TEXT SUMMARIZER/MACHINE TRANSLATION

A Project Report submitted in partial fulfilment

of the requirement of undergraduate degree

**Bachelor of Technology**

**In**

**Computer Science Engineering**

By

PALADUGU SIRIVANTH

AP18110010245

CSE-D

****

**Department of Computer Science Engineering**

**SRM University, AP**

**Amaravathi-522502**

**ABSTRACT**

In this new era, where tremendous information is available on the internet, it is most important to provide the improved mechanism to extract the information quickly and most efficiently . It is very difficult for human beings to manually extract the summary of a large documents of text. There are plenty of text material available on the internet. So there is a problem of searching for relevant documents from the number of documents available, and absorbing relevant information from it.In order to solve the above two problems, the automatic text summarization is very much necessary. Text summarization is the process of identifying the most important meaningful information in a document or set of related documents and compressing them into a shorter version preserving its overall meanings.

**Introduction**

A summary is a text that is produced from one or more texts, that conveys important information in the original text, and it is of a shorter form. The goal of automatic text summarization is presenting the source text into a shorter version with semantics.

The most important advantage of using a summary is ,it reduces the reading time. Text Summarization methods can be classified into extractive and abstractive summarization. An extractive summarization method consists of selecting important sentences, paragraphs etc. from the original document and concatenating them into shorter form. An Abstractive summarization is an understanding of the main concepts in a document and then express those concepts in clear natural language.

There are two different groups of text summarization : indicative and informative.Inductive summarization only represent the main idea of the text to the user. The typical length of this type of summarization is 5 to 10 percent of the main text .On the other hand, the informative summarization systems gives concise information of the main text .The length of informative summary is 20 to 30 percent of the main text .

**Literature Survey**

I studied many websites, research papers and books for implementing this project. Extractive summarizers have been mostly based on scoring sentences in the source document since past. The most common and recent text summarization techniques use either statistical approaches, or linguistic techniques.The high frequency words ,standard keyword ,Cue Method, Title Method ,Location Method are used for weighting the sentences.

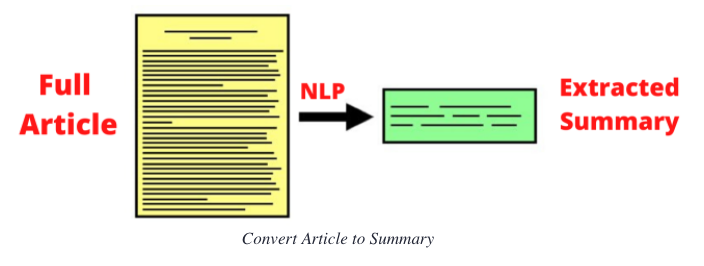
**Proposed Work**

In order to build the text summarization tool, we first need to install a NLP toolkit. Here I have used spacy to run the project. First we need to import the libraries from spacy to the notebook we use and then we need to list all the stop words. After listing the stop words we now need to list the tokens used in our text and this is known as text tokenization. Now we need to list the punctuations used in our text.

Now we need to find out the word frequency i.e. how many times each word occurs in our text this process is known as word tokenization. After finding the word frequency we now need to find out the frequency of the sentence that is how many similar sentences are occurring in our text and this process is known as sentence tokenization. And then after this we need to find out the sentence score.

After finding out the sentence score we now need to select the length of our summary in terms of percentage and then we need to adjust it with the original text. And after getting our summary we also can compare and find out the length of the summary of the text to the previous text.

