الجامعة المصرية للتعلم الإلكتروني الأهلية
THE EGYPTIAN E-LEARNING UNIVERSITY

Language_Engineering_Project

(Text Summarizer)

Team members

1- Salaheldin Mohamed Salah _2000343

2- Hamdy Hamada Ahmed_ 2001689

3- Hatem Ali Hassan_2000479

Project Link :

https://github.com/DragonsEG/NLP-Text-Summarization

Text Summarizer

Text summarization is the process of turning more significant documents into shorter and more precise paragraphs or sentences. The process brings out crucial information and ensures that the paragraph's meaning stays the same. This helps reduce the time to understand large papers like research articles, without skipping any vital information. The benefits of using text summarization include:

- ✓ They make reading easier.
- ✓ It saves time.
- ✓ It helps memorize information easily.
- ✓ It boosts the work rate efficiency.

Algorithm used in the project

1- Basic Sum Algorithm

Sum Basic is an algorithm to generate multi-document text summaries. Basic idea is to utilize frequently occurring words in a document than the less frequent words to generate a summary that is more likely in human abstracts. It generates n length summaries, where n is user specified number of sentences. Sum Basic has the following advantages:

- Used to easily understand the purpose of a document.
- Provides greater convenience and flexibility for the reader.
- Generates shorter and concise forms from multiple documents.

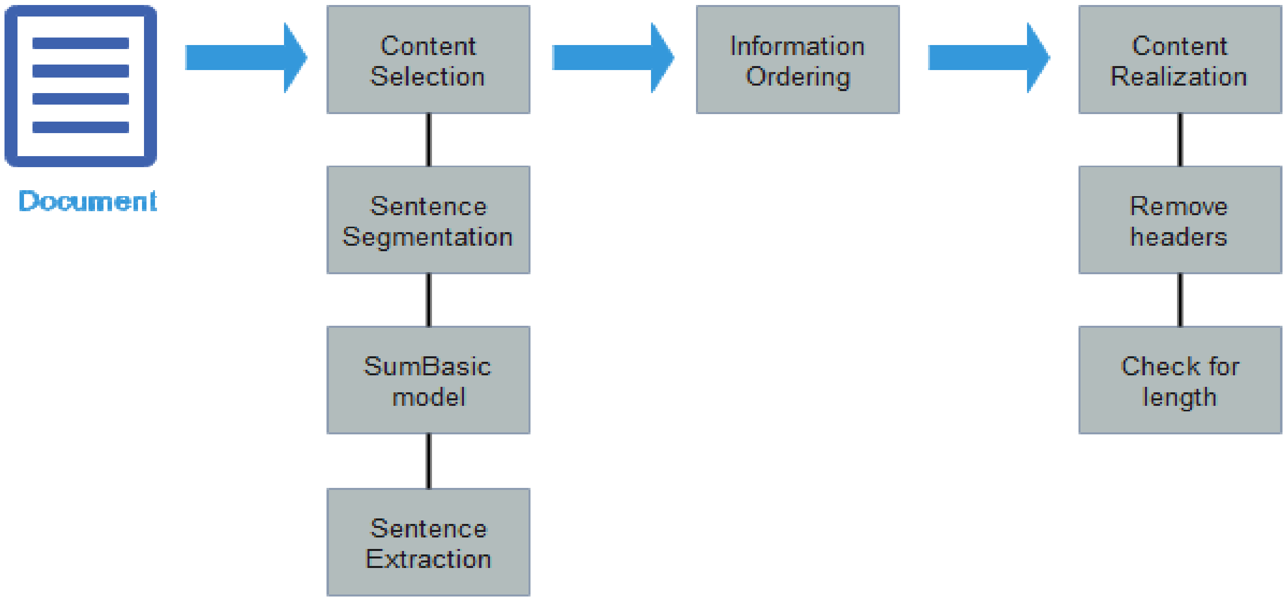


Fig. SumBasic implementation

2- Text Rank Algorithm

Text Rank is a text summarization technique that is used in Natural Language Processing to generate Document Summaries. Text Rank uses an extractive approach and is an unsupervised graph-based text summarization technique. Overview of PageRank

-PageRank is an algorithm used to calculate the rank of web pages and is used by search engines such as Google.

Text Rank is based on the PageRank Algorithm.

This algorithm gets its name from Larry Page, one of the co-founders of Google.

Algorithm

- The rank/importance of a page is decided by the number and quality of links to that page.
- It applies several iterations on the pages to arrive at a final value.

