A

Project Report

On

TEXT-SUMMARIZATION USING NLP

Submitted to

RAJIV GANDHI UNIVERSITY OF KNOWLEDGE AND TECHNOLOGIES

RK VALLEY

in partial fulfilment of the requirement for the award of the Degree of

BACHELOR OF TECHNOLOGY

In

COMPUTER SCIENCE & ENGINEERING

Submitted by

BANDLA SNEHA (R180133)

BEDADALA JAYASURYA NARAYANA REDDY (R180818)

Under the Guidance of

Mr. P.SANTOSH KUMAR, Assistant Professor

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

RAJIV GANDHI UNIVERSITY OF KNOWLEDGE TECHNOLOGIES

(catering the Educational Needs of Gifted Rural Youth of AP)

R.K Valley, Vempalli(M), Kadapa(Dist) – 516330

2020 – 2024

RAJIV GANDHI UNIVERSITY OF KNOWELDEGE TECHNOLOGIES

(A.P.Government Act 18 of 2008)

RGUKT -RK Valley

Vempalli, Kadapa , Andhra Pradesh -516330

CERTIFICATE OF PROJECT COMPLETION

This is to certify that the work entitled **"Text-Summarization Using NLP"** is bonafide work of **Bandla Sneha(R180133), Bedadala Jaya Surya Narayana Reddy(R180818)** carried out under our guidance and supervision for the partial fulfilment for the degree of Bachelor of Technology in Computer Science and Engineering during the academic session August 2023-December 2023 at RGUKT-RK VALLEY.

Project Guide
Mr P Santhosh Kumar
Asst.Prof. in Dept of CSE,
RGUKT-RK Valley.

Head of the Department
Mr. N.Satyanandaram,
Lecturer in Dept of CSE,
RGUKT-RK Valley

RAJIV GANDHI UNIVERSITY OF KNOWELDEGE TECHNOLOGIES

(A.P.Government Act 18 of 2008)

RGUKT -RK Valley

Vempalli, Kadapa , Andhra Pradesh -516330

DECLARATION

We, **Bandla Sneha(R180133),Bedadala Jaya Surya Narayana Reddy(R180818)** hereby declare that the project report entitled "**Text-Summarization using NLP**" done under guidance of **Mr. P.Santhosh Kumar** is submitted in partial fulfilment for the degree of Bachelor of Technology in Computer Science and Engineering during the academic session February 2023 – July 2023 at RGUKT-RK Valley. we also declare that this project is a result of our own effort and has not been copied or imitated from any source. Citations from any websites are mentioned in the references. To the best of my knowledge, the results embodied in this dissertation work have not been submitted to any university or institute for the award of any degree or diploma.

Date:

Place:

Bandla Sneha - R180133

Bedadala Jaya Surya Narayana Reddy - R180818

ACKNOWLEDGEMENT

I would like to express my deep sense of gratitude & respect to all those people behind the screen who guided, inspired and helped me crown all my efforts with success. I wish to express my gratitude to **P.Santhosh kumar** for his valuable guidance at all stages of study, advice, constructive suggestions, supportive attitude and continuous encouragement, without which it would not be possible to complete this project.

I would also like to extend our deepest gratitude & reverence to the Director of RGUKT, RK Valley **Dr.A V S S Kumar Swami Gupta** and HOD of Computer Science and Engineering **Mr. N. Satyanandaram** for their constant support and encouragement.

Last but not least I express my gratitude to my parents for their constant source of encouragement and inspiration for me to keep my morals high

With Sincere Regards,

Bandla Sneha – R180133

Bedadala Jaya Surya Narayana Reddy -R180818

TABLE OF CONTENT Page no

ABSTRACT

The amount of data on the Internet has increased exponentially over the past decade. Therefore, we need a solution that converts this massive amount of raw information into useful information that the human brain can understand. One such common technique in research that helps when dealing with large amounts of data is text summarization. Automatic summarization is a well-known approach to reduce documents to key ideas. This works by storing important information by creating a shortened version of the text. Text summaries are divided into extraction and abstraction methods. The extraction summary method minimizes the summarization burden by selecting a subset of relevant sentences from the actual text. There are many methods, but researchers specializing in **natural language processing (NLP)** are particularly attracted to the extraction method. The meaning of the sentence is calculated using linguistic and statistical features. In this work, extractive and abstract methods for summarizing texts were examined. This white paper uses a **spacey algorithm** to analyze the above methods, resulting in fewer iterations and a more focused summary

Chapter 1

Introduction

1.1. INTRODUCTION

Due to the massive volume of textual content that is generated on the Internet and in numerous archives of news stories, scientific papers, legal documents, etc., text summarization is becoming much more crucial. With the enormous amount of textual content, manual text summarizing takes a lot of time, effort, money, and even becomes impracticable.

