International Journal of Engineering Science Invention (IJESI)
ISSN (Online): 2319-6734, ISSN (Print): 2319-6726
www.ijesi.org ||Volume 10 Issue 7 Series II || July 2021 || PP 07-11

Text Summarization Using NLP

[1.]M.Monika Rani [2.]A.Harika Sweta [3.]K Jaswani [4.]K Pavan Sidhu
[5.]Dr.D.N.V.S.L.S Indira
[1,2,3,4]B.Tech. Student, [5]Associate Professor
Department of Information Technology, Gudlavalleru Engineering College, Gudlavalleru, AP, INDIA

ABSTRACT
In this new era, where tremondous 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 problem of creating a short, accurate, and fluent summary of a longer text document. Automatic text summarization methods are greatly needed to address the ever-growing amount of text data available online to both better help discover relevant information and to consume relevant information faster.There is an enormous amount of textual material, and it is only growing every single day.Think of the internet, comprised of web pages, news articles, status updates, blogs and so much more. There is a great need to reduce much of this text data to shorter, focused summaries that capture the salient details.
KEY WORDS: *Text summarization, Natural Language Processing*

Date of Submission: 12-07-2021 Date of Acceptance: 28-07-2021

I. INTRODUCTION

Before going to the Text summarization, first we, have to know that what a summary is. 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.

- **Extraction-based summarization**

The extractive text summarization technique involves pulling keyphrases from the source document and combining them to make a summary. The extraction is made according to the
defined metric without making any changes to the texts.
Example: Joseph and Marry rode on a donkey to **attend** the annual **event** in **Jerusalem**. In the city, **Mary** gave **birth** to a child named **Jesus.**
Extractive summary: Joseph and Marry attend event Jerusalem. Mary birth Jesus.
As you can see above, the words in bold have been extracted and joined to create a summary although sometimes the summary can be grammatically strange.

- **Abstraction-based summarization**

The abstraction technique entails paraphrasing and shortening parts of the source document. When abstraction is applied for text summarization in deep learning problems, it can overcome the grammar inconsistencies of the extractive method.
The abstractive text summarization algorithms create new phrases and sentences that relay the most useful information from the original text — just like humans do.
Here is an example:
Therefore, abstraction performs better than extraction. However, the text summarization algorithms required to do abstraction are more difficult to develop; that's why the use of extraction is still popular.
Here is an example
Abstractive summary: Joseph and Mary came to Jerusalem where Jesus was born.

1.1Natural Language Processing

Simply and in short, natural language processing (NLP) is about developing applications and services that are able to understand human languages. We are talking here about practical examples of natural language processing (NLP) like speech recognition, speech translation, understanding complete sentences, understanding synonyms of matching words, and writing complete grammatically correct sentences and paragraphs.

1.2 Text Summarization

Text summarization is a subdomain of Natural Language Processing (NLP) that deals with extracting summaries from huge chunks of texts. There are two main types of techniques used for text summarization: NLP-based techniques and deep learning-based techniques. we will see a NLP-based technique for text summarization. We will not use any machine learning library in this article. Rather we will simply use Python's NLTK library for summarizing Wikipedia articles.

1.3 Flask

Flask is a lightweight WSGI web application framework. It is designed to make getting started quick and easy, with the ability to scale up to complex applications. It began as a simple wrapper around Werkzeug and Jinja and has become one of the most popular Python web application frameworks.

II. PROPOSED SYSTEM

This project basic idea is to reduce the large chunks of data into important points,so that reader can reduce time of reading entire paragraph. This project uses python and Flask web framework. Text summarizer extracts the major points and provides a required sentence synopsis for easy reading. Text summarizer picks out the most important sentences from the text so the user only have to read through what's important. Here the user give http link or any text data in text area and user also specify the required number of lines to html page. Then the required data is displayed by using NLP based techniques.

The main objective of this project is to to reduce the time of reading the entire paragraph. The user gives any http link or any text in text area and the user is also allowed to specify the required number of lines to html page. Then the important points of the paragraph is extracted and displayed by using NLP based techniques.