3- Luhn's Algorithm

Luhn's algorithm is an approach based on TF-IDF. It selects only the words of higher importance as per their frequency. Higher weights are assigned to the words present at the beginning of the document. It considers the words lying in the shaded region in this graph:

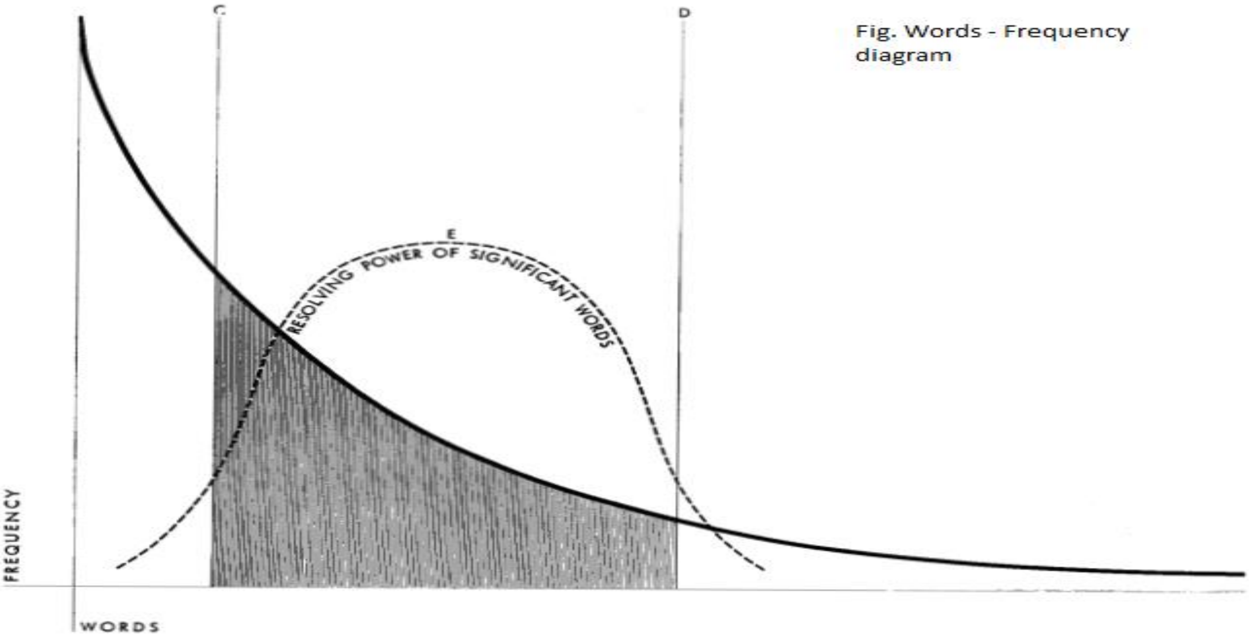
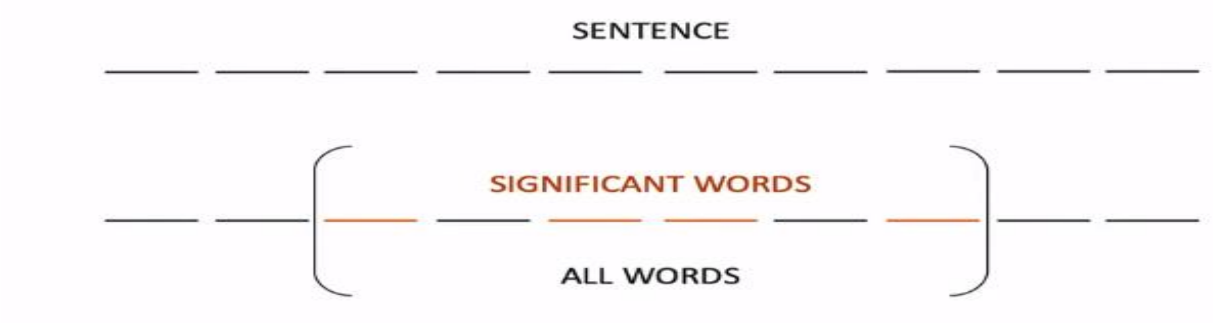


Fig. Words - Frequency diagram

The region on the right signifies the highest occurring elements while words on the left signifies least occurring elements. Luhn introduced the following criteria during text preprocessing:

- Removing stop words
- Stemming (Likes->Like)

In this method we select sentences with the highest concentration of salient content terms. For example, if we have ten words in a sentence and four of the words are significant.



For calculating the significance instead of number of significant words by all words here we divide them by the span that consist of these words.