NCON-PGR_2022_064

KDA: An Unsupervised Approach for Analyzing Keyphrases Distance from News Articles as a Feature of Keyphrase Extraction

Mohammad Badrul Alam Miah[1,3] *and* ***Suryanti Awang***[2]*

[1]Faculty of Computing, Universiti Malaysia Pahang, 26600, Pekan, Pahang, Malaysia

[2]Center of Excellence for Artificial Intelligence & Data Science, Universiti Malaysia Pahang, 26300, Gambang, Kuantan, Pahang, Malaysia

[3]Dept. of Information and Communication Technology, Mawlana Bhashani Science and Technology University, Tangail-1902, Bangladesh

**Corresponding author*: badrul.ict@gmail.com; suryanti@ump.edu.my

Abstract

Automatic keyphrase extraction remains a significant and difficult issue in the current research domain because of the exponential explosion of information and internet sources. Various activities involving natural language processing and information retrieval systems greatly benefit from the use of keyphrases. To extract the best keyphrases and summarize the documents to the highest standard, feature extractions for those keyphrases are crucial. This paper proposes an unsupervised region-based KDA technique for analyzing the distance of keyphrases from news articles as feature of keyphrase extraction. The proposed technique is divided into eight phases: data collection, data pre-processing, data processing, keyphrase searching, distance calculating, distance averaging, curve-plotting, and curve-fitting. At first, the proposed technique collects two different datasets that contain the news articles; it is then applied to the data pre-processing step that uses a few preprocessing algorithms. Then this pre-processing data is used in the data processing stage, where it is sent to the keyphrase searching step, the distance calculation process, and then the distance averaging steps. Curve plotting analysis is then applied, and finally the curve fitting technique is used. Afterwards, the performance of the proposed technique is put to test and evaluated using two of the most accessible benchmark datasets. The proposed method is then compared to other available methods in order to demonstrate its efficiency, advantages, and importance. Lastly, the results of the experiment demonstrated that the proposed approach efficiently analyzed the keyphrase distance from news articles, produced an F1-score of 96.91%, and presented keyphrases of 94.55%, as well as greatly improved the effectiveness of the current keyphrase extraction methods.

Keywords: Curve fitting technique; Data pre-processing; Data processing; Feature extraction; KDA technique; Keyphrase extraction.