Automatic Text Summarization Using NLP

Dharmaiah Devarapalli[1], Pavani Pasupuleti[2], Sathi Navya Vahini Reddy[3], Srija Padmini Guduri[4], Raja Pranathi Vemuri[5], Jaya Madhuri Pericharla[6]

[1]Professor, Department of Computer Science and Engineering, Shri Vishnu Engineering College for Women, Bhimavaram, India" [2-6]"Department of Computer Science and Engineering, Shri Vishnu Engineering College for Women, Bhimavaram, India,"
{pavanipasupuleti9, navyasathi444,
guduri.srija, rajapranathi456,

Abstract: The process of compressing a piece of text into a smaller version is known as summarising, reducing the size of the original text while maintaining important informational aspects and content meaning. Summary sentences from input documents can now be easily retrieved by automatic text summarizing systems. However, it has a number of flaws, including erroneous extraction of crucial sentences, inadequate coverage, poor sentence coherence, and duplication. The spacy package in Python is used to propose a new concept of summarizer technique in this work. It takes the most important information from the text and extracts it.In order to compute the word frequency, the scoring system is also employed to calculate the score for the words.The results reveal that the proposed method takes less time to complete the summary process than the existing algorithm.The text to summary converter is a web-based utility that assists in material summarization. We can upload our data, and this application will provide us with a summary of it. The main purpose is to create accurate summaries of the information entered. Extraneous sentences will be deleted in order to get to the most important sentences.

Keywords: Text to summary converter, Spacy, NLP, stop words, Word frequencies

1. INTRODUCTION

We've noticed an upsurge in the amount of textual information available in recent years. The amount of textual data being produced is continually growing[2]. The user's ability to read the textual information gets progressively difficult, and as a result, the user loses interest. To overcome this problem, Text Ssummarization was devised. As a result of huge breakthroughs in software and hardware technology, data mining has experienced rapid expansion in recent years. As technology advances, more sorts of data become available, which is especially beneficial for text data.[7]. The rapid production of enormous stores of diverse types of data has been facilitated by social network and web software and hardware platforms. Structured data is often maintained by a database system, whereas text data is typically managed by a search engine due to the lack of structures.[9]. The search engine allows the online user to find the relevant information from the gathered works using a keyword query.[5]. A way of compressing a lengthy original text into a more brief format, resulting in a summary of the original topic, is known as text summarizing[14]. The summary is based on the most important parts and key ideas of the original text. As a result, the reader is given a sense of the original text as well as a focused view of it.

2. PROBLEM DEFINITION

There has been a recent increase in the amount of text data available from a variety of sources. This volume of literature is a great source of knowledge and information that, in order to be useful, must be properly summarized. The problem's primary purpose is to automatically summarize the text[5]. As the Internet has developed tremendously, people are becoming overwhelmed by the large amount of online information and articles[3]. The increased supply of papers necessitates further research into computerized text summarization. It will also say the count of words before and after summarizing.

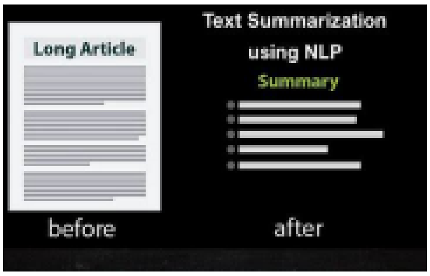


Fig 1: Text To Summary

.

3. METHODOLOGY

3.1 Natural Language Processing(NLP)

NLP is simply defined as teaching an algorithm to read and analyze human (natural) languages in the same way that a human does, but faster, more correctly, and on much larger datasets[5]. Producing a summary of textual content used to be a manual process.

