Summarization of Text using Natural Language Processing in Machine Learning:Analyzing, Interpreting, Shortening of Enormous Input Field Content

*V Karthick, Associate Professor Subbalakshmi N, UG Student Sudhashree M, UG Student Swetha P, UG Student*

*Department of CSE Department of CSE Department of CSE Department of CSE*

*Rajalakshmi Engineering College Rajalakshmi engineering College Rajalakshmi Engineering College Rajalakshmi Engineering College*

*Chennai, India Chennai, India Chennai, India Chennai, India*

[*vkarthick86@gmail.com*](mailto:vkarthick86@gmail.com)[*210701263@rajalakshmi.edu.in*](mailto:210701263@rajalakshmi.edu.in)[*210701268@rajalakshmi.edu.in*](mailto:210701268@rajalakshmi.edu.in)[*210701277@rajalakshmi.edu.in*](mailto:210701277@rajalakshmi.edu.in)

**ABSTRACT:**

**Text summarization is a one time solution to cut short the enormous content of any text. In the modern world, people find it very difficult to read extensive textual material as it makes people lose interest often and moreover understanding massive volume of text is also sometimes rocket science as people might not be in a mood to comprehend text as mood plays a vital role in comprehending texts. People when not in a good state of mind or when they are confused, find it difficult to understand massive text content. Here text summarization comes as a boon when people want crisp content and when they hardly have any time to spare. Texts in newspaper articles may contain massive or enormous volume of content, while reading them will make people forget the important topics. People start to loose interest if they read large volume of text within a short span of time. A crisp text helps to improve people’s understanding of the content. This doesn’t lead to confusion of the text, they help in precise and accurate understanding of content of the text.The methodology used to summarize the content of the text is NLP (Natural Language Processing). Here the content from enormous volumes of text is fed from articles or newspapers or even Wikipedia. That is the first step. The next step is click the button that cuts the content of text into little text. This project depicts how a ML model cut shorts massive content to small content of text. It retains all the important topics and text that are necessary and leaves all unnecessary information that can lead to confusion. First the text is extracted from the original text then, keywords are extracted using naive bayes and SVM(Support Vector Machine). The next is data preprocessing that is carried out using Tokenization, Stop\_words, Punctuation libraries from NLP**[**(6)**](https://paperpile.com/c/8cREju/rJaG)**. Then frequency calculation is done by calculating weights of words and key terms. Priority of each sentence is calculated along with summary quality. All these are fed to a GUI (Graphical User Interface) that provides a front end website for our text summarization project. Summary of the original content is given once the summarize button is clicked**[**(4)**](https://paperpile.com/c/8cREju/O3Fr)**.**

***Keywords- Support Vector Machine(SVM), Natural Language Processing(NLP),Machine Learning(ML), Parts of Speech (POS), LLM (Large language models), Sentence scoring, Graphical User Interface(GUI).***

**1. INTRODUCTION:**

In this gadget centric world, people are often so involved in their gadgets which could be smart phones or laptops or even tablets be it children or adults or even the elderly [(Fiori and Alessandro 2019)](https://paperpile.com/c/8cREju/b9zs). In olden days there was no other source of information except newspapers so people relied fully on newspaper to know information. But now that is not the case, there are so many other source of information like news apps in smart phones and google news. Information available online be it google news or news apps provide us with crisp information and massive information running to 1-2 pages. Due to the enormous content written in newspapers, people tend to lose interest naturally and shortly, they get bored very easily as they require a glimpse of news. That is why people nowadays are relying mostly on new apps rather than newspapers. Sometimes language used to write a newspaper may be of high level which may provide poor understanding of text. But the language used in these apps is simple which provides better understanding. That is also the reason why people are switching over to news apps. Here we are using a text summarizer to shorten massive content of text which provides us crisp content of information so that people do not tend to loose interest as well they can have proper understanding of content. This model is trained well on large amount of text so that model can understand the contents, terminologies properly to give summarized content. Natural Language Processing(NLP) is the main technique that is used in this text summarization project. NLP is the method that is used analyze, interpret, generate language that is meaningful. Therefore, we are using this technique to summarize large content of text. NLP is also used to understand human voice and speech and interpret those which is speech recognition and even chatbots.

