Text Summarization

Submitted in partial fulfillment of the requirements of

NLP Project

for

Final Year of Computer Engineering

By

Ayush Gupta **19102B0033**

Ashmeet Arora **19101A0080**

Under the Guidance of

Mrs. Suja Jayachandran

Department of Computer Engineering

Vidyalankar Institute of Technology
Wadala(E), Mumbai-400437

University of Mumbai

2022-23

CERTIFICATE OF APPROVAL

This is to certify that the project entitled

Text Summarization

is a bonafide work of

Ayush Gupta	**19102B0033**
Ashmeet Arora	**19101A0080**

submitted to the University of Mumbai in partial fulfillment of

NLP Project

For Final Year of Computer Engineering

Guide Head of Department Principal(Name)

NLP Project Report
Approval

This project report entitled ***Text Summarization*** by

1. ***Ayush Gupta*** ***19102B0033***
2. ***Ashmeet Arora*** ***19101A0080***

is approved for NLP Project for Final Year of Computer Engineering.

Internal Examiner External Examiner

Date:

Place:

Declaration

I declare that this written submission represents my ideas in my own words and where others' ideas or words have been included, I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

	Name of student	Roll No.	Signature
1)	Ayush Gupta	19102B0033	
2)	Ashmeet Arora	19101A0080	

Date:

Place:

Acknowledgements

We would like to thank our mentor Mrs. Suja Jayachanran ma'am for providing useful information forthis project, also for guiding us on the right path. We would like to thank ma'am for conducting meeting whenever necessary.

I would like to thank our team members for helping us in completion of the project.
'Coming together is beginning. Keeping together is progress. Working together is Success'.

Abstract

In this project, Automatic text summarization is basically summarizing of the given paragraph using natural language processing and machine learning. There has been an explosion in the amount of text data from a variety of sources. This volume of text is an invaluable source of information and knowledge which needs to be effectively summarized to be useful. In this review, the main approaches to automatic text summarization are described. We review the different processes for summarization and describe the effectiveness and shortcomings of the different methods. The system works by assigning scores to sentences in the document to be summarized, and using the highest scoring sentences in the summary. Score values are based on features extracted from the sentence. A linear combination of feature scores is used. Almost all of the mappings from feature to score and the coefficient values in the linear combination are derived from a training corpus. Some anaphor resolution is performed. The system was submitted to the Document Understanding Conference for evaluation. In addition to basic summarization, some attempt is made to address the issue of targeting the text at the user. The intended user is considered to have little background knowledge or reading ability. The system helps by simplifying the individual words used in the summary and by drawing the pre-requisite background information from the web.

Table of Contents

Sr No	Description	Page No
1	Introduction	
2	Potential Applications	
3	Objective	
4	Scope	
5	Working of project	
6	Conclusion	
7	References	

Introduction

In the modern Internet age, textual data is ever increasing. Need some way to condense this data while preserving the information and meaning. We need to summarize textual data for that. Text summarization is the process of automatically generating natural language summaries from an input document while retaining the important points. It would help in easy and fast retrieval of information. There are two prominent types of summarization algorithms.

- Extractive summarization systems form summaries by copying parts of the source text through some measure of importance and then combine those part/sentences together to render a summary. Importance of sentence is based on linguistic and statistical features.

- Abstractive summarization systems generate new phrases, possibly rephrasing or using words that were not in the original text. Naturally abstractive approaches are harder. For perfect abstractive summary, the model has to first truly understand the document and then try to express that understanding in short possibly using new words and phrases. Much harder than extractive. Has complex capabilities like generalization, paraphrasing and incorporating realworld knowledge. Majority of the work has traditionally focused on extractive approaches due to the easy of defining hard-coded rules to select important sentences than generate new ones. Also, it promises grammatically correct and coherent summary. But they often don't summarize long and complex texts well as they are very restrictive.

Potential Applications

Potential applications Possible current uses of summarization:

1. People need to learn much from texts. But they tend to want to spend less time while doing this.

2. It aims to solve this problem by supplying them the summaries of the text from which they want to gain information.

3. Goals of this project are that these summaries will be as important as possible in the aspect of the texts' intention.

4. The user will be eligible to select the summary length.

5. Supplying the user, a smooth and clear interface.

6. Configuring a fast replying server system.

Objectives

The objective of the project is to understand the concepts of natural language processing and creating a tool for text summarization. The concern in automatic summarization is increasing broadly so the manual work is removed. The project concentrates creating a tool which automatically summarizes the document.