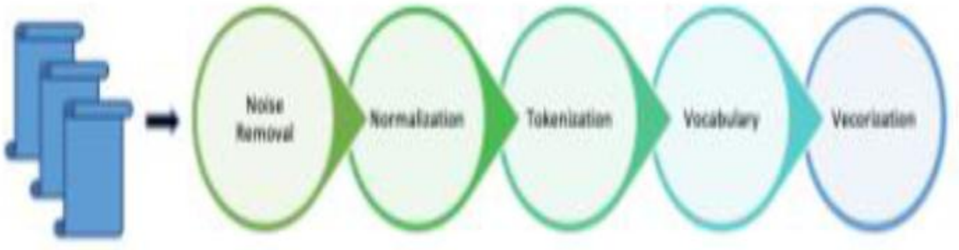


Fig 2: NLP Workflow

3.2 SpaCy:

SpaCy is a brand-new Python library for "Industrial-strength Natural Language processing" SpaCy is a very young NLP library that hasn't gained as much traction as NLTK [13]. It is intended for usage in production environments, and it can assist us in developing applications that efficiently process large amounts of text [15].

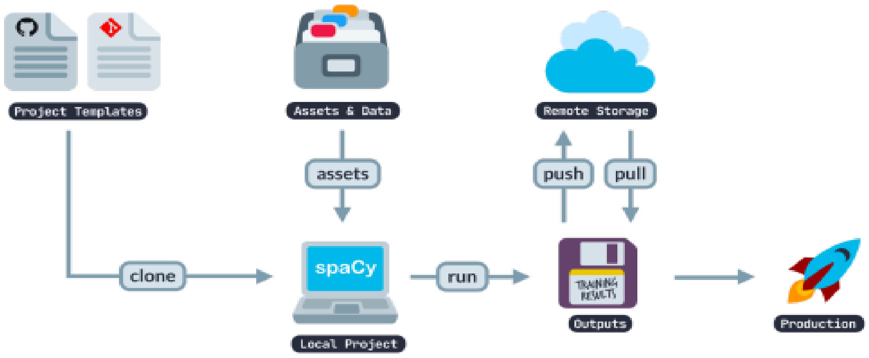


Fig 3.spaCy

3.3 Heapq Library:

Heaps are binary trees in which each parent node has the same or lower value than any of its descendants.[12]. This method uses arrays with heap[k]= heap[2*k+1] for all k, counting items from zero. [14]Non-existing items are treated as infinite for the sake of comparison. The smallest element of a heap is always the root.

3.4 String:

The Python String module contains string manipulation constants, utility functions, and classes[17].

4. DATA SET DESCRIPTION

4.1 Data Gathering:

Data collection requires obtaining information from a variety of sources. In our project, we provide a large amount of text from other sources such as Wikipedia, newspapers, and so on:

4.2 A Pre−Processing Step:

Text is an incredibly rich source of data. Hundreds of millions of new emails and text messages are sent every minute[16]. There's a mound of text data just ready to be mined for information. There are different steps in pre−processing of data i.e Stop words, punctuation marks, and uppercase words were removed, Tokenization, Parts of Speech (POS) Tagging, Entity Detection etc[1]:

4.3 Tokenization:

Tokenization is the process of separating text into tokens and omitting characters such as punctuation marks (,. " ') and spaces. The tokenizer in spaCy accepts unicode text as input and produces a sequence of token objects[13]. .Breaking up the text into individual words is referred to as word tokenization. Many language processing systems require input in the form of individual words rather than lengthy strings of text, therefore this is an important step[11].