The work done by Mani I, Maybury MT [(Mani and Maybury 1999)](https://paperpile.com/c/8cREju/zIwx) describes the advances in automatic text summarization process that uses NLP with frequency calculation process and information retrieval.It also describes why text summarization is important in this modern world. Torres-Moreno JM [(Torres-Moreno 2014)](https://paperpile.com/c/8cREju/ojnR) in his work demonstrates why majority of documents are unstructured and how unstructured data confuses people . It also explains how text summarization condenses the text while retaining all the important information that provides better comprehension. This work also explains the recent advances in text summarization also its application and trends. Fiori, Alessandro [(Fiori and Alessandro 2019)](https://paperpile.com/c/8cREju/b9zs) in their work describes how content analysis and text extraction is done in a text summarization project. It also demonstrates why those are done to increase user familiarity with this project. Priya V and Uma Maheswari K [(Priya and Umamaheswari 2023)](https://paperpile.com/c/8cREju/1CfF) in their work explains how known problems in text summarization is solved using cognitive methods. It also explains how the text documents are shortened very easily using cognitive approaches. It also demonstrates how text summarization helps people in having better understanding of text and how people can understand can read at faster speed. Wang T, Yang C, Zou M, Liang J, Xiang D, Yang W, Wang H, Li J [(Wang et al. 2024)](https://paperpile.com/c/8cREju/LJQj) in their work explains all the limitations involved in text summarization project. First limitation is that the text is also considered in a linear manner that ignores structured data. Second limitation is how and why it is difficult to extract information for local topics. Nenkova A, McKeown K [(Nenkova and McKeown 2011)](https://paperpile.com/c/8cREju/glSv) in their research work explains how the traditional methods and also recent methods are used for getting necessary information and leaving out unnecessary content. It is also describes the evolution of summarization of text over the years.

**Large Language Models (LLMs):**

* **Architecture:** Large Language Models (LLMs) boast a complex neural network architecture with a deep learning approach. They have a significantly larger number of parameters, enabling them to capture intricate patterns in data.
* **Training:** LLMs require extensive training on large, diverse datasets. This process is resource-intensive but crucial for the model’s ability to perform complex language tasks.
* **Predictive Performance:** LLMs are known for their high accuracy and ability to generate refined, comprehensive text. Their deep understanding of language nuances allows for more accurate disease associations.
* **Resource Requirements:** The advanced capabilities of LLMs come at the cost of higher computational resource demands. They require powerful hardware and more energy, which can be a limiting factor in resource-constrained environments.

The aim of the study is to :

1. Extract keywords from the text
2. Increase the readability of long text
3. Easy understanding of the long texts
4. Less time consuming process
5. Ability to read more topics with important key content

The research gap in this study, Text summarization projects can cut short large text content. It can also give a glimpse of content when we insert a video from youtube. But it cannot summarize text from videos taken from other apps like Netflix, Disney+Hotstar and other entertainment apps.

**3. MATERIAL AND METHODS**

Text Summarization provides us the easiest method, less time consuming way for reading the text from any source such as articles, newspaper, books, journals and from research papers. This approach is proven to be useful for common people who cannot be able to spend enough time to read the whole context of a topic. It helps in reducing the time consumed in reading the whole text, improves the clarity and understanding of the text, focuses on key points and within a limited span of time people will be able to read a large context of text. In this project we accomplish the task of Text Summarization using NLP. NLP also known as Natural Language Processing is used as a Communication medium between the user and the Computer. It performs a wide range of tasks such as Text Preprocessing, Understanding the text sentence by sentence, Generating the text by giving a prompt and Summarizing the text when a text is provided. There are various methodologies that are used for text summarization - Rule based approach, abstractive or extractive approach, graph based and statistical approach. We have accomplished this project by adopting a Rule based and extractive approach for summarization of the text. This approach helps in extracting the key features and key points of the text, frequency of the word that appears in the text and finally collaborating all these the summarized text is displayed. Generally the text summarization is classified into two types of summarization. They are Abstractive and Extractive summarization techniques. Extractive summarization is a type of summarization which is implemented by selecting some main sentences of the text, joining all the main parts of the text will give the full summarized content of the text. Abstractive summarization technique is a another type of summarization technique that helps in summarizing the whole content of the text by generating a new content of the text by analyzing the given text, it helps in paraphrasing and generating new sentences of the given text.