Scope

The project is wide in scope | all of the limitations stated below may seem to contradict that, but they are the only restrictions applied. This project looks at single document summarization - the area of multi document summarization is not covered. Also, the summaries produced are largely extracts of the document being summarized, rather than newly generated abstracts. The parameters used are optimal for news articles, although that can be changed easily.

Code:

Import Necessary Libraries

```
import bs4 as bs
import urllib.request
import re
import nltk
```

Scrap Data From Wikipedia

```
scraped_data = urllib.request.urlopen('https://en.wikipedia.org/wiki/Data_science')
article = scraped_data.read()
parsed_article = bs.BeautifulSoup(article, 'lxml')
paragraphs = parsed_article.find_all('p')
article_text = ""
for p in paragraphs:
    article_text += p.text
```

Preprocessing

```
article_text = re.sub(r'\[[0-9]*\]', ' ', article_text)
article_text = re.sub(r'\s+', ' ', article_text)
formatted_article = re.sub('[^a-zA-Z]', ' ', article_text)
formatted_article = re.sub(r'\s+', ' ', formatted_article)
tokenize_sentence = nltk.sent_tokenize(article_text)
```

Frequency of Each Word

```
stopwords = nltk.corpus.stopwords.words('english')
word_frequency = {}
for word in nltk.word_tokenize(formatted_article):
    if word not in stopwords:
        if word not in word_frequency.keys():
            word_frequency[word] = 1
        else:
            word_frequency[word] += 1
maximum_frequncy = max(word_frequency.values())
for word in word_frequency.keys():
    word_frequency[word] = (word_frequency[word] / maximum_frequncy)
```

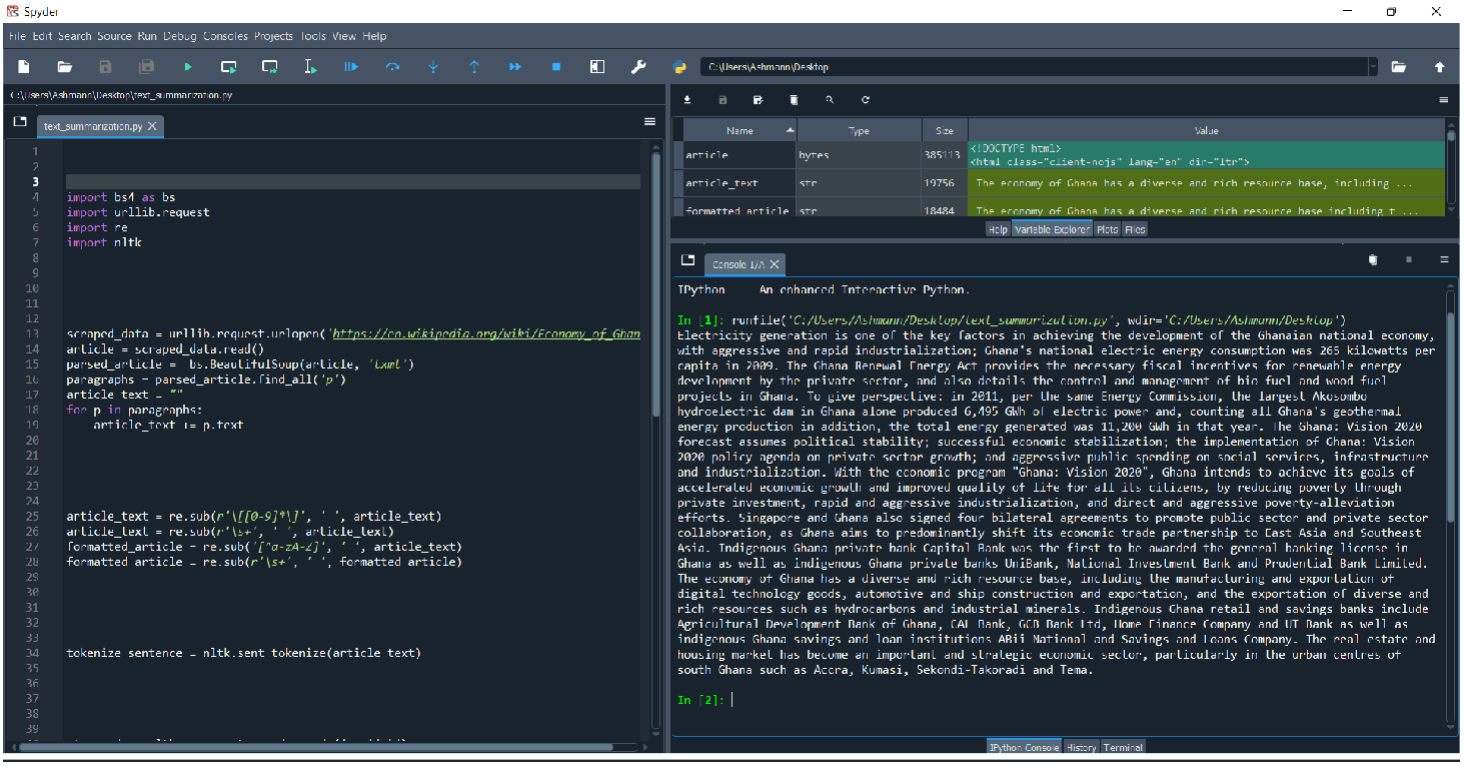
Sentence Score

```
sentence_score = {}
for sent in tokenize_sentence:
    for word in nltk.word_tokenize(sent.lower()):
        if word in word_frequency.keys():
            if len(sent.split(' ')) < 50:
                if sent not in sentence_score.keys():
                    sentence_score[sent] = word_frequency[word]
                else:
                    sentence_score[sent] += word_frequency[word]
```

