Conceitos do Acesso Aberto a Dados de Pesquisa: identificação através da Mineração de Textos

Luciana Monteiro Krebs[1]

Rita do Carmo Ferreira Laipelt[2]

Resumo

Objetivo. O artigo tem como objetivo explorar o tema Acesso Aberto a Dados de Pesquisa e suas tendências, através da identificação dos principais termos presentes em corpus formado por artigos científicos.
Método. A metodologia escolhida é a mineração de textos, e foram utilizados os softwares WordCounter, TagCrowd, WordClouds, Voyant e Sobek para o processamento automático de dez artigos coletados para este fim.
Resultados. A análise permitiu visualizar as características descritivas do *corpus* (quanto ao tamanho dos textos, média e proporções), validação da eficácia da estratégia de busca utilizada, a descoberta dos principais termos formados por até três unidades lexicais (em relação à sua frequência e relevância no documento), além de similaridades e diferenciações feitas a partir dos termos e, por fim, a descoberta de novos termos não previstos, porém relacionados ao tema.
Conclusões. No decorrer da pesquisa foi possível identificar as características do *corpus* em termos de quantidade de palavras e proporção que cada artigo representa no *corpus*. Também foi possível explorar títulos e palavras-chave, identificando termos preferidos e variantes terminológicas. As descobertas a partir da mineração do texto integral concentraram-se nos termos mais frequentes do *corpus* com uma a três unidades lexicais, o que é útil para explorar palavras não previstas na busca inicial. Além disso, foram analisadas similaridades e distinções, o que permite entender e visualizar o que une os artigos (ou o que eles têm em comum), e o que cada artigo traz de novidade em relação aos demais, ou seja, o que os diferencia. Por fim, apresentam-se as correlações entre os termos que representam os conceitos no *corpus* analisado.

Palavras-chave

Gestão de dados de pesquisa; Acesso aberto; Compartilhamento de dados; Mineração de textos

Concepts of Open Access to Research Data: identification through Text Mining

Abstract

Objective. The paper aims to explore the domain Open Access to Research Data and its trends, by identifying the main terms present in a corpus formed by scientific articles.
Method. The methodology chosen is text mining, and WordCounter, TagCrowd, WordClouds, Voyant and Sobek software were used for automatic processing of ten articles collected for this purpose.
Results. The analysis allowed to visualize the descriptive characteristics of the corpus (regarding the size of texts and proportions), validation of the effectiveness of the search strategy used, the discovery of the main terms formed by up to three lexical units (in relation to their frequency and relevance in the document), as well as similarities and differentiations made from the terms and, finally, the discovery of new terms not foreseen, but related to the theme.
Conclusions. In the course of the research it was possible to identify the characteristics of the corpus in terms of the number of words and the proportion that each article represents in the corpus. It was also possible to explore titles and keywords, identifying preferred terms and terminological variants. The discoveries from the mining of the full text were concentrated in the most frequent terms of the corpus with one to three lexical units, which is useful to explore words not predicted in the initial search. In addition, similarities and distinctions were analyzed, which allows one to understand and visualize what unites the articles (or what they have in common), and what each article brings of novelty in relation to the others, that is, what differentiates them . Finally, we present the correlations between the terms that represent the concepts in the analyzed corpus.

Keywords

Research data management; Open access; Data management; Data sharing; Text mining.

[1] Doutoranda no Programa de Pós-Graduação em Comunicação e Informação da Universidade Federal do Rio Grande do Sul (UFRGS). luciana.monteiro@ufrgs.br
[2] Professora Adjunta do Departamento de Ciências da Informação da Faculdade de Biblioteconomia e Comunicação da Universidade Federal do Rio Grande do Sul (UFRGS). ritacarmo@yahoo.com.br