The first step in the text summarization in GUI, is extraction of text from any source such as newspaper, articles, journal, research paper, books, etc. the text may contain any amount of line of content it may contain even 500 or 1000 words, text summarizer summarizes it to 150 words minimum. Using the naive bayes and SVM feature extraction is done, all the important keywords are extracted and sentence is formed by using these important key words. Second step is Parts of speech, it is also an extraction step in which all the unwanted or wasted words are removed thereby making the text more concise and precise. Frequency is calculated for each word in the sentence and weights are allotted to the words in each sentence. Depending upon the weights and frequency calculation of the words in the sentence, sentence scoring is done. Sentence scoring is the next step in which the sentences in the text are given weights in the frequency calculation and the scoring of the sentence is done based on the weights and the frequency of the words. Based on the sentence scoring and frequency calculation the sentences are framed for the summarized text. At the last step in summarization the summarized text is displayed.

**HARDWARE AND SOFTWARE REQUIREMENTS:**

**HARDWARE REQUIREMENTS**:

* Disc Space: 2 to 5GB minimum of RAM
* Processor: Intel processor
* GPU that supports CUDA (Compute unified device architecture)
* Network connectivity of the internet with higher bandwidth is recommended for faster processing and generating.
* Storage of minimum 2 gb is required for processing of the texts.

**SOFTWARE REQUIREMENTS:**

* Operating System: Windows 11
* Python Version: 3.8 or higher.
* Jupyter notebook, spyder to be installed.
* Python Libraries: transformers, torch, numpy, matplotlib, pandas, tkinter, spacy, stop\_words, sentence\_scoring, nlp libraries, tokenization, punctuation, heapq, word\_freq, frequency\_counter, scikit-learn.
* Data Processing: Pandas for data manipulation, scikit-learn for additional machine learning tools.

**4.EXISTING SYSTEM:**

The existing system of the study includes both extractive and abstractive method of summarization, The extractive method of summarization tells us that for eg. if a user gives an input of 10 lines of sentence then the output will be concise to only 5 lines of sentences in a text. The key words or the main words are selected from the text and is given as a summary based on the most frequently used word that is calculated based on the method of frequency calculation. The meaning of the sentence is generally analyzed by the technique of semantic search. They make use various algorithms such as Text-rank, page-rank , TF-ID algorithm and graph based algorithms for generating concise text.

The difficulties faced in existing system is:

1. It gives a different meaning of the text.
2. Difficult to handle complex sentences in a text.
3. Limited readability of the text.
4. The text taken from various resources gives different meaning.
5. Cost of computation and time taken

* TEXT EXTRACTION:

The existing system such as textrank or pagerank is an algorithm used by Google search engine to rank web pages in search engine results. It assigns numerical weights to each word in the text. Page rank works by assigning the weights the words and hyperlinks. The textrank algorithm works by extracting the words from the text and assigning weights to the text and treats each word as nodes and edges representing the relationship of similarity between the nodes. Textrank works by ranking the important key words or feature in a sentence. This technique is implemented to the whole text and thereby the summarized text is generated.

* **MODEL TRAINING:**

The model is trainedusing various text and the summarized text as a output is verified by using extractive text summarizer. The loss of the training model must be high it means that the model is having a high learning rate. High learning rate helps in testing of the text easily and gives high testing accuracy.

* MODEL TESTING:

The model is tested using various other texts taken from other resources. The model is tested against training text. If the model is trained against various training data the testing results will be quick and more accurate.

* MODEL EVALUATION:

The model is evaluated based on two metrics:

**ROUGE(Recall-Oriented Understudy for Gisting Evaluation):**

It is a method of measuring the quality of summaries by referring to the reference summarized text.

variants include ROUGE-N (measuring overlap of n-grams), ROUGE-L (measuring longest common subsequence), and ROUGE-W (measuring weighted longest common subsequence).

**FORMULAS:**

ROUGE-N=Count of overlapping n-grains/total number of n-grains in the reference summary

ROUGE-L=Longest common subsequence/total number of words in the reference summary.

**BLEU(Bilingual Evaluation Understudy)**

It helps in measuring the overlap of n-grams between the generated summary and the reference summaries.

* BLEU score: BLEU=BP×exp⁡(∑𝑛=1𝑁𝑤𝑛⋅log⁡(precision𝑛))

These metrics gives us the quantitative measure of the quality of generated summarized text.

**5.PROPOSED SYSTEM:**

Proposedsystem of the study includes, aiming at summarizing of the text using spacy, nlp libraries, stop\_words, punctuation, tokenizer, etc.