The two types of Text summarization techniques are

1.extractive : The extractive method chooses the key phrases from the source documents and then concatenates them to create the summary

2.abstractive: The abstractive approach generates the summary using sentences that are distinct from the original sentences after representing the input documents in an intermediary form.

Therefore, generating a system for providing a summary of the enormous amount of text material, machine learning and natural language processing techniques will be used. To use SpaCy for natural language processing to create a system that recognizes contexts in a document and provides the best possible summarization.

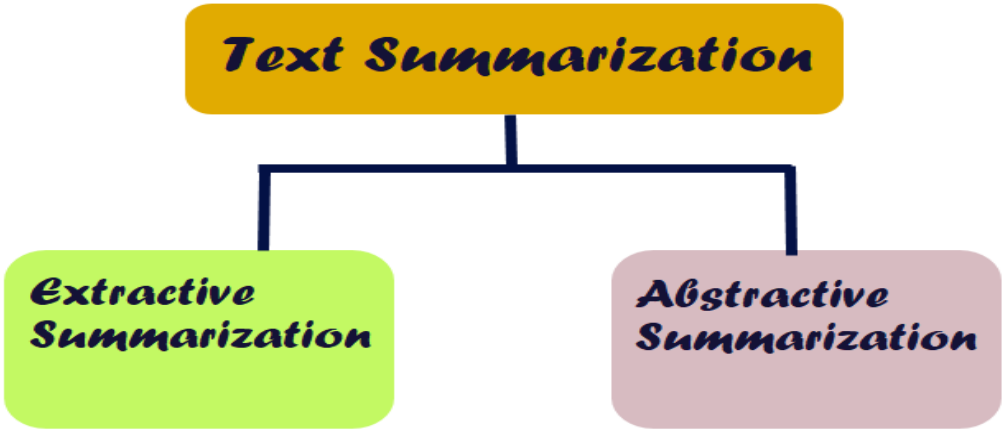


Fig 1.1

1.2.Purpose

- Automatically condense and rewrite a large chunk of text to create a small, crisp summary.
- Give the reader most of the information present in the original document while also ensuring that no information has been lost during condensation.
- Bring out information that is crucial, and also ensure that the meaning of the paragraph stays the same.
- Reduce the time to understand large papers like research articles, without skipping any vital information.
- Reduce reading time.
- Make the selection process easier when researching documents.
- Improve the effectiveness of indexing.
- Provide personalized information in question-answering systems.
- Summarizing strategies is adopted in almost every area of studies or industry.

1.3.Technologies used

Python

Natural Language processing(NLP)

Chapter 2

Technologies and Libraries

2.1 Natural Language Processing(NLP)

Natural language processing (NLP) is a subfield of linguistics, computer science, and artificial intelligence , in particular how to program computers to process and analyze large amounts of natural language data. NLP is currently the focus of significant interest in the machine learning community. Some of the use cases for NLP are listed here:

- Chatbots
- Search(text and Audio)
- Text Classification
- Sentiment Analysis
- Recommendation System
- Question Answering
- Speech recognization
- NLU (Natural Language Understanding)
- NLG (Natural Language Generation)

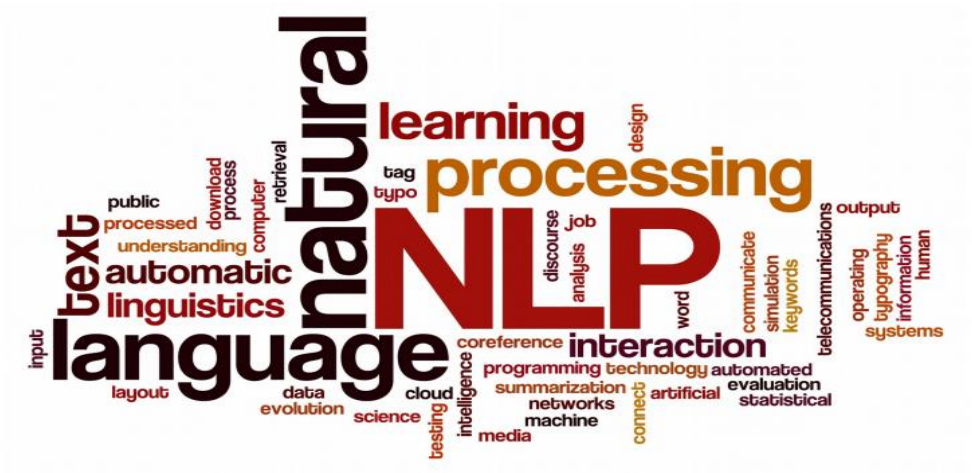


Fig 2.1

2.2. SpaCy Library

spaCy is a free, open-source advanced natural language processing library, written in the programming languages Python and Cython. spaCy mainly used in the development of production software and also supports deep learning workflow via statistical models of PyTorch and TensorFlow. spaCy provides a fast and accurate syntactic analysis, named entity recognition and ready access to word vectors. We can use the default word vectors or replace them with any you have. spaCy also offers :

