Text Summarizer using NLP

Submitted in partial fulfillment of the requirements of the degree

BACHELOR OF ENGINEERING IN **COMPUTER ENGINEERING**

By

Pokharkar Akshay Hanumant Chandrabhanga	**VU1F1920053**
Dhumal Pratik Pramod Smita	**VU1F1920058**
Singh Aman Shashidhar Sangeeta	**VU1F1920064**
Hadawale Hrithik Kashinath Archana	**VU1F1920059**

Name of the Mentor

Prof. Sumit Shinde

Department of Computer Engineering

VasantDada Patil Pratishthan's

College of Engineering

Vasantdada Patil Education Complex, Eastern
Express Highway Near Everard Nagar,
Chunabhatti, Sion, Mumbai, Maharashtra
400022

University of Mumbai

(AY 2021-22)

Contents

Abstract

Acknowledgments

Introduction

1.1 Introduction

1.2 Motivation

1.3 Problem Statement & Objectives

1.4 Organization of the Report

2 Literature Survey

2.1 Survey of Existing System/**SRS**

2.2 Limitation Existing system or Research gap

2.3 Mini Project Contribution

3 Proposed System (eg New Approach of Data Summarization)

3.1 Introduction

3.2 Architecture/ Framework

3.3 Algorithm and Process Design

3.4 Details of Hardware & Software

3.4 Experiment and Results for Validation and Verification

3.5 Analysis

3.6 Conclusion and Future work.

References

Abstract

Text Summarization is condensing the source text into a shorter version preserving its information content and overall meaning. It is very difficult for human beings to manually summarize large documents of text.

Text Summarization methods can be classified into extractive and abstractive summarization. An extractive summarization method consists of selecting important sentences, paragraphs etc. from the original document and concatenating them into shorter form. The importance of sentences is decided based on statistical and linguistic features of sentences. An abstractive summarization method consists of understanding the original text and re-telling it in fewer words. It uses linguistic methods to examine and interpret the text and then to find the new concepts and expressions to best describe it by generating a new shorter text that conveys the most important information from the original text document

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.

To perform abstractive summarization an encoder-decoder neural network with an attention model (though this is in primitive stages currently and needs an immense amount of computing power) is used.

In this project we have built extractive summarization tool from scratch using TF-IDF vectorization and pairwise cosine similarity

Acknowledgements

I extend my sincere thanks to Vasantdada Patil college of Engineering which provided me with the opportunity to fulfill our wish and achieve our goal.

I would like to express deep debt to Prof. Sumit Shinde, project guide for her vital suggestion, meticulous guidance and constant motivation, makes this project successful uptill this point. They spent a lot of time with us and gave all the related information and expertise about report writing

1.1 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.

Stages of Text Summarization:-

(i)Content Selection:- Choose Sentences to extract from large Chunks Text.

(ii) Information Ordering:- Choose an order to place Summary.

(iii) Sentence Realization:- Clean up the sentences

We have used TF-IDF vectorization and pairwise cosine similarity. TF-IDF (Term Frequency — Inverse Document Frequency) gives weights to individual words based on their uniqueness compared to the document's overall vocabulary. Words with higher weights (more unique) often have more importance or provide more meaning to the document. To use this,we built a function that takes in an article's text, tokenizes each sentence (dataframe rows), creates a vocabulary without stop words for the individual document (dataframe columns) and finally gives TF-IDF weights to each individual word in the vocab for each sentence. Using the TF-IDF weights for each sentence, we convert each row into a vector and store them in a matrix. Next, we find the cosine-similarity of each TF-IDF vectorized sentence pair.

Finally, after finding the cosine-similarity for all vectorized pairs,we average the weights of each vector, and return the indexes of the vectors with the highest averages. These indexes are then used to pull out the sentences from the original text for the summarization. The sentences with the highest average weights will capture the unique and important sentences from the original text (although like everything, it's not always perfect).