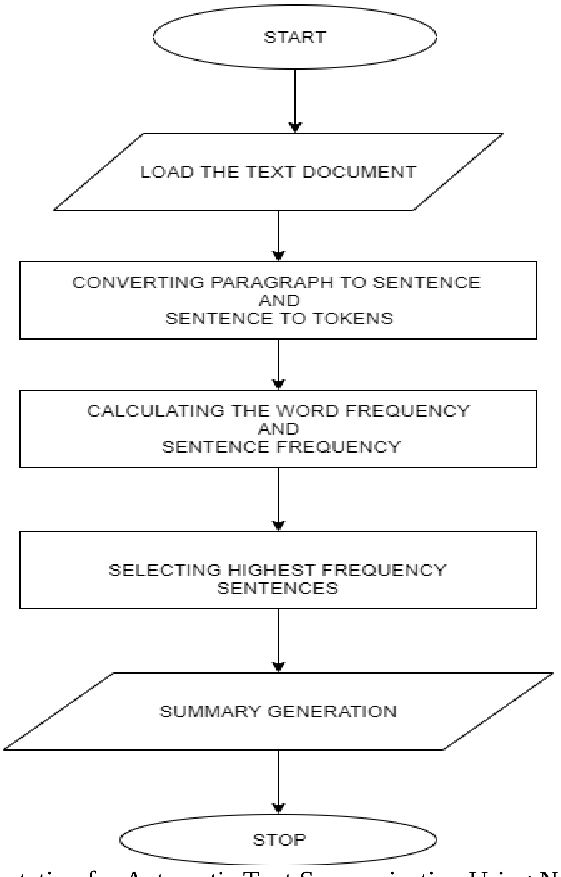


Fig 1: Overall Representation for Automatic Text Summarization Using Natural Language Processing.

2.1 System Architecture Of The Proposed System

The proposed system depicts the two stages for Text Summarization and they are listed below.
Stage 1: Data Pre-Processing
Stage 2: Processing

Stage 1: Data Pre-Processing

Pre Processing is structured representation of the original text. It usually includes:
a) Sentences boundary identification. In English, sentence boundary is identified with presence of dot at the end of sentence.
b) Stop-Word Elimination—Common words with no semantics
c) Stemming—The purpose of stemming is to obtain the stem or radix of each word, which emphasize its semantics.

Stage 2: Processing

In Processing step, features influencing the relevance of sentences are decided and calculated and the weights are assigned to these features using weight learning method. Final score of each sentence is determined using Feature-weight equation. Top ranked sentences are selected for final summary. Summary evaluation is a very important aspect for text summarization. Generally, summaries can be evaluated using intrinsic or extrinsic measures. While intrinsic methods attempt to measure summary quality using human evaluation and extrinsic methods measure the same through a task-based performance measure such the information retrieval oriented task.

Text Summarization Steps

Convert Paragraphs to Sentences

First the user need to convert the whole paragraph into sentences. The most common way of converting paragraphs to sentences is to split the paragraph whenever a period is encountered. So if the user split the paragraph under discussion into sentences.

Text Preprocessing

After converting paragraph to sentences, user need to remove all the special characters, stop words and numbers from all the sentences.

Tokenizing the Sentences

The user need to tokenize all the sentences to get all the words that exist in the sentences. After tokenizing the sentences.

Find Weighted Frequency of Occurrence

Next user need to find the weighted frequency of occurrences of all the words. User can find the weighted frequency of each word by dividing its frequency by the frequency of the most occurring word.