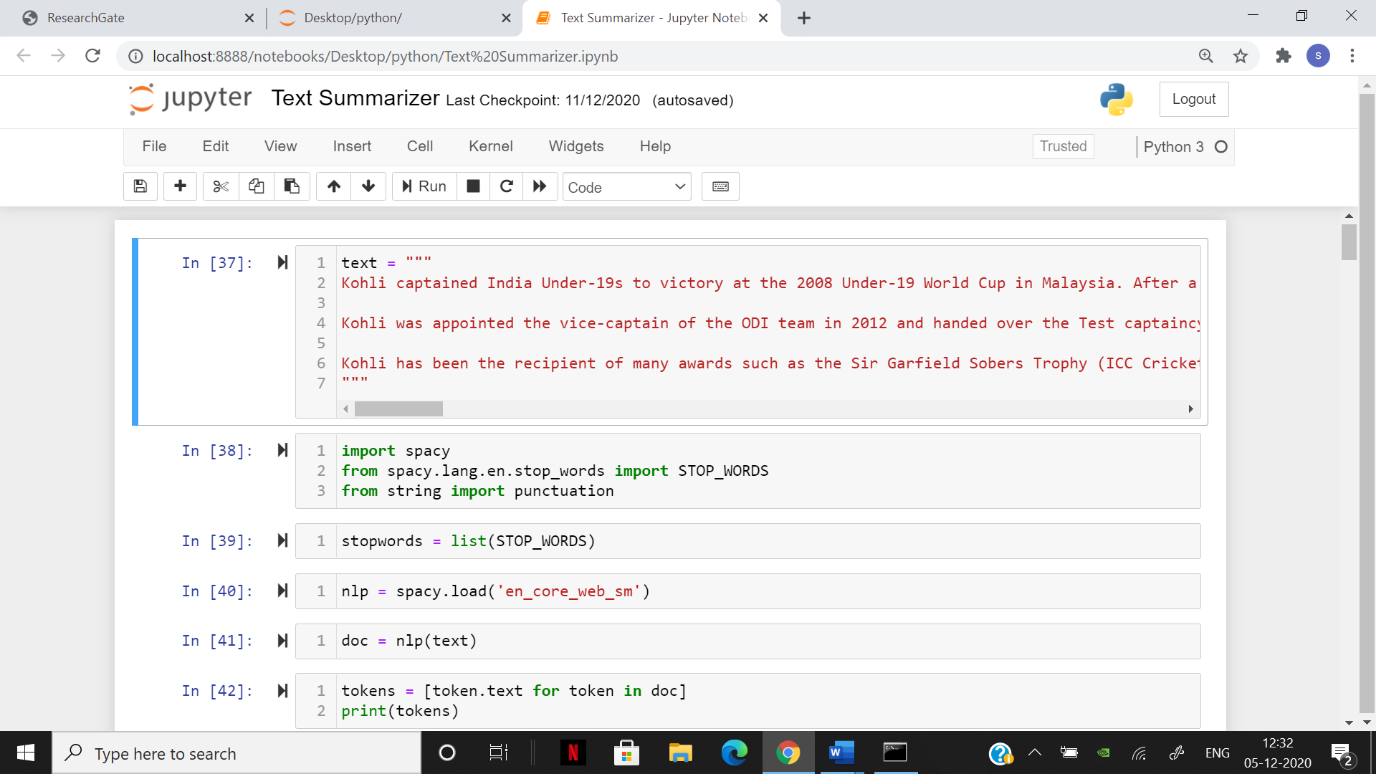
Below is the pictorial representation of the text summarization

