Abstract

Text summarization is an important activity in the analysis of a high volume text documents. Text summarization has number of applications; recently number of applications uses text summarization for the betterment of the text analysis and knowledge representation. In this paper a frequent term based text summarization algorithm is designed and implemented in java.
The designed algorithm works in three steps. In the first step the document which is required to be summarized is processed by eliminating the stop word and by applying the stemmers. In the second step term-frequent data is calculated from the document and frequent terms are selected, for these selected words the semantic equivalent terms are also generated. Finally in the third step all the sentences in the document, which are containing the frequent and semantic equivalent terms, are filtered for summarization. The designed algorithm is implemented using open source technologies like java, DISCO, Porters stemmer etc. and verified over the standard text mining corpus.

Reference

- DISCO (extracting DIstributionally related words using CO-occurrences) - http://www.linguatools.de/disco/disco_en.html the British National Corpus (BNC)
Index Terms

Computer Science         Text Mining

Key words

Text Summarization         Frequent Words
Semantic Similarity

Summarization Algorithm