Replace Words by Weighted Frequency in Original Sentences

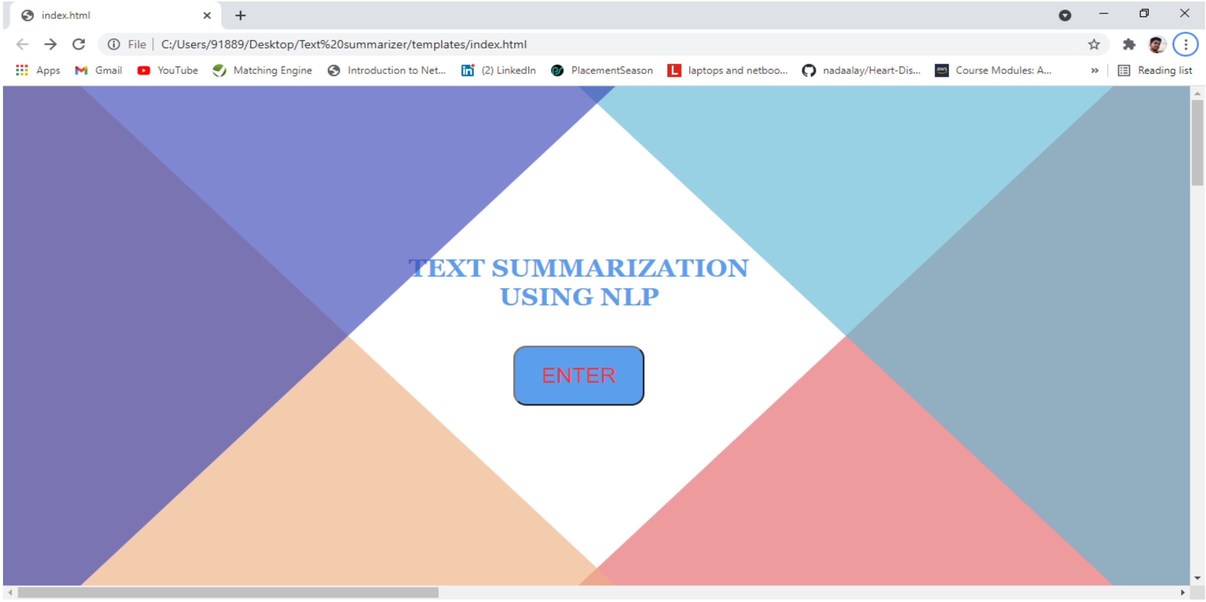
The final step is to plug the weighted frequency in place of the corresponding words in original sentences and finding their sum. It is important to mention that weighted frequency for the words removed during preprocessing (stop words, punctuation, digits etc.) will be zero and therefore is not required to be added.

Sort Sentences in Descending Order of Sum

The final step is to sort the sentences in inverse order of their sum. The sentences with highest frequencies summarize the text.

III. OUTPUT AND DISCUSSION

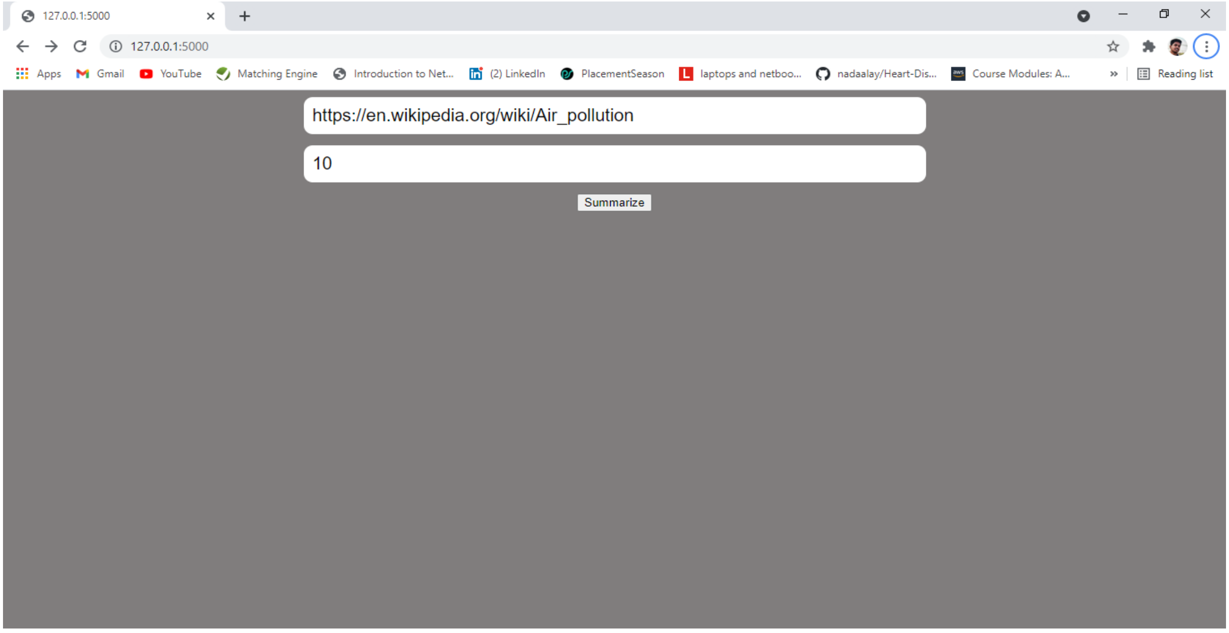
To execute the project an IDE is required. In this project spyder has been used. First load the project into the IDE. Now all the packages which are required should be installed. Clean and build the project and run the whole project. The results can be seen in http://127.0.0.1:5000 address.