EXTRACTION OF TEXT:

The text is extracted from any source such as newspaper, research paper, article, journal and other resources. The text is extracted from any source and is used as an input in the model.

TOKENIZATION OF THE TEXT:

The extracted text is tokenized part by part. The tokenized word is then assigned weights in terms of frequency. The weights is calculated in terms of the frequency measure. The frequency is assigned to every words and then the words are classified based on the frequency.

FREQUENCY CALCULATION:

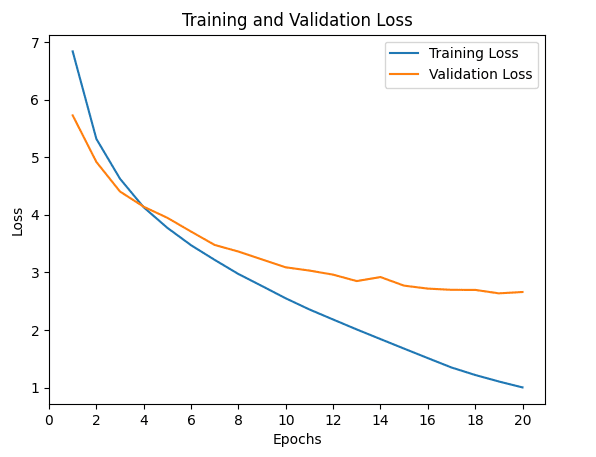
The words are classified based on the frequency assigned to them, while tokenizing the words in the sentence.

PARTS OF SPEECH (POS):

The frequency allotted words in the sentence are divided into small chunks of wanted and unwanted words. The unwanted words are removed from the sentence in the text thereby shortening and rearranging and generating new sentences of the text is done here to summarize the text. Thereby parts of speech (POS) helps by reducing the noise.

SCORING SENTENCE:

The sentence is scored based on the parts of speech (POS). Based on the scoring of the sentence a new sentence is generated and thereby the text is summarized.

****

**6.METHODOLOGY:**

DATA COLLECTION:

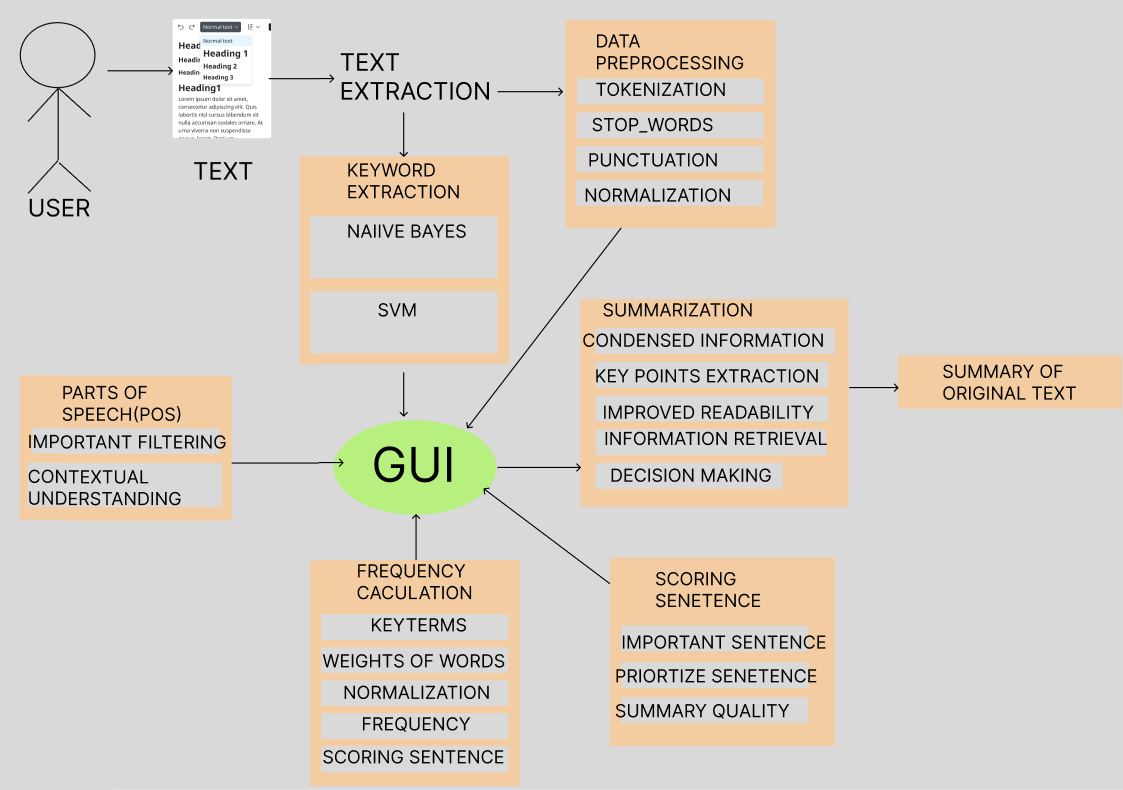
The initial step is to collect the enormous volume of data that is cut short into precise information. It is then fed to the text extraction part. Full stop is used to indicate the end of a sentence.