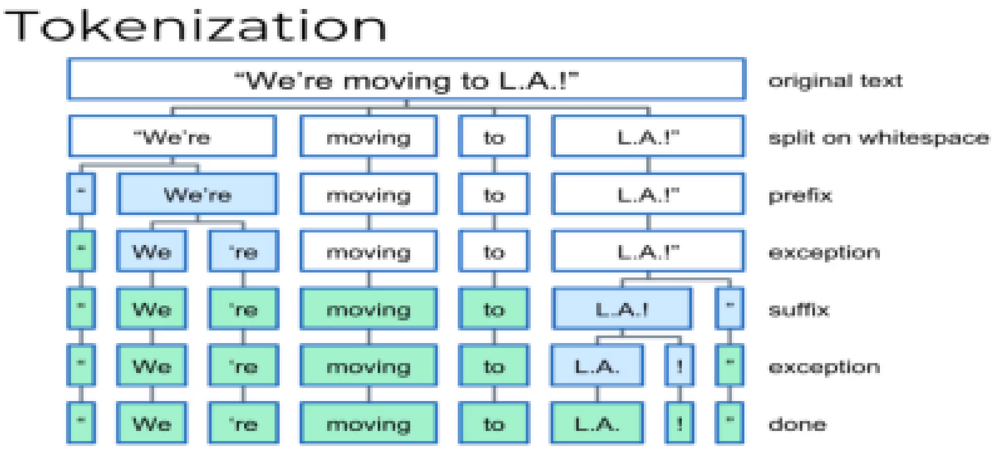


Fig 4: tokenization

4.4. Designing the model:

The model's design is the next step.

1) Text Cleaning : Stop words, punctuation marks were removed and uppercase words were removed and replaced with lowercase[3].

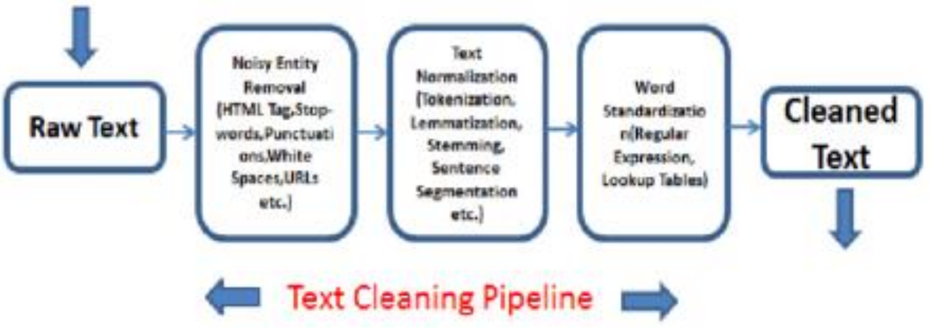

Raw Text
Noisy Entity Removal (HTML Tag,Stop-words,Punctuations,White Spaces,URLs etc.)
Text Normalization (Tokenization, Lemmatization, Stemming, Sentence Segmentation etc.)
Word Standardization(Regular Expression, Lookup Tables)
Cleaned Text
Text Cleaning Pipeline

Fig 5: Text cleaning

2) Word Tokenization: Tokenize each word from sentences[7].

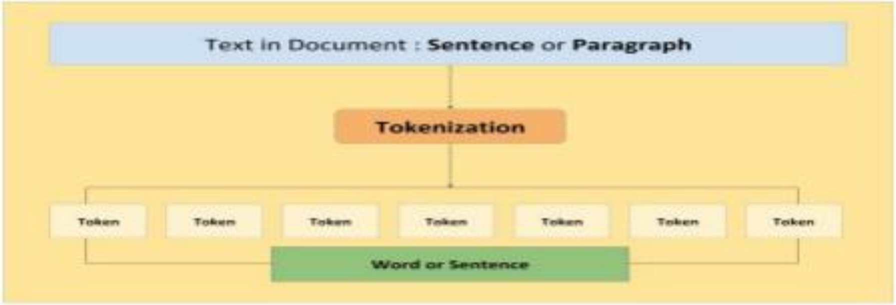


Fig 6: Word Tokenization

3) Word Frequency table: To get the normalised word frequency count, count the frequency of each word and divide the maximum frequency by each frequency.[8]