TEXT SUMMARIZATION USING NLP

Fig 2:Starting Page

As shown in Fig 2 after execution of project this is the first web page of after user click ENTER it redirects to next web page.

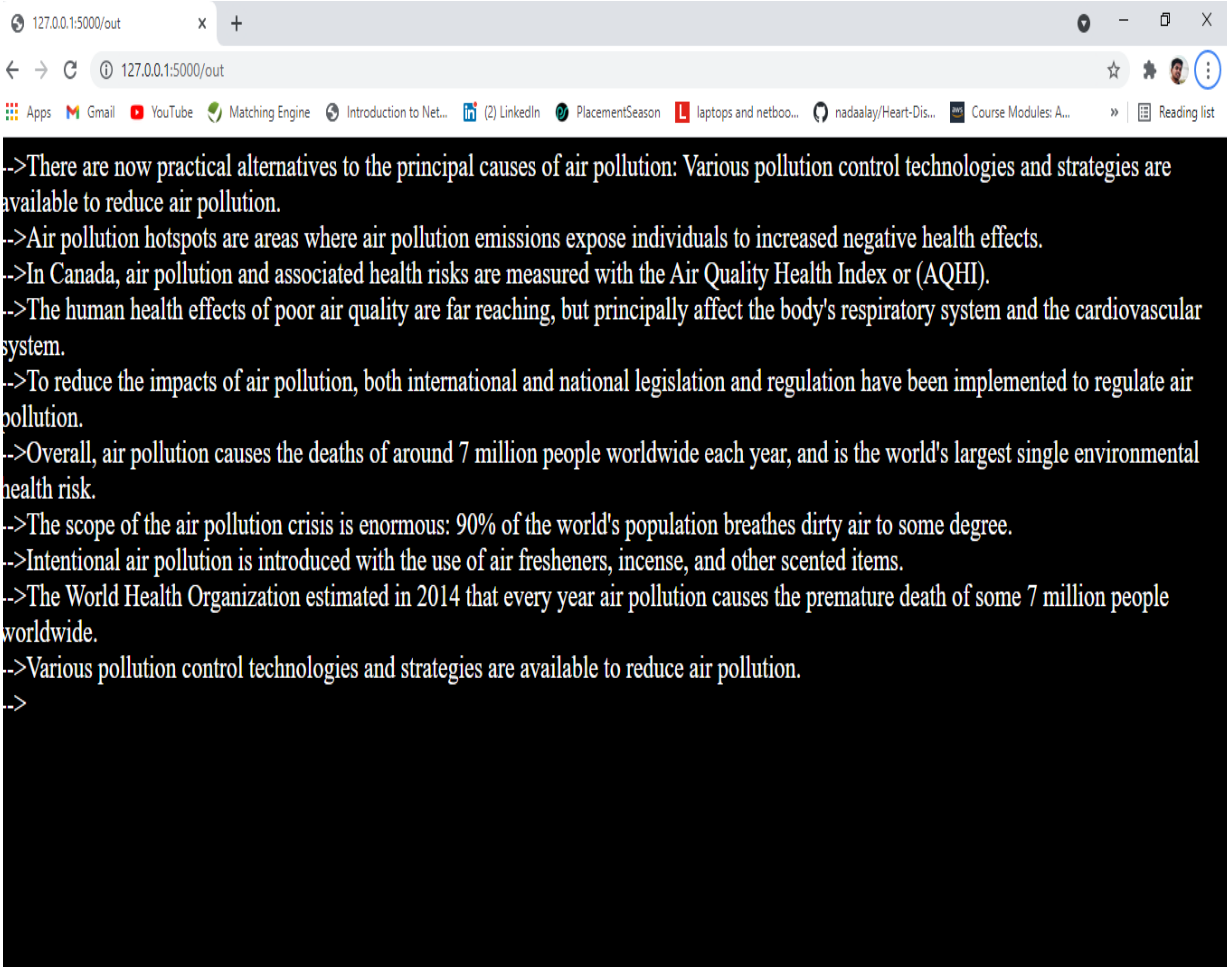


https://en.wikipedia.org/wiki/Air_pollution

10

Fig 3: Input Page

As shown in Fig 3 here the user give any paragraph or http link in the first text field and user specify the number of lines in the next text filed and we click summerize button.



-->There are now practical alternatives to the principal causes of air pollution: Various pollution control technologies and strategies are available to reduce air pollution.
-->Air pollution hotspots are areas where air pollution emissions expose individuals to increased negative health effects.
-->In Canada, air pollution and associated health risks are measured with the Air Quality Health Index or (AQHI).
-->The human health effects of poor air quality are far reaching, but principally affect the body's respiratory system and the cardiovascular system.
-->To reduce the impacts of air pollution, both international and national legislation and regulation have been implemented to regulate air pollution.
-->Overall, air pollution causes the deaths of around 7 million people worldwide each year, and is the world's largest single environmental health risk.
-->The scope of the air pollution crisis is enormous: 90% of the world's population breathes dirty air to some degree.
-->Intentional air pollution is introduced with the use of air fresheners, incense, and other scented items.
-->The World Health Organization estimated in 2014 that every year air pollution causes the premature death of some 7 million people worldwide.
-->Various pollution control technologies and strategies are available to reduce air pollution.
-->

Fig 4 Output page

As shown in Fig 4 after clicking summerize button the user get the above web page by summarizing the paragraph into specified number of lines based on NLP technique.

IV. CONCLUSION AND FUTURE SCOPE