1.2 Motivation

Creating a summary from a given piece of content is a very abstract process that everyone participates in. Automating such a process can help parse through a lot of data and help humans better use their time to make crucial decisions. With the sheer volume of media out there, one can be very efficient by reducing the fluff around the most critical information.

Due to the increasingly busy lifestyle of the people,they are not able to read their favourite articles,documents,reviews,books,web articles,blogs,etc because doing these process is time consuming.

Our motivation is to create a user friendly text summarization website to summarize web articles,passages and PDFs,which is free to use.

1.3 Problem Statement & Objectives

Problem Statement

To create a text summarizer which summarises the text or the content of the paragraph in minimum words without changing its meaning.This system is made using NLP based model which is branch of machine learning. This text summarizer also summarizes text from the weblinks and also summarises text from PDF document.

Objectives

- Summaries reduce reading time.

- When researching documents, summaries make the selection process easier.

- Automatic summarization improves the effectiveness of indexing.

- Automatic summarization algorithms are less biased than human summarizers.

- Personalized summaries are useful in question-answering systems as they provide personalized information.

- Using automatic or semi-automatic summarization systems enables commercial abstract services to - increase the number of text documents they are able to process.

1.4 Organization of Report

Part 1: This part declares the Introduction part about the project and executive summary about the Website used in the project. The organization of project is mentioned as well. Finally the project objectives.

Part 2: Text Summarization System using NLP Overview is mentioned as a full summary about the used technique.

Part 3: This part describes the full analysis for the proposed system. The Use case and The Class Diagram are giving in detail.

Part 4: System Implementation which gives the full documented of the code ~ use to implement the proposed system.

Part 5: A full summary about the project

2. Literature Survey

2.1 Survey of Existing System

We have investigated the existing surveys of the ATS domain,and a few of them are presented to prove the significanceof this paper. Most surveys covered the former methodsand research on ATS. However, recent trends, applicability,effects, limitations, and challenges of ATS techniques werenot present. Table 1 summarizes and compares the existingsurvey on ATS.Mishra reviewed (2000-2013) years of studiesand found some methods such as hybrid statistical and MLapproaches. The researchers did not include cognitive aspectsor evaluations of the impact of ATS. Allahyari investigated different processes such as topic representation,frequency-driven, graph-based, and machine learning meth-ods for ATS. This research only includes the frequently usedstrategies. El-Kassas described graph-based, fuzzylogic-based, concept-oriented, ML approaches, etc., with heir advantages or disadvantages. This research did not in-clude abstractive or hybrid techniques. Saranyamol offered a thorough survey for analysts by introducing variousaspects of ATS such as structure, strategies, datasets, evalua-tion metrics, etc. Gambhir attempted to analyze ahybrid approach including two text summarization methods.This study missed many contemporary techniques for review.The research of Gholamrezazadeh represents acomprehensive and comparative study of extractive methodsin ATS of the last decade. Several multilingual approacheshave also been discussed. Andhale provided ataxonomy of text summarization methods and a variety oftechniques. Although the author has covered some time-consuming processes of ATS, recent, more efficient methodssuch as machine learning were missed. Abualigah conducted research on how to handle multiple documents andmassive web data for text summarization. Lastly, the papercontains a comparative table with recent studies withoutdetails. Bharti presented a survey of researchpapers based on automated keyword extraction methods andtechniques. It covers ideas about multiple databases that areused for document summarization

