ABSTRACT

In this current era, relevant information is scattered across multiple documents and there might be redundancy in the data. It is difficult for the user to go through all the documents by allotting a lot of time to collect the entire information. To overcome this Multiple Document Summarization is proposed.

The proposed framework consists of 2 phases. The first phase is to apply pre-processing steps to all the documents. In the pre-processing stage, pre-processing steps like stopword removal, stemming are applied. In stopword removal, all the stop words in the document are removed. In stemming, the given word is converted to its root word.

In the second phase Feature Extraction methods like frequency score, position score, length score, headline score are calculated to summarize the multiple documents into a single document. Frequency score is based on word probability. Length score is based on the ideal length given. Position score is based on the position and length of the sentence. Headline score is based on percentage of words matching with the heading.

To check the efficiency, documents are summarized manually and then compared with the system generated summaries. The average efficiency achieved in this project is 80.34% .With the use of multiple Document Summarization, users can input multiple documents and get the important information in a single document based on the compression ratio and thereby saving their time.