TEXT EXTRACTION:

The next step is text extraction in which keywords are extracted using naive bayes and SVM(Support Vector Machine). Keywords extraction is important because they help in providing better comprehension. Therefore it plays a vital role in this project.

DATA PREPROCESSING:

The next step is data preprocessing in which the data collected is processed using Tokenization[(Hvitfeldt and Silge 2021)](https://paperpile.com/c/8cREju/Me4u), stop\_words, punctuation and normalization libraries of NLP(Natural Language Processing). This part takes care of punctuation(Punctuation)[(Hardeniya et al. 2016)](https://paperpile.com/c/8cREju/O3Fr), end of one sentence(stop\_wotrds)[(Elaziz et al. 2019)](https://paperpile.com/c/8cREju/3vec)[(Elaziz et al. 2019; Kulkarni and Shivananda 2021)](https://paperpile.com/c/8cREju/3vec+rJaG), breaking down of text into smaller text(Tokenization)[(Hardeniya et al. 2016; Millstein 2020)](https://paperpile.com/c/8cREju/O3Fr+Dnf7). The model performance, consistency, standardization, removal of unwanted words, converting text to lowercase so as to reduce the vocabulary weight are taken care by Normalization library[(Hao et al. 2018)](https://paperpile.com/c/8cREju/bmEW).



PARTS OF SPEECH(POS):

The next step is parts of Speech (POS). This past is used to retain important information and leaves out unwanted information that leads to confusion. This also facilitates a contextual understanding of summary of text that is provided as output.

FREQUENCY CALCULATION:

This is the next step that includes calculating frequency of key terms, weights of terms. At the end of the summary the total of words is also mentioned. This is done by frequency calculation step of this project. Normalization and scoring sentence too are included in this step where the unwanted terms are removed and consistency along with standardization are maintained[(Kulkarni and Shivananda 2021)](https://paperpile.com/c/8cREju/rJaG).

SCORING SENTENCE:

The next step is scoring sentences that involves calculating frequency of terms in each sentence so as to check the importance of the terms in a text. Scoring sentence part also facilitates important terms filtering, priority of sentence and quality of summary that is provided as output.

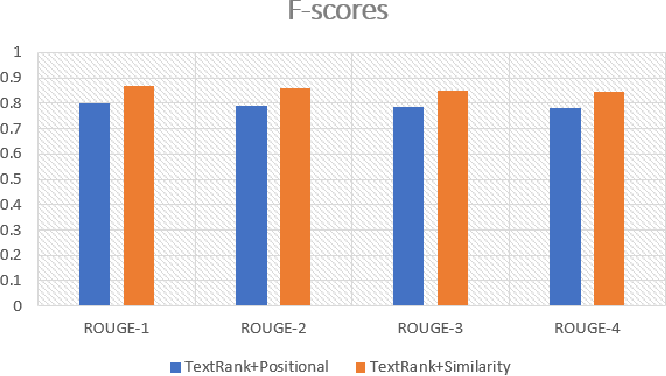
SUMMARIZATION:

This the next step in text summarization project where the information from original text are condensed, key terms are extracted and retained, readability of the text is improved so that the summary provides improved comprehension so as to reduce confusion, decision is made between words that are important and words that are less important and whom to retrain and whom to leave out.