This project is designed such a way that the user is allowed to summerize any http link or text data. This project mainly helps to reduce the large chunks of data into brief summary.It selects the highest frequency words in the paragraph and selects the sentence score and highest sentence score will be selected and generates required summary. At present it just takes the frequency of the words in text and returns the output.

In future it can be developed using advanced machine learning techniques to give its exact meaning of the paragraph .This project can also deployed into android app with many features like saving in the device, google drive. To this project login page can be added and previous summarised data can be retrieved using login credentials. User can add the notes to the end of each page after summarisation.

REFERENCES

[1]. Li, Ailin, et al. "The Mixture of Text rank and Lexrank Techniques of Single Document Automatic Summarization Research in Tibetan." 2016 8th International Conference on Intelligent HumanMachine Systems and Cybernetics (IHMSC). Vol. 1. IEEE, 2016.
[2]. https://glowingpython.blogspot.com/2014/09/text-summarization-with-nltk.html
[3]. M. Allahyari, S. Pouriyeh, M. Assefi et al., “Text summarization techniques: a brief survey,” International Journal of Advanced Computer Science and Applications, vol. 8, no. 10,2017.
[4]. A. B. Al-Saleh and M. E. B. Menai, “Automatic Arabic text summarization: a survey,” Artificial Intelligence Review, vol. 45, no. 2, pp. 203–234, 2016.
[5]. A. Mahajani, V. Pandya, I. Maria, and D. Sharma, “A comprehensive survey on extractive and abstractive techniques for text summarization,” in Ambient Communications and Computer Systems, Y.-C. Hu, S. Tiwari, K. K. Mishra, and M. C. Trivedi, Eds., vol. 904, pp. 339–351, Springer, Singapore,2019.
[6]. T. Shi, Y. Keneshloo, N. Ramakrishnan, and C. K. Reddy, Neural Abstractive Text Summarization with Sequence-ToSequence Models: A Survey, http://arxiv.org/abs/1812.02303,2020.
[7]. T. Cai, M. Shen, H. Peng, L. Jiang, and Q. Dai, “Improving transformer with sequential context representations for abstractive text summarization,” in Natural Language Processing and Chinese Computing, J. Tang, M.-Y. Kan, D. Zhao, S. Li, and H. Zan, Eds., pp. 512–524, Springer International Publishing, Cham, Switzerland, 2019.