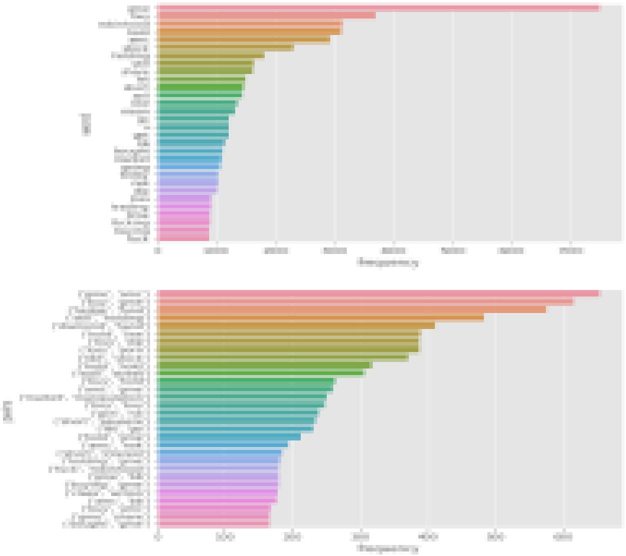


Fig 7: Word Frequencies

4) Sentence Tokenization: As per frequency of sentence then

5) Summarization

These are the five steps that go into creating our model.

4.5.Testing the model:

These are the several stages to evaluate that this the model guarantees that it can accurately forecast. The first step is to anticipate the testing set.. Figure 8,9,10 shows the results we got during testing the model almost the same as we predicted[10].

4.6. Model implementation:

```
ALGORITHM:
INPUT : Text
OUTPUT : Summary
STEP 1 :import packages
STEP 2 : Text preprocessing
STEP 3 : word Tokenization, split the text into words
STEP 4 : Word Frequency table: To acquire the normalised word frequency count,
count the frequency of each word and divide the maximum frequency by each
frequency.[8].
STEP 5 : Sentence Tokenization: As per frequency of sentence
STEP 6: Summarize
STEP 7 : Save the model for future use
```

5. RESULTS

We were so close to the projected outcomes. This simple web application makes extracting a summary from text like a cake walk[20].These are results obtained. We observed that these results are up to our expectations, but also missing some important points. Hopefully in the future, we are going to update it for a better experience. We can also obtain the word count before and after summarizing.

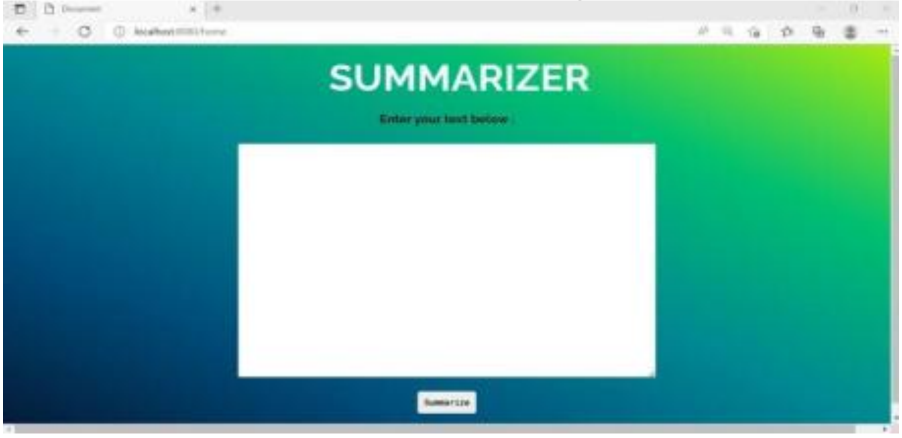
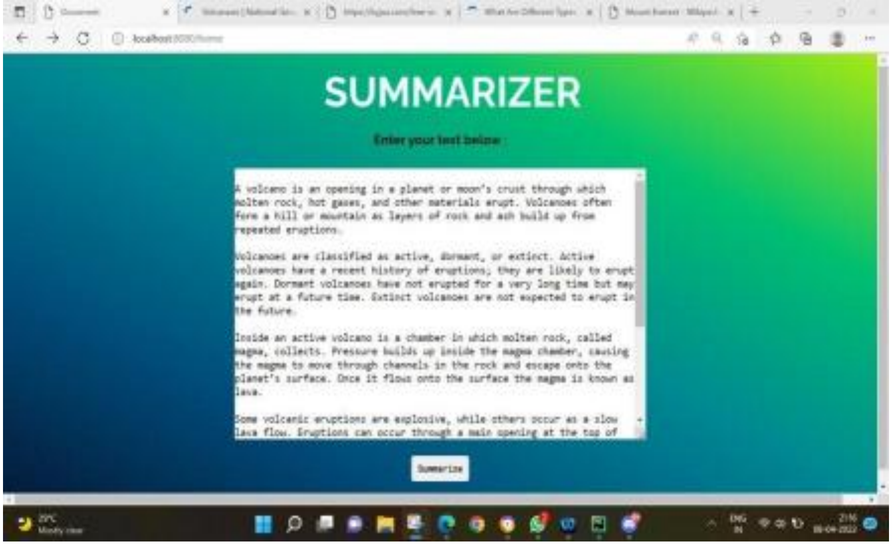


