Mineração de texto no auxílio a construção de referencial teórico

Natalia Souza Vieira[1]
Wagner Lourenzi Simões[2]

Resumo: De acordo com Faro, Giordano e Spampinato (2012, p. 62), o objetivo primordial da mineração de textos é a descoberta de conhecimento oculto em documentos de texto, com a subsequente apresentação desse conhecimento de maneira coerente e concisa. A mineração de texto pode ser considerada uma vertente específica da mineração de dados, focada na extração de conhecimento de conjuntos de dados textuais não estruturados, tais como artigos, notícias e outros documentos textuais. O propósito final é a estruturação dos dados e a identificação de padrões presentes nos textos processados. Isso não visa substituir a leitura dos documentos, mas sim servir como uma orientação para a leitura, direcionando aprofundamentos e permitindo a classificação dos documentos avaliados na pesquisa. Conforme Gajzer (2010, p. 223-224), o processo de mineração de textos compreende quatro estágios fundamentais: a) Transformação do Texto: Este estágio envolve a conversão dos documentos para um formato textual, eliminando símbolos desnecessários; b) Separação de Palavras (ou Tokenização): Trata-se de um método que identifica as características relevantes de um texto, segmentando o fluxo contínuo de caracteres e removendo elementos textuais como pontuação, separação de sílabas, marcações e números, que individualmente agregam pouco valor à informação.; c) *Stemming*: Neste estágio, a quantidade de tokens é reduzida pela extração de sufixos e prefixos que compõem cada token, resultando em uma "normalização linguística" que converte formas variantes de um termo em uma forma comum denominada "*stem*."; d) Matriz de Frequências: Conforme Fan et al. (2006, p. 79), esta etapa é responsável por categorizar os *stems* e associá-los às suas frequências de ocorrência nos textos analisados. Isso permite inferências sobre proximidade, distância, sinônimos e termos relacionados. Após esse processo de pré-processamento, o texto está pronto para a aplicação de técnicas de mineração, como análise de sentimentos, classificação de documentos, resumos automáticos e extração de informações, como autores, palavras-chave e termos emergentes. O presente estudo tem como objetivo o desenvolvimento de uma ferramenta de mineração de texto para a classificação de artigos relacionados ao varejo digital. Esse objetivo é alcançado por meio do uso de scripts em linguagem R. A finalidade principal é fornecer insights sobre os termos emergentes no campo do varejo digital. Os artigos, provenientes de diversas fontes e em diferentes formatos, são extraídos e submetidos a um processo que

[1] Discente do Curso de Graduação em Engenharia de Produção do Centro Universitário Cesuca. E-mail: natieandy@gmail.com
[2] Docente do Curso de Engenharia de Produção do Centro Universitário Cesuca. Doutor em Engenharia de Produção e Sistemas. E-mail: wagner.lourenzi@cesuca.edu.br

envolve a geração da matriz de frequência para identificar os termos mais citados em artigos que tratam do tema do varejo digital. Essa análise visa identificar termos que podem enriquecer o referencial teórico sobre o setor, especialmente aqueles que foram negligenciados nas buscas iniciais para a composição do referencial teórico do trabalho. Além disso, a realização de uma análise de agrupamento desses artigos pode contribuir para a identificação de tópicos que requerem aprofundamento ao longo do estudo. É importante ressaltar que a ferramenta em questão ainda está em processo de desenvolvimento. Este ferramental desempenhará um papel fundamental no refinamento da pesquisa sobre o varejo 4.0, que está sendo conduzida como parte do projeto. É relevante mencionar que a ferramenta está em fase de desenvolvimento contínuo.

Palavras-chaves: Mineração de textos; Classificação de textos; *Text Minning*.