Abstract

In the Information era, optimization of processes for Information Retrieval, Text Summarization, Text and Data Analytic systems becomes utmost important. Therefore in order to achieve accuracy, extraction of redundant words with low or no semantic meaning must be filtered out. Such words are known as stopwords. Stopwords list has been developed for languages like English, Chinese, Arabic, Hindi, etc. Stopword list is also available for Sanskrit language.

Stop-word removal is an important preprocessing techniques used in Natural Language processing applications so as to improve the performance of the Information Retrieval System, Text Analytics & Processing System, Text Summarization, Question-Answering system, stemming etc. In this paper, a simple approach is used to design stop-word removal algorithm and its implementation for Sanskrit language. The algorithm and its implementation uses dictionary based approach. In dictionary based approach predefined list of stopwords is compared to the target text on which removal is required.
References

10. Kaur J. and Saini J. R., “A Natural Language Processing Approach for Identification of Stop Words in Punjabi Language”, published in International Journal of Data Mining and Emerging Technologies; ISSN: 2249-3212 (eISSN: 2249-3220); Indian Journals, New Delhi, India; vol. 5, issue 2, November 2015; pages 114-120; DOI: 10.5958/2249-3220.2015.00015.4
Index Terms

Computer Science

Algorithms

Keywords

Information Retrieval (IR), Natural Language Processing (NLP), Sanskrit, Stopword, Tokenization.