- Tokenization
- Sentence boundary detection
- POS tagging
- Syntactic parsing
- integrated word vectors
- alignment into the original string with high accuracy.

Chapter 3

Text-Summarization Procedure

3.1. Steps to Text Summarization:

1) Text Cleaning: Removing stop words, punctuation marks and making the words in lower case.

2) Word Tokenization: Tokenize each word from sentences.

3) Word Frequency table: Count the frequency of each word and then divide the maximum frequency with each frequency to get the normalized word frequency count.

4) Sentence Tokenization: As per frequency of sentence then

5) Summarization

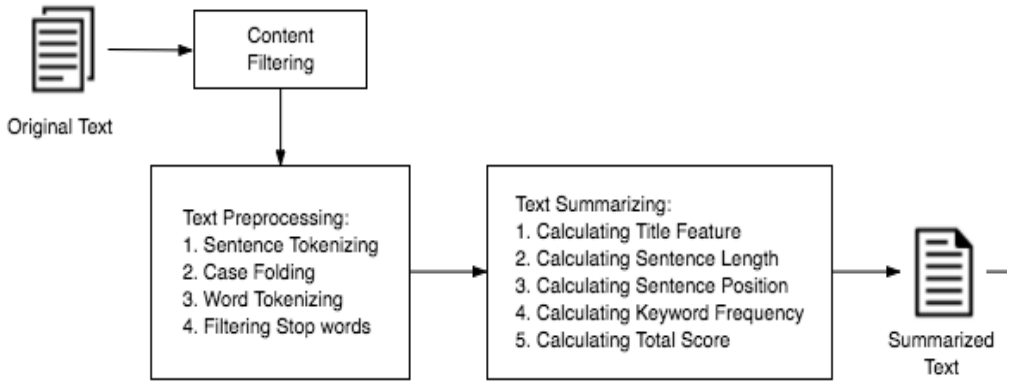
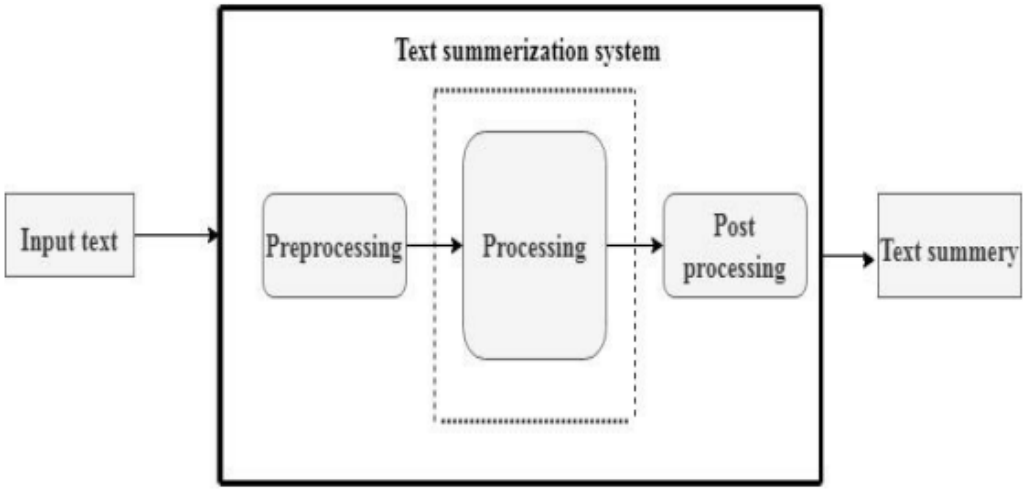


Fig 3.1

3.2.Text summarization Work flow



Architecture of project

Fig 3.2

1.Input:

Taking input

```
text=input()
```

2.Document pre-processing

Due to the Excess sources of information in today's world, the input documents we receive may not be in the correct English format, which may contain audio. Sounds include various special characters, unwanted spaces, newlines, stops, and more. Therefore, perform the following tasks on the input file to get only the useful parts of the document.

Step 1: All line breaks are removed.

