Abstract

The reduced text of a document is the collection of sentences that contains the important sentences containing keywords of the document. The authentic keywords extraction is the primary target for any text reduction algorithm. The presented survey shows the primary algorithm used for document summarization based on keywords. Also, the work presents a novel approach for keywords identification and in turn text reduction based on words histogram, the no. of sentences containing the words and knowledge corpus. The text summary is extracted using the sentence vectorization process. The sentence vectorization gives the sentences that have at least one of the key words in the sentence from the entire document. The algorithm works fine for the textual matter in the document in MS Notepad format. Factual information that is normally covered under double inverted comas is also given due attention in text summary.

References

- Egitim Fakültesi, Mehmet Akif Ersoy, "Quality of written summary texts: An analysis in the context of gender and school variables", 1877-0428 © 2010 Published by Elsevier Ltd.
- Kushal Bafna, Durga Toshniwal, "Feature Based Summarization of Customers’ reviews of Online Products", Reviews of Online Products, 2013 The Authors. Published by Elsevier B. V.
- Tiedan Zhu, Kan Li, "The Similarity Measure Based on LDA for Automatic Summarization", 2011 Published by Elsevier Ltd

**Index Terms**

Computer Science

Algorithms

**Keywords**

Text Reduction  Text summary  Sentence Vectorization  Word Histogram

Reduction algorithm
Synonyms