Fig 8:Initial web page



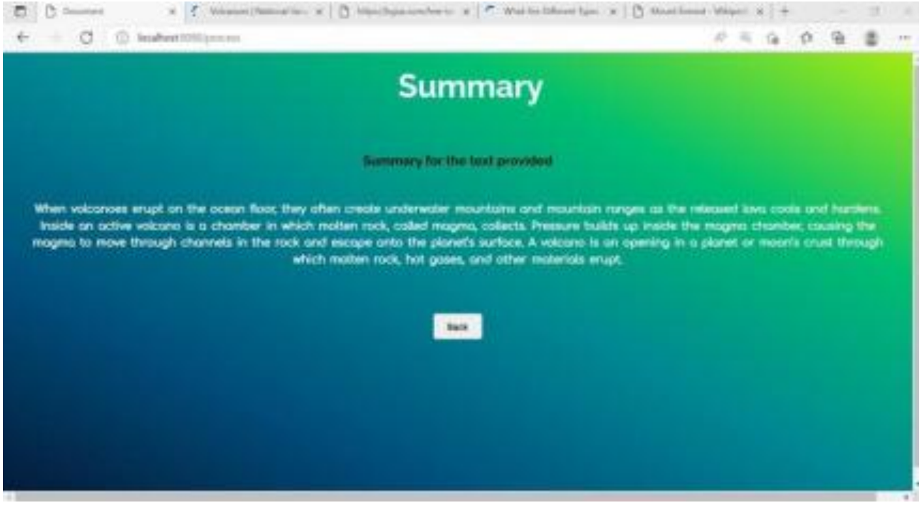
A volcano is an opening in a planet or moon's crust through which molten rock, hot gases, and other materials erupt. Volcanoes often form a hill or mountain as layers of rock and ash build up from repeated eruptions.

Volcanoes are classified as active, dormant, or extinct. Active volcanoes have a recent history of eruptions; they are likely to erupt again. Dormant volcanoes have not erupted for a very long time but may erupt at a future time. Extinct volcanoes are not expected to erupt in the future.

Inside an active volcano is a chamber in which molten rock, called magma, collects. Pressure builds up inside the magma chamber, causing the magma to move through channels in the rock and escape onto the planet's surface. Once it flows onto the surface the magma is known as lava.

Some volcanic eruptions are explosive, while others occur as a slow lava flow. Eruptions can occur through a main opening at the top of

Fig 9 : Entering text into it



When volcanoes erupt on the ocean floor, they often create underwater mountains and mountain ranges as the released lava cools and hardens. Inside an active volcano is a chamber in which molten rock, called magma, collects. Pressure builds up inside the magma chamber, causing the magma to move through channels in the rock and escape onto the planet's surface. A volcano is an opening in a planet or moon's crust through which molten rock, hot gases, and other materials erupt.