Paper No.	Publication year and publisher	Factors analysed	Scope of Improvement
1	2018, arXiv (Cornell University)	They Presented a fully data driven approach for automatic text summarization.They proposed and evaluated the model on standard datasets which show results comparable to the state of the art models without access to any linguistic information	they have assumed that summary length to be generated should be less than ‘page_len’.
2	2017,arXiv (Cornell University)	They first pre-train the generative model by generating summaries given the source text. Then they pre-train the discriminator by providing positive examples from the human-generated summaries and the negative examples produced from the pre-trained generator. After the pre-training, the generator and discriminator are trained alternatively.	The evaluation metric is different from the training loss. The input of the decoder in each time step is often from the true summary during the training. In the testing phase, the input of the next time step is the previous word.
3	2019, arXiv (Cornell University)	They apply thier model to CNN/Daily Mail dataset (Hermann et al., 2015; Nallapati et al., 2016), which contains news articles (39 sentences on average) paired with multi-sentence summaries, and show that they outperform the state of-the-art abstractive system by at least 2 ROUGE points.	It is unable to attain higher level of abstraction.
4	2020,JMLR (Journal of Machine Learning Research)	It reduces each document in the corpus to a vector of real numbers, each of which represents ratios of counts. After suitable normalization, this term frequency count is compared to an inverse document frequency count, which measures the number of occurrences of a word in the entire corpus .The end result is a term-by-document matrix X whose columns contain the tf-idf values for each of the documents in the corpus. The tf-idf scheme reduces documents of arbitrary length to fixed length lists of numbers.	They could also consider partially exchangeable models in which they condition on exogenous variables; thus, for example, the topic distribution could be conditioned on features such as “paragraph” or “sentence,” providing a more powerful text model that makes use of information obtained from a parser.

2.2 Limitation of Existing Survey

- There are some of the sites or interfaces that are available in market made for stock prediction like paraphraser.io , quillbot, Cruxe.in, prepostseo.com etc.
- But they lack some of these features as per survey:

1.**Data privacy issues**:Data privacy, sometimes also referred to as information privacy, is an area of data protection that concerns the proper handling of sensitive data including, notably, personal data[1] but also other confidential data, such as certain financial data and intellectual property data, to meet regulatory requirements as well as protecting the confidentiality and immutability of the data.

2.**Not free but with subscription:**There are many text summarizer available on the internet,but they require monthly and yearly subscription.

3.**Not so user friendly interface:**Some websites have dull interface and are difficult to navigate.

2.3 Mini Project Contribution

Name	Id	Contribution
Pokharkar Akshay Hanumant Chandrabhaga	VU1F1920053	Planning of model, Documentation, Resource handling
Dhumal Pratik Pramod Smita	VU1F1920058	Frontend Development and tokenization function
Hadawale Hrithik Kashinath Archana	VU1F1920059	NLP Summarizer model and cosine similarity
Singh Aman Shashidhar Sangeeta	VU1F1920064	Web Scrapping and pdf text extraction

