|  |  |  |  |
| --- | --- | --- | --- |
| Title | Evaluating keyphrase extraction algorithms for finding similar news articles using lexical similarity calculation and semantic relatedness measurement by word embedding. | | |
| Author(s) Name | Sarwar, T. B.; Noor, N. M.; and Miah, M. S. U | | |
| Contact Email(s) | saef@aiub.edu | | |
| Published Journal Name | Peerj CS | | |
| Type of Publication | Journal | | |
| Volume | 8 | Issue |  |
| Publisher | PEERJ | | |
| Publication Date | Jul 2022 | | |
| ISSN |  | | |
| DOI | http://doi.org/10.7717/peerj-cs.1024 | | |
| URL |  | | |
| Other Related Info. |  | | |
|  | | | |

|  |  |
| --- | --- |
| Abstract |  |
| A textual data processing task that involves the automatic extraction of relevant and salient keyphrases from a document that expresses all the important concepts of the document is called keyphrase extraction. Due to technological advancements, the amount of textual information on the Internet is rapidly increasing as a lot of textual information is processed online in various domains such as offices, news portals, or for research purposes. Given the exponential increase of news articles on the Internet, manually searching for similar news articles by reading the entire news content that matches the user’s interests has become a time-consuming and tedious task. Therefore, automatically finding similar news articles can be a significant task in text processing. In this context, keyphrase extraction algorithms can extract information from news articles. However, selecting the most appropriate algorithm is also a problem. Therefore, this study analyzes various supervised and unsupervised keyphrase extraction algorithms, namely KEA, KP-Miner, YAKE, MultipartiteRank, TopicRank, and TeKET, which are used to extract keyphrases from news articles. The extracted keyphrases are used to compute lexical and semantic similarity to find similar news articles. The lexical similarity is calculated using the Cosine and Jaccard similarity techniques. In addition, semantic similarity is calculated using a word embedding technique called Word2Vec in combination with the Cosine similarity measure. The experimental results show that the KP-Miner keyphrase extraction algorithm, together with the Cosine similarity calculation using Word2Vec (Cosine-Word2Vec), outperforms the other combinations of keyphrase extraction algorithms and similarity calculation techniques to find similar news articles. The similar articles identified using KPMiner and the Cosine similarity measure with Word2Vec appear to be relevant to a particular news article and thus show satisfactory performance with a Normalized Discounted Cumulative Gain (NDCG) value of 0.97. This study proposes a method for finding similar news articles that can be used in conjunction with other methods already in use. | |