Abstract

Text Data Mining, also known as Text Mining or Knowledge Discovery from Textual Data, refers to the process of extracting interesting and non-trivial patterns or knowledge from text documents. It is considered by many as the next wave of knowledge discovery. Text Mining framework consists of two components: Text cleansing that transforms unstructured text documents into an intermediate form; and knowledge distillation that deduces patterns or knowledge from the intermediate form.

Under the broader spectrum of Text Mining is a more specific one related to Text Summarization. Text Summarization is a process to reduce the text in a system and to generate a good and relevant summary of the given contents. It is a process of generating a concise but adequate version of a larger source so that the essence is retained but the size of the content gets shortened.

The primary objective of the thesis is to create a succinct and effective summary for a single document. Overall, Text Summarization is divided into two categories: abstraction and extraction. The extraction selects sentences that have the highest weight in the retrieved document and put them together to generate a summary version of the original document without changing or altering the main text. In abstraction, the original text is converted into another semantic form with the help of linguistic methods to get a shorter summary of the original document. The extraction based Text Summarization is widely accepted by users and the research community. The abstraction based Text Summarization generates summaries wherein the semantics of the document are interpreted and a new smaller version describing the content of the input bigger document is generated. The sentences of the summary in abstraction are not necessarily part of the main document. The work carried out in this research is related to the extraction based concept.

In this thesis, two different approaches have been proposed for Text Summarization. In both the models, Latent Semantic Analysis (LSA) has been used in combination with other methods. In the first approach, after pre-processing the input document, the Naïve Bayes classifier has been used with recursive feature elimination to generate the summary.

In the second approach, the concept of Deep Learning using Self-Organizing Maps – an unsupervised learning and Artificial Neural Network – a supervised learning has been used to select sentences from the input document.

In both the models, the standard datasets available online for research in this area – the DUC dataset and the Opinosis dataset have been used. To evaluate the summaries generated by the models, the ROUGE (Recall-Oriented Understudy for Gisting Evaluation) toolkit has been used. The models are developed in Python. The papers based on these models have been published in UGC recognized journals and Springer publications.