Text Data Preprocessing Library: Bilingual Approach

Authors : Kabil Boukhari

Abstract : In the context of information retrieval, the selection of the most relevant words is a very important step. In fact, the text cleaning allows keeping only the most representative words for a better use. In this paper, we propose a library for the purpose text preprocessing within an implemented application to facilitate this task. This study has two purposes. The first, is to present the related work of the various steps involved in text preprocessing, presenting the segmentation, stemming and lemmatization algorithms that could be efficient in the rest of study. The second, is to implement a developed tool for text preprocessing in French and English. This library accepts unstructured text as input and provides the preprocessed text as output, based on a set of rules and on a base of stop words for both languages. The proposed library has been made on different corpora and gave an interesting result.

Keywords : text preprocessing, segmentation, knowledge extraction, normalization, text generation, information retrieval
Conference Title : ICAIL 2022 : International Conference on Artificial Intelligence in Law
Conference Location : Copenhagen, Denmark
Conference Dates : July 19-20, 2022

Open Science Index, Computer and Information Engineering Vol:16, No:07, 2022 publications.waset.org/abstracts/150846.pdf