3. Proposed system

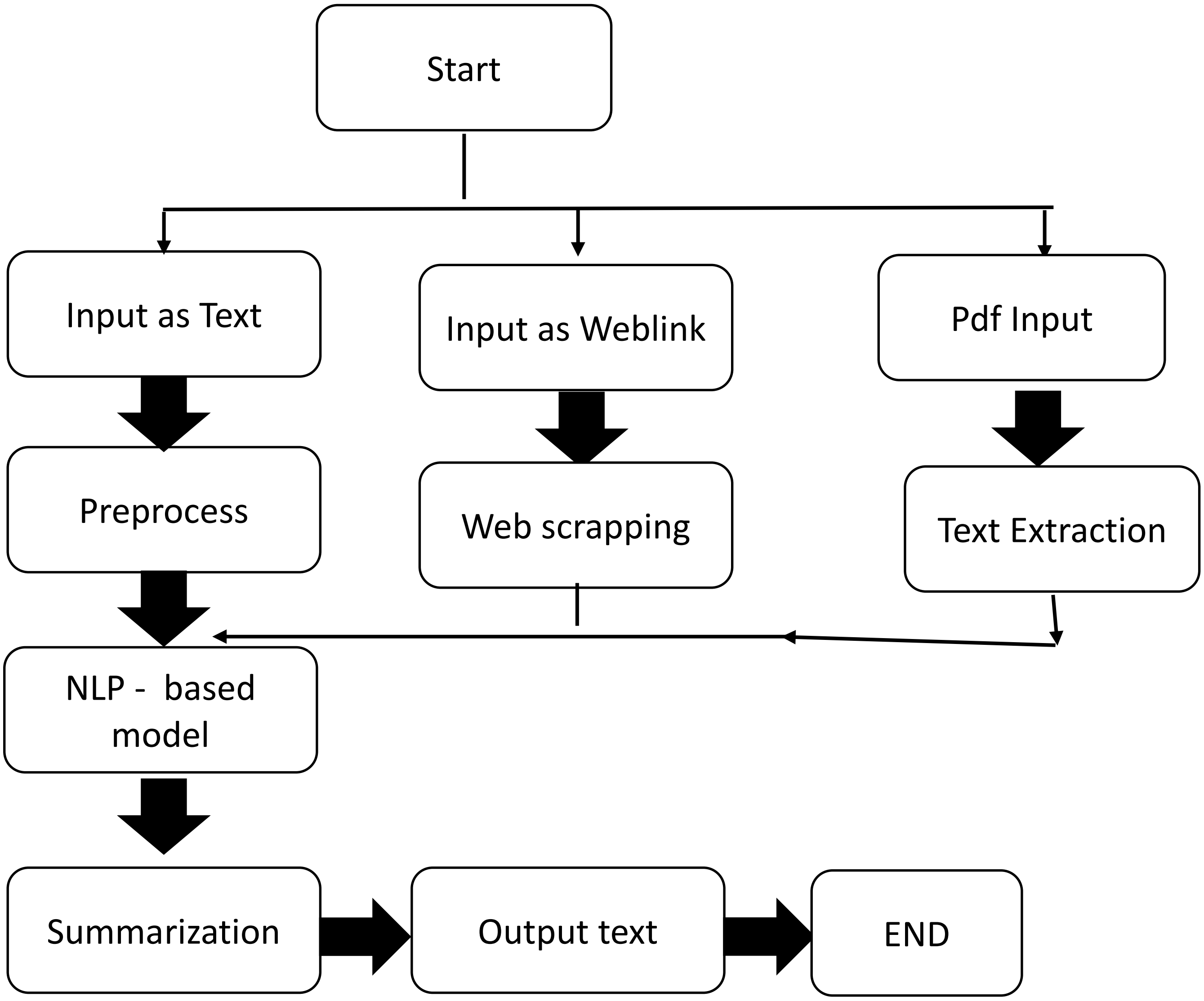
3.1 Introduction

Text summarization takes care of choosing the most significant portions of text and generates coherent summaries that express the main intent of the given document. The services offered by our text summarizer is , summarizing the text from input, summarizing web articles or weblinks and summarizing from the PDF. Our system does not ask for user details. It provides a platform to get summary without creating an account.