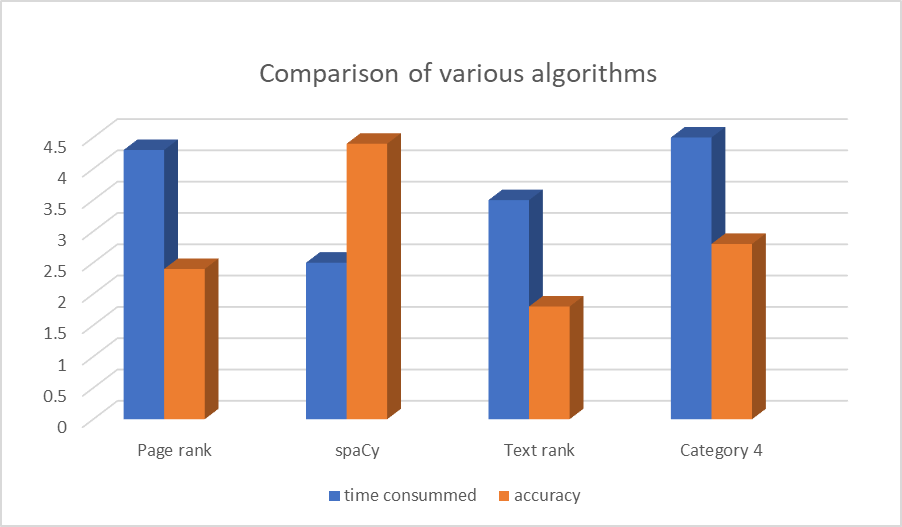
SUMMARY OF TEXT:

This is the last step of our project where the data collection, text extraction, data preprocessing, parts of speech, frequency calculation, scoring sentence, and summarization steps are combined along with GUI(Graphical User Interface) to provide the output that is a summary of original text.

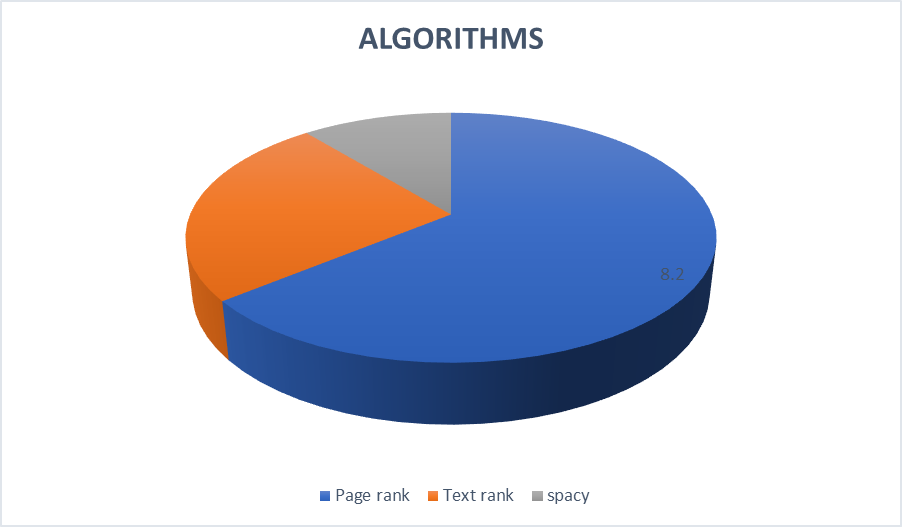
**7.RESULTS:**

****

Textrank+positional rank gives us the frequency of the words that were detected by spaCy. Whereas Textrank+similarity gives us the similarity of the text that were found in the model of the text. The similar words in the text are found and analyzed.



This graph shows us the comparison of various algorithms especially page rank, text rank and spaCy. The graph depicts us that the time consumed is more in page rank compared to text rank and SpaCy. Whereas the time consumed in spaCy is less compared to pagerank and textrank. The accuracy level of spaCy is higher compared to other algorithms both the text rank and page rank algorithms.



The above figure describes the algorithms that are in usage till now. Pagerank and text rank algorithms are widely in use. In future the demand for spacy will increase for text summarization depending upon the needs.

| Model | Training Loss | Validation Loss | Inference Time |
| --- | --- | --- | --- |
| PAGE RANK | Higher | 9.2942 | 0.008518 secs |
| SPACY (NLP) | Lower | 12.4217 | 0.015379 secs |

**9.DISCUSSION:**

Text summarizer helps in summarizing the content of the text from any sources such as article, newspaper, journal, researc paper, books using extractive summarization technique. The Extraction summarization technique is used for extracting the words from various resources such as books, etc and the sentences in the resources are tokenized and are assigned weights. Based upon the weights assigned to them the frequency is calculated, for each word in the sentence. The frequency is allotted based on the number of times the word appears on the text. Finally by using Parts of Speech (POS) and Sentence scoring the summarized text is displayed in the GUI. Parts of Speech helps in reducing the noise, helps in removing the unwanted words from the sentence thereby helps in shortening of the text. Sentence scoring is done after Parts of speech [POS] which helps in assigning scores for the sentences of the text. Thus the summarized text is displayed.

