominais. Belo Horizonte: ECI/UFMG, 2005.