Fig 10: Summary Obtained

6. CONCLUSION

Increasing your productivity by reducing the amount of time you spend reading can have a significant impact. Whether you're reading papers or academic journals, Python and SpaCy's natural language processing capabilities can help you save time without sacrificing information quality**.**This is simply one of the methods for obtaining text summary by calculating the most essential phrases using the most frequently used words. Other options include utilizing the nltk library to perform lexical analysis, a part of speech tagger, and n-grams. As additional resources become available, we intend to continue to maintain and expand these packages.

FUTURE SCOPE

The proposed work does not involve a summarizer for the entire notebook. Hopefully, with potential future work we can increase the quality of the summarizer and make it work efficiently[21].

ACKNOWLEDGEMENT

"This work is supported by the Department of Computer Science and Engineering, Shri Vishnu Engineering College for Women, Bhimavaram, India."

REFERENCES

1. Deepa, R., Konshi, J., Haritha, A. and Shobini, K. (n.d.). Automatic Text Summarization System. [online] Available at: https://www.ripublication.com/ijaerspl2019/ijaerv14n5spl_04.pdf [Accessed 19 Jun. 2022].

2. Johnson, M.E. (2018). Automatic Summarization of Natural Language. arXiv:1812.10549 [cs, stat]. [online] Available at: https://arxiv.org/abs/1812.10549 [Accessed 30 Jun. 2022].

3.Maybury, M. (1999). Advances in Automatic Text Summarization. [online] Google Books. MIT Press. Available at: https://books.google.co.in/books?hl=en&lr=&id=YtUZQaKDmzEC&oi=fnd&pg=PA81&dq=EduardHovyand Chin+Yew+Lin.Automated+text+summarization+in+SUMMARIST.+MIT+Press [Accessed 30 Jun. 2022].

4.Mahdipour, E. (2014). Automatic Persian Text Summarizer Using Simulated Annealing and Genetic Algorithm. International Journal of Intelligent Information Systems, 3(6), p.84. doi:10.11648/j.ijiis.s.2014030601.26.

5.Srikanth, P. and Deverapalli, D. (2017). CFTDISM:Clustering Financial Text Documents Using Improved Similarity Measure. [online] IEEE Xplore. doi:10.1109/ICCIC.2017.8524466.

6. S.C. and Joshi, A. (2017). Information and Communication Technology for Intelligent Systems (ICTIS 2017) - Volume 2. [online] Google Books. Springer. Available at:https://www.google.co.in/books/edition/Information_and_Communication_Technology/fE8xDwAAQBAJ?hl=en&gbpv=1&dq=J.+Leskovec [Accessed 30 Jun. 2022].

7.Kupiec, J.M. and Schuetze, H. (n.d.). System for genre-specific summarization of documents. [online] Available at: https://patents.google.com/patent/US6766287B1/en [Accessed 30 Jun. 2022].

8..Doran, W., Stokes, N., Carthy, J. and Dunnion, J. (n.d.). Comparing Lexical Chain-based Summarisation Approaches Using an Extrinsic Evaluation. [online] Available at: http://www.oriyana.cz/id32402/jazyk/jazykove(2da/aplikovana(1_lingvistika/Ontologie/WordNet/Conference_2004/103.pdf [Accessed 30 Jun. 2022].

9. Goldstein, J., Kantrowitz, M., Mittal, V. and Carbonell, J. (1999). Summarizing text documents. Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '99. doi:10.1145/312624.312665.

10. Luhn, H.P. (1958). The Automatic Creation of Literature Abstracts. IBM Journal of Research and Development, [online] 2(2), pp.159–165. doi:10.1147/rd.22.0159.

11.Kedar Bellare, Anish Das Sharma, Atish Das Sharma, Navneet Loiwal and Pushpak Bhattacharyya.Generic Text Summarization Using Wordnet. Language Resources Engineering Conference (LREC 2004), Barcelona, May, 2004.

12 . Satapathy, S.C. and Joshi, A. (2017). Information and Communication Technology for Intelligent Systems (ICTIS 2017) - Volume 2. [online] Google Books. Springer.