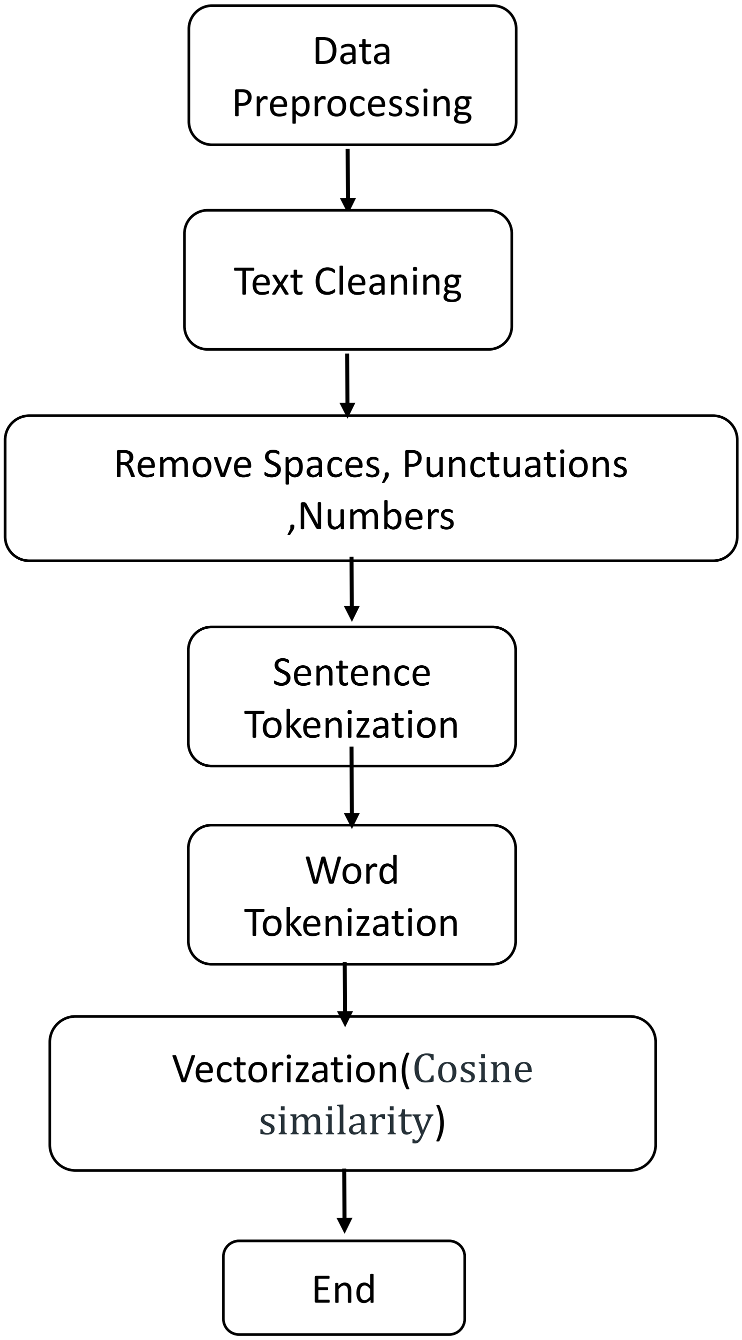
Proposed system involves the different module to generate the summary for given multiple documents. Previous system has some drawback such that it can take only text file as input. If we give other files such as PDF or word file as input then it cannot accept that file and shows the message only text files are allowed. To overcome these problems we proposed a new system that takes the input as text, PDF and weblinks. The system involves the following basic three phases.

This model uses NLP based algorithms for solving problem. It does webscrapping for getting text from the web links. The text from the pdf is extracted using python Py2pdf library. Summarizer file is made which creates different functions for different purpose. The system model is defined below.

3.2. Architecture / Framework



Model Working:



3.3 Algorithm and Process Design

NLP:

Natural language processing (NLP) is the ability of a computer program to understand human language as it is spoken and written -- referred to as natural language. It is a component of artificial intelligence

TF-IDF (Term Frequency — Inverse Document Frequency)

TF-IDF stands for Term Frequency Inverse Document Frequency of records. It can be defined as the calculation of how relevant a word in a series or corpus is to a text. The meaning increases proportionally to the number of times in the text a word appears but is compensated by the word frequency in the corpus (data-set).

TF-IDF is used with cosine similarity. TF-IDF (Term Frequency — Inverse Document Frequency) gives weights to individual words based on their uniqueness compared to the document's overall vocabulary. Words with higher weights (more unique) often have more importance or provide more meaning to the document.

Tokenization

Tokenization is the process of tokenizing or splitting a string, text into a list of tokens. One can think of token as parts like a word is a token in a sentence, and a sentence is a token in a paragraph

To use this, we built a function that takes in an article's text, tokenizes each sentence (dataframe rows), creates a vocabulary without stop words for the individual document (dataframe columns) and finally gives TF-IDF weights to each individual word in the vocab for each sentence.