Summary of Whole Text

```
import heapq
sentence_summary = heapq.nlargest(10, sentence_score, key = sentence_score.get)
summary = ' '.join(sentence_summary)
print(summary)
```

Output:


Spyder
File Edit Search Source Run Debug Consoles Projects Tools View Help
text_summarization.py
import bs4 as bs
import urllib.request
import re
import nltk
scraped_data = urllib.request.urlopen('https://en.wikipedia.org/wiki/Economy_of_Ghan
article = scraped_data.read()
parsed_article = bs.BeautifulSoup(article, 'lxml')
paragraphs = parsed_article.find_all('p')
article_text = ""
for p in paragraphs:
article_text += p.text
article_text = re.sub(r'\[[0-9]*\]', ' ', article_text)
article_text = re.sub(r'\s+', ' ', article_text)
formatted_article = re.sub('[^a-zA-Z]', ' ', article_text)
formatted_article = re.sub(r'\s+', ' ', formatted_article)
tokenize_sentence = nltk.sent_tokenize(article_text)
Name
Type
Size
Value
article
bytes
385113
<!DOCTYPE html>
<html class="client-nojs" lang="en" dir="ltr">
article_text
str
19756
The economy of Ghana has a diverse and rich resource base, including ...
formatted_article
str
18484
The economy of Ghana has a diverse and rich resource base including t ...
Help
Variable Explorer
Plots
Files
Console 1/A
IPython
An enhanced Interactive Python.
In [1]: runfile('C:/Users/Ashmann/Desktop/text_summarization.py', wdir='C:/Users/Ashmann/Desktop')
Electricity generation is one of the key factors in achieving the development of the Ghanaian national economy, with aggressive and rapid industrialization; Ghana's national electric energy consumption was 265 kilowatts per capita in 2009. The Ghana Renewal Energy Act provides the necessary fiscal incentives for renewable energy development by the private sector, and also details the control and management of bio fuel and wood fuel projects in Ghana. To give perspective: in 2011, per the same Energy Commission, the largest Akosombo hydroelectric dam in Ghana alone produced 6,495 GWh of electric power and, counting all Ghana's geothermal energy production in addition, the total energy generated was 11,200 GWh in that year. The Ghana: Vision 2020 forecast assumes political stability; successful economic stabilization; the implementation of Ghana: Vision 2020 policy agenda on private sector growth; and aggressive public spending on social services, infrastructure and industrialization. With the economic program "Ghana: Vision 2020", Ghana intends to achieve its goals of accelerated economic growth and improved quality of life for all its citizens, by reducing poverty through private investment, rapid and aggressive industrialization, and direct and aggressive poverty-alleviation efforts. Singapore and Ghana also signed four bilateral agreements to promote public sector and private sector collaboration, as Ghana aims to predominantly shift its economic trade partnership to East Asia and Southeast Asia. Indigenous Ghana private bank Capital Bank was the first to be awarded the general banking license in Ghana as well as indigenous Ghana private banks UniBank, National Investment Bank and Prudential Bank Limited. The economy of Ghana has a diverse and rich resource base, including the manufacturing and exportation of digital technology goods, automotive and ship construction and exportation, and the exportation of diverse and rich resources such as hydrocarbons and industrial minerals. Indigenous Ghana retail and savings banks include Agricultural Development Bank of Ghana, CAL Bank, GCB Bank Ltd, Home Finance Company and UT Bank as well as indigenous Ghana savings and loan institutions ABii National and Savings and Loans Company. The real estate and housing market has become an important and strategic economic sector, particularly in the urban centres of south Ghana such as Accra, Kumasi, Sekondi-Takoradi and Tema.
In [2]:
IPython Console
History
Terminal