****

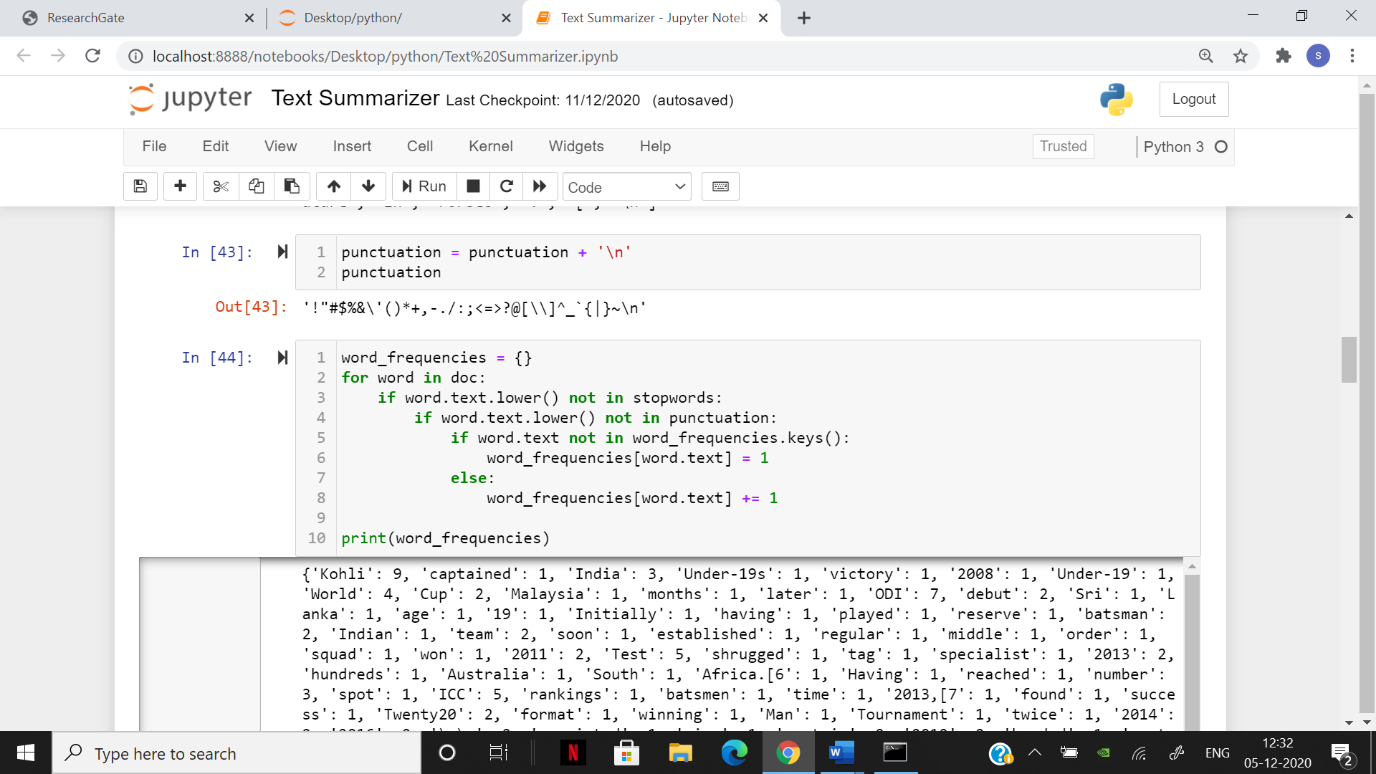
**Experimental Setup**

* Anaconda prompt is the tool which is used to download SpaCy and Jupyter notebook.
* SpaCy is a tool used for Natural language processing.
* Jupyter notebook is the tool used to run the whole program in python using Natural language processing.

