Abstract

For the retrieval of information from documents of different natural languages, pre-processing of the document is the main task. During pre-processing, words which occur too frequently and have little semantic in the document should be identified. Such words are called Stopwords. Stopwords list for different world languages like English, Chinese, Hindi, Arabic Sanskrit etc. are identified. But as I long as I know there is no standard method to identify these words for the Amharic language. In this paper, we proposed the automatic identification of Stopwords for the Amharic text by an aggregate based methodology of words frequency, inverse document frequency, and entropy value measure. Available works on Stopwords identification techniques are based on static or dictionary based Stopwords lists. This method inefficient and very expensive and it is a time-consuming task as the searching process takes a long time. The proposed work will overcome these problems using aggregated methods of both frequency measures and entropy measures of words in the Amharic text for the automatic Stopwords identification.
References


**Index Terms**

Computer Science Information Systems

**Keywords**

Natural language processing, information retrieval, document pre-processing, Stopwords, Amharic