**8.CONCLUSION:**

Thus our solution aims at providing a summary of enormous volumes of text that gives better comprehension of text that is easy to read and does not confuse people. This is very useful in this world as people are busy in their jobs and they have little time to spare. To all these people this project is a one time solution that provides crisp information. The goal of this project is to cut short massive text and to provide a glimpse of text that is given as input so as to improve understanding along with saving time and improving user satisfaction. Further the project can be adopted in various fields such as in education, law, etc. Those people who find it difficult to read large content of text,

generally the common public people who at times feel very lethargic to read large content of the text can use this tool to read large text in less time. Thereby text summarizer is useful and is adopted in various domains[(6)](https://paperpile.com/c/8cREju/rJaG).The future scope of Text summarizer is it can be adopted with videos such as youtube, netflix, entertainment videos such as OTT channel videos can be uploaded and the text can be summarized. Especially for those people who feel very lazy to watch the full video or people who need to know the climax of a video soon can use the extension of Text Summarizer[(4)](https://paperpile.com/c/8cREju/O3Fr).

**10.REFERENCES:**

[1] [Elaziz, Mohamed Abd, Mohammed A. A. Al-qaness, Ahmed A. Ewees, and Abdelghani Dahou. 2019. *Recent Advances in NLP: The Case of Arabic Language*. Springer Nature.](http://paperpile.com/b/8cREju/3vec)

[2] [Fiori, and Alessandro. 2019. *Trends and Applications of Text Summarization Techniques*. IGI Global.](http://paperpile.com/b/8cREju/b9zs)

[3] [Hao, Tianyong, Xiaoyi Pan, Zhiying Gu, Yingying Qu, and Heng Weng. 2018. “A Pattern Learning-Based Method for Temporal Expression Extraction and Normalization from Multi-Lingual Heterogeneous Clinical Texts.” *BMC Medical Informatics and Decision Making* 18 (Suppl 1): 22.](http://paperpile.com/b/8cREju/bmEW)

[4] [Hardeniya, Nitin, Jacob Perkins, Deepti Chopra, Nisheeth Joshi, and Iti Mathur. 2016. *Natural Language Processing: Python and NLTK*. Packt Publishing Ltd.](http://paperpile.com/b/8cREju/O3Fr)

[5] [Hvitfeldt, Emil, and Julia Silge. 2021. *Supervised Machine Learning for Text Analysis in R*. CRC Press.](http://paperpile.com/b/8cREju/Me4u)

[6] [Kulkarni, Akshay, and Adarsha Shivananda. 2021. *Natural Language Processing Recipes: Unlocking Text Data with Machine Learning and Deep Learning Using Python*. Apress.](http://paperpile.com/b/8cREju/rJaG)

[7] [Mani, Inderjeet, and Mark T. Maybury. 1999. *Advances in Automatic Text Summarization*. MIT Press.](http://paperpile.com/b/8cREju/zIwx)

[8] [Millstein, Frank. 2020. *Natural Language Processing With Python: Natural Language Processing Using NLTK*. Frank Millstein.](http://paperpile.com/b/8cREju/Dnf7)

[9] [Nenkova, Ani, and Kathleen McKeown. 2011. *Automatic Summarization*. Now Publishers Inc.](http://paperpile.com/b/8cREju/glSv)

[10] [Priya, V., and K. Umamaheswari. 2023. *Computational Techniques for Text Summarization Based on Cognitive Intelligence*. CRC Press.](http://paperpile.com/b/8cREju/1CfF)

[11] [Torres-Moreno, Juan-Manuel. 2014. *Automatic Text Summarization*. John Wiley & Sons.](http://paperpile.com/b/8cREju/ojnR)

[12] [Wang, Ting, Chuan Yang, Maoyang Zou, Jiaying Liang, Dong Xiang, Wenjie Yang, Hongyang Wang, and Jia Li. 2024. “A Study of Extractive Summarization of Long Documents Incorporating Local Topic and Hierarchical Information.” *Scientific Reports* 14 (1): 10140.](http://paperpile.com/b/8cREju/LJQj)