Step 2: All corner brackets and special numbers are removed.

Step 3: All commas, extra spaces and repeating sentences are removed

3. Removal of stop words

In this step it will remove all subtitles from your input according to your native language. These stop words do not provide reliable information about a particular context. It does not convey any information about this emotion as it builds a collection of emotions like "is", "am", "who" to create an illustration.

```
stopwords=list(STOP_WORDS)
from string import punctuation
punctuation=punctuation+ '\n'
```

4.Tokenization

Previously, sentences were split into several words. Basically, this token model is used to do the activity in the form of a pipelined NLP natural language processing process. This is useful at two stages, word level and sentence level. The first is a standard word mark that restores a set of words in a given sentence.

```
nlp = spacy.load('en_core_web_sm')
doc= nlp(text)
tokens=[token.text for token in doc]
print(tokens)
```

5.Extraction of important sentences

We need a way to verify the value of the text in the scroll. The following calculations are performed to extract the key phrase from the text

Step 1 Calculate word frequencies from the text after removing stopword and punctuations.

```
word_frequencies={}
for word in doc:
    if word.text.lower() not in stopwords:
        if word.text.lower() not in punctuation:
            if word.text not in word_frequencies.keys():
                word_frequencies[word.text] = 1
            else:
                word_frequencies[word.text] += 1
```

Step 2: Calculate the maximum frequency and divide it by all frequencies to get normalized word frequencies.

```
max_frequency=max(word_frequencies.values())
for word in word_frequencies.keys():
    word_frequencies[word]=word_frequencies[word]/max_frequency
```

Print normalized word frequencies.

```
print(word_frequencies)
```

Step 3: Get sentence Scores.

```
sentence_tokens= [sent for sent in doc.sents]
print(sentence_tokens)
```

Step 4: Calculate the most important sentences by adding the word frequencies in each sentence.

```
sentence_scores = {}
for sent in sentence_tokens:
    for word in sent:
        if word.text.lower() in word_frequencies.keys():
            if sent not in sentence_scores.keys():

sentence_scores[sent]=word_frequencies[word.text.lower()]
            else:

sentence_scores[sent]+=word_frequencies[word.text.lower()]
```

Step 5: From '*heapq*' import '*nlargest*' and calculate %(how much percentage i.e.30% or 40%) of text with maximum score.

```
from heapq import nlargest
select_length=int(len(sentence_tokens)*0.3)
select_length
summary=nlargest(select_length,
sentence_scores,key=sentence_scores.get)
summary
```

Step 6: Get the summary of the text.

```
final_summary=[word.text for word in summary]
final_summary
summary=''.join(final_summary)
summary
```

6.Result

The text is then preprocessed, including the removal of stop words and punctuation, by finding the word Frequency and the Sentence Frequency. Finally, create a text summary.

Chapter 4

Requirements Specification

4.1 packages

pip install -U spacy

python -m spacy download en_core_web_sm

import spacy

from spacy.lang.en.stop_words import STOP_WORDS

from string import punctuation

from heapq import nlargest

Chapter 5

Conclusions and Future scope

5.1.Conclusion

Text summarization is a fascinating academic subject with numerous practical applications in industry. Summaries are helpful in a variety of downstream applications, such as news summaries, reporting, news summaries, and headlines, by condensing enormous volumes of information into brief bursts. Therefore, using spaCy in natural language processing the system identifies text of a document and gives the best possible summary.

This is just one of the ways to get text summarization by use of most frequently used words and then calculating the most important sentences.

Chapter 6

Future Scope

Future Scope

- The current model works for text data in the form of different documents.

- It can be further improvised to accepting inputs in form of audio files, Images and extracting speech text from videos for summarization These can be implemented by using various other libraries like *'nltk'* to do it by using lexical analysis, part of speech tagger, and n-grams, name entity recognition, and cosine similarities between sentences

References

[1] Source Information of NLP :https://monkeylearn.com/what-istext-classification/

[2]Source Information of SpaCy: https://realpython.com/natural-language-processing-spacy-python/

[3] G. Silva, R. Ferreira, S. J. Simske, L. Rafael Lins, M. Riss, and H. O. Cabral, "Automatic text document summarization based on machine learning," DocEng 2015 - Proc. 2015 ACM Symp. Doc. Eng., pp. 191–194, 2015, doi: 10.1145/2682571.2797099

[4] J. N. Madhuri and R. Ganesh Kumar, "Extractive Text Summarization Using Sentence Ranking," 2019 Int. Conf. Data Sci. Commun. IconDSC 2019, pp. 1–3, 2019, doi: 10.1109/IconDSC.2019.8817040.