Original Text:

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

Summarized Text:

Electricity generation is one of the key factors in achieving the development of the Ghanaian national economy, with aggressive and rapid industrialization; Ghana's national electric energy consumption was 265 kilowatts per capita in 2009. The Ghana Renewal Energy Act provides the necessary fiscal incentives for renewable energy development by the private sector, and also details the control and management of bio-fuel and wood fuel projects in Ghana. To give perspective: in 2011, per the same Energy Commission, the largest Akosombo hydroelectric dam in Ghana alone produced 6,495 GWh of electric power and, counting all Ghana's geothermal energy production in addition, the total energy generated was 11,200 GWh in that year. The Ghana: Vision 2020 forecast assumes political stability; successful economic stabilization; the implementation of Ghana: Vision 2020 policy agenda on private sector growth; and aggressive public spending on social services, infrastructure and industrialization. With the economic program "Ghana: Vision 2020", Ghana intends to achieve its goals of accelerated economic growth and improved quality of life for all its citizens, by reducing poverty through private investment, rapid and aggressive industrialization, and direct and aggressive poverty-alleviation efforts. Singapore and Ghana also signed four bilateral agreements to promote public sector and private sector collaboration, as Ghana aims to predominantly shift its economic trade partnership to East Asia and Southeast Asia. Indigenous Ghana private bank Capital Bank was the first to be awarded the general banking license in Ghana as well as indigenous Ghana private banks UniBank, National Investment Bank and Prudential Bank Limited. The economy of Ghana has a diverse and rich resource base, including the manufacturing and exportation of digital technology goods, automotive and ship construction and exportation, and the exportation of diverse and rich resources such as hydrocarbons and industrial minerals. Indigenous Ghana retail and savings banks include Agricultural Development Bank of Ghana, CAL Bank, GCB Bank Ltd, Home Finance Company and UT Bank as well as indigenous Ghana savings and loan institutions ABii National and Savings and Loans Company. The real estate and housing market has become an important and strategic economic sector, particularly in the urban centres of south Ghana such as Accra, Kumasi, Sekondi-Takoradi and Tema.

Conclusion

Text summarization is one of the major problems in the field of Natural Language Processing. Methods such as Deep Understanding, Sentence Extraction, Paragraph Extraction, Machine Learning, and even some which employ all these methods along with Traditional NLP Techniques(Semantic Analysis, etc.). As such, keeping these accomplishments in mind, there is still ample amount of research left in the domain of Text Summarization, as a meaningful summary is still difficult to attain in all domains and languages. As with time internet is growing at a very fast rate and with it data and information is also increasing. it will going to be difficult for human to summarize large amount of data. Thus there is a need of automatic text summarization because of this huge amount of data. Until now, we have read multiple papers regarding text summarization, natural language processing. There are multiple automatic text summarizers with great capabilities and giving good results. We have learned all the basics of Extractive and Abstractive Method of automatic text summarization and tried to implement extractive one. We have made a basic automatic text summarizer using nltk library using python and it is working on small documents. We have used extractive approach to do text summarization.

References

[1] Jiwei Tan, XiaojunWan,Jianguo Xiao Institute of Computer Science and Technology,Peking University "Abstractive document summarization with a GraphBased attentional neural model. "

[2] SeonggiRyang, Graduate school of Information science and technology, University of Tokyo Takeshi Abekawa, National institute of informatics "Framework of automatic text summarization using Reinforcement learning" 48 Vol 11, Issue 4 , April/2020 ISSN NO:0377-9254 www.jespublication.com PageNo:14

[3] Tianshi, YaserKeneshloo, Narenramakrishnan, Chandan K. Reddy, Senior member, IEEE " Neural Abstractive text summarization with sequence-to - sequence models"

[4] Josef Steinberger, KarelJežek, "Using latent Semantic analysis In Text Summarization and Summary Evaluation", Department of Computer Science and Engineering, UniverzitníCZ-306 14 Plzeň.