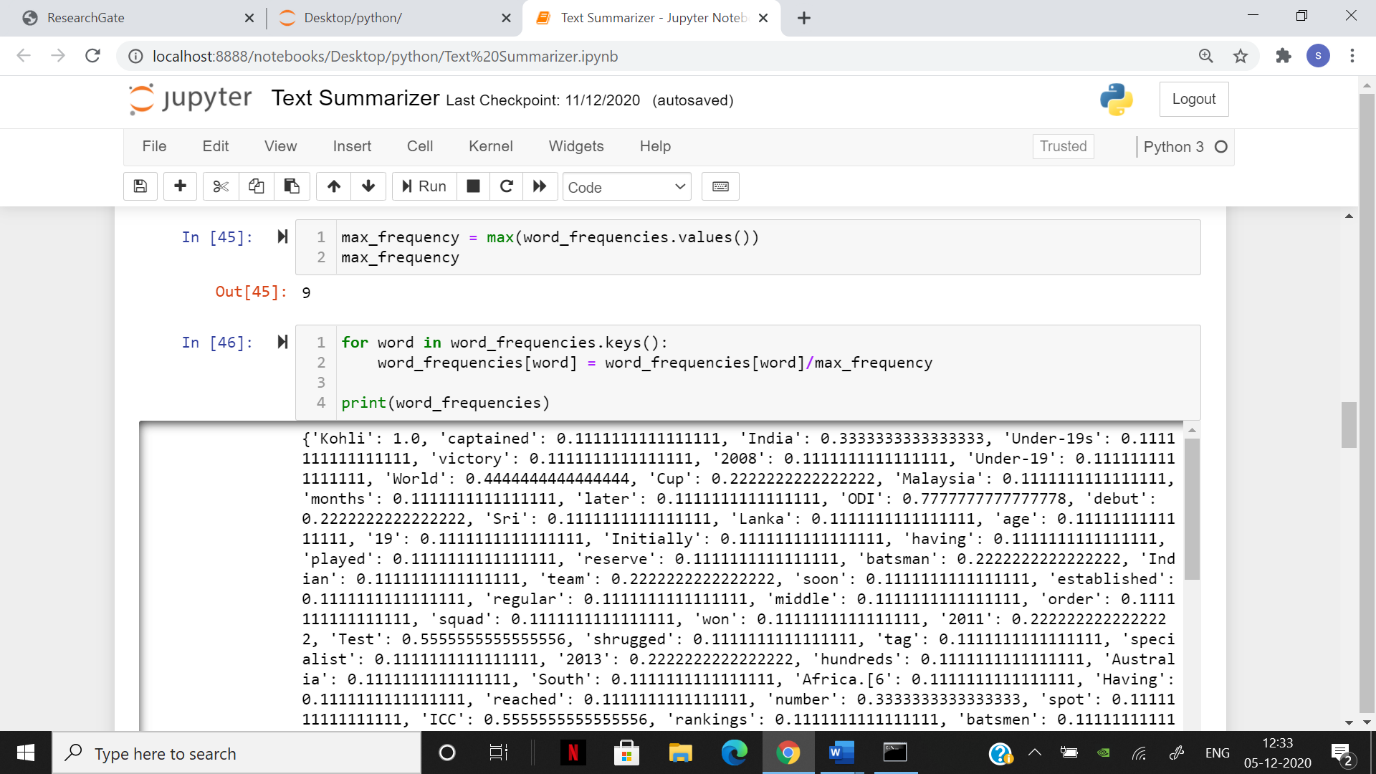
**Outputs:**



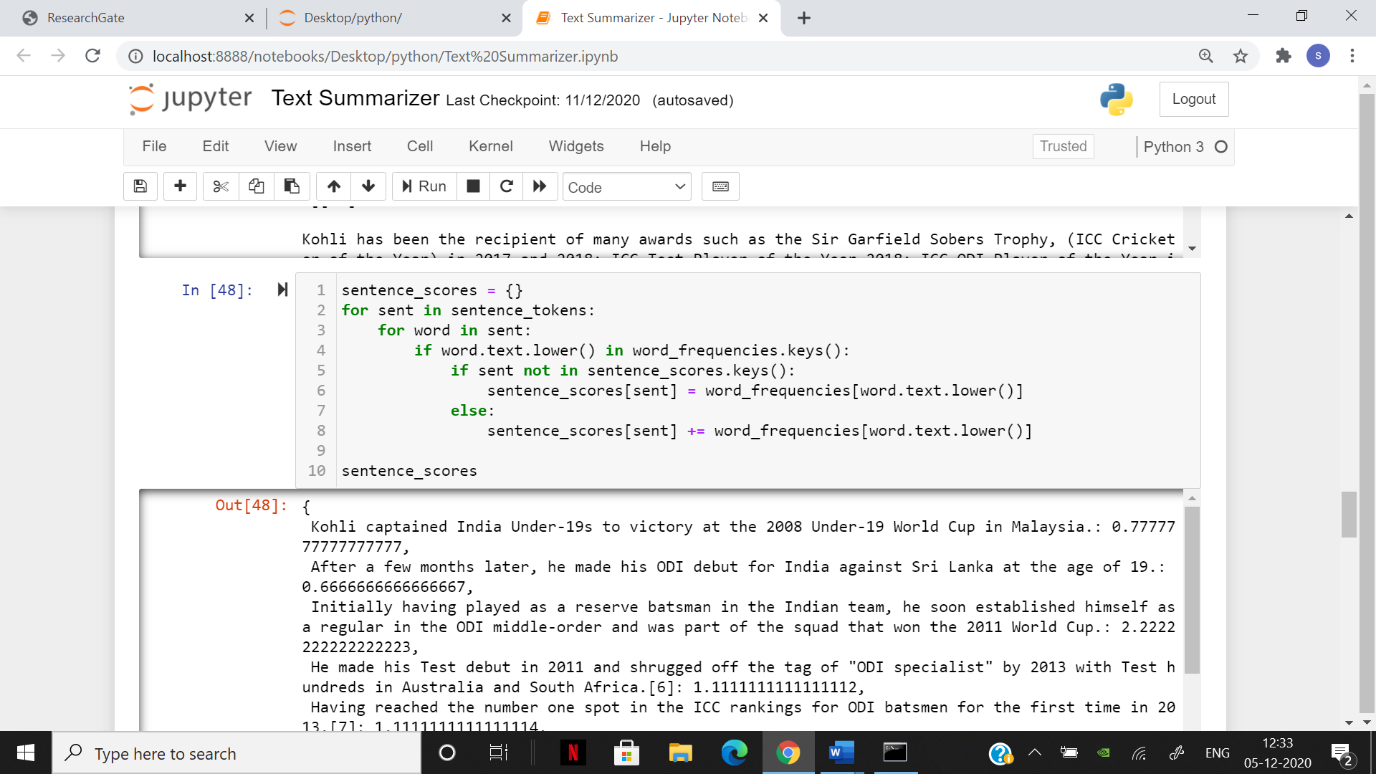
GIVEN TEXT, STOP WORDS AND TOKENS



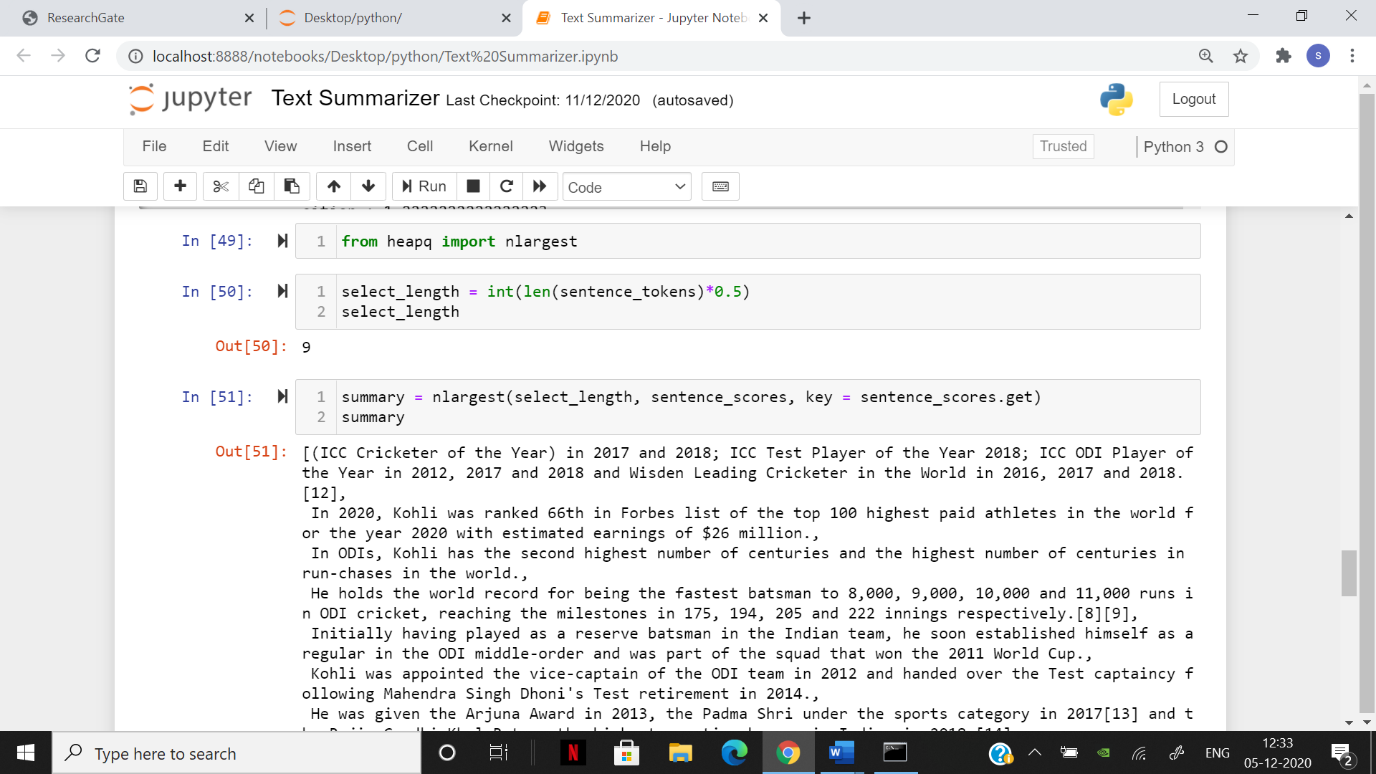
PUNCTUATION AND WORD FREQUENCIES



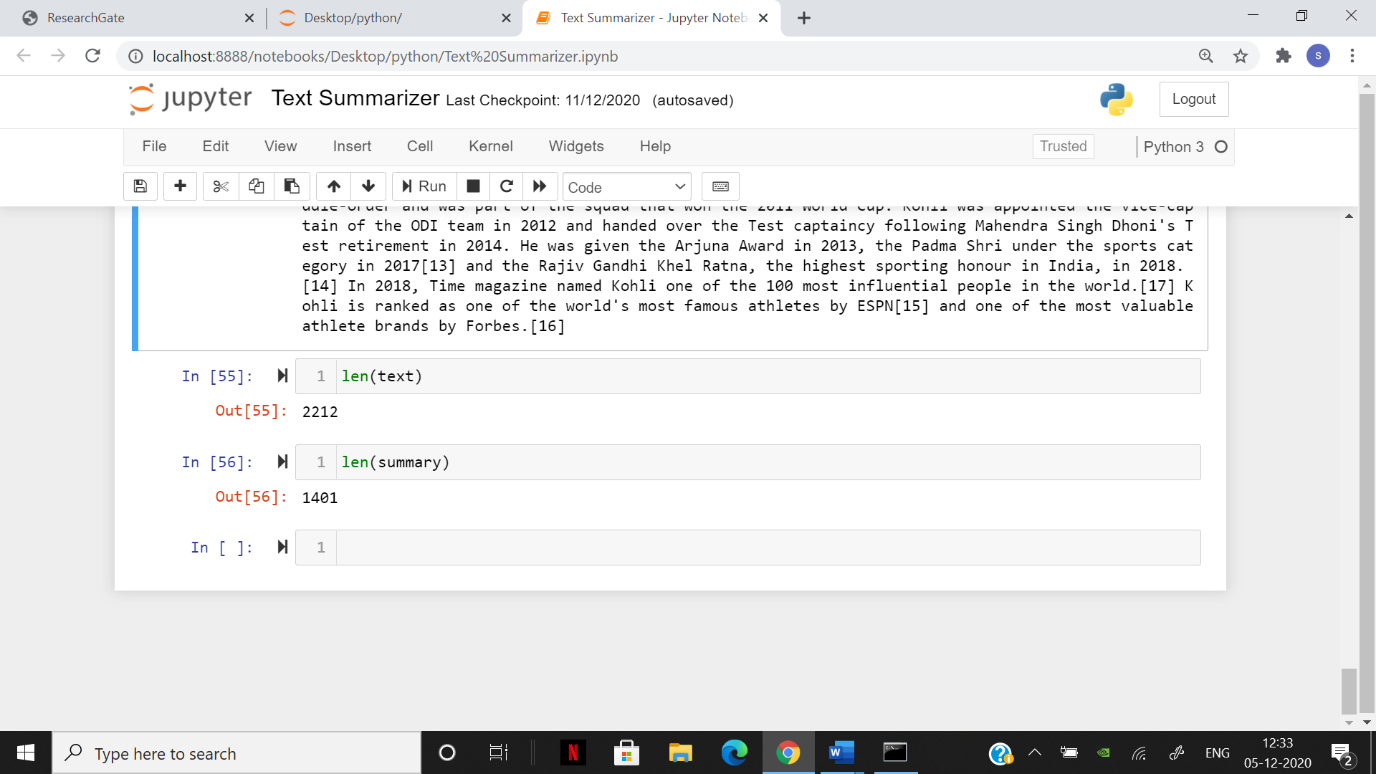
MAXIMUM FREQUENCY AND WORD FREQUENCY



SENTENCE SCORE



ADJUSTING LENGTH OF THE SUMMARY



COMPARING LENGTH OF TEXT AND SUMMARY

**Conclusion and Future work**

Finally our summary is displayed and we got the output of the text in form of summary. Now in future I will try to implement it with image processing to find out the summary of the article.

**References**

[**https://www.analyticsvidhya.com/blog/2018/11/introduction-text-summarization-textrank-python/**](https://www.analyticsvidhya.com/blog/2018/11/introduction-text-summarization-textrank-python/)

[**https://algorithmia.com/blog/introduction-automatic-text-summarization**](https://algorithmia.com/blog/introduction-automatic-text-summarization)