COSINE Similarity

Cosine similarity is a measure of similarity between two non-zero vectors of an inner product space that measures the cosine of the angle between them.

Similarity = (A.B) / (||a||.||B||) where A and B are vectors.

Specifically, it measures the similarity in the direction or orientation of the vectors ignoring differences in their magnitude or scale. Both vectors need to be part of the same inner product space, meaning they must produce a scalar through inner product multiplication. The similarity of two vectors is measured by the cosine of the angle between them.

Using the TF-IDF weights for each sentence, I convert each row into a vector and store them in a matrix. Next, I find the cosine-similarity of each TF-IDF vectorized sentence pair.

Process Flow:

Finally, after finding the cosine-similarity for all vectorized pairs, I average the weights of each vector, and return the indexes of the vectors with the highest averages. These indexes are then used to pull out the sentences from the original text for the summarization. The sentences with the highest average weights will capture the unique and important sentences from the original text (although like everything, it’s not always perfect).

3.4 DETAILS OF HARDWARE AND SOFTWARE

Operating environment:

Software requirements	
Programming Language:	python
Operating system:	Windows 7, Windows 8 and higher versions, Linux, MacOs
Interpreter:	Web application

Hardware Requirements
Laptop / PC with 4 GB RAM
Processor – i3 or higher version, AMD 3 or higher version
Storage – 5 GB max
Internet Connection
1 GHz speed

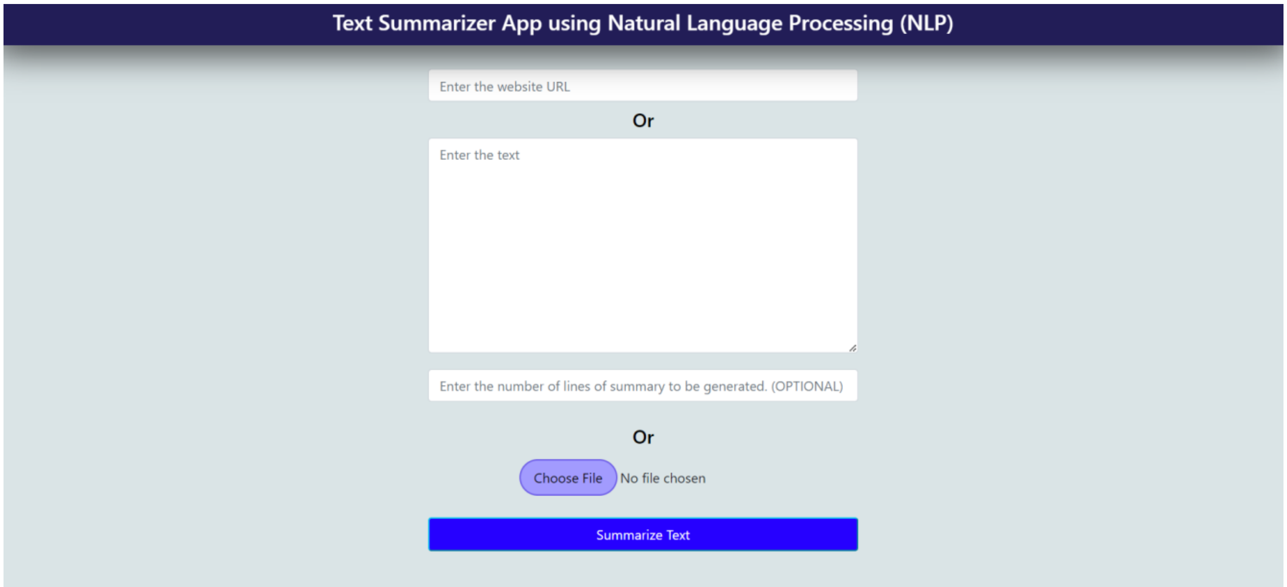
Hardware interfaces: The solution involves extensive use of several hardware devices. These devices include;

- Internet modem
- LAN
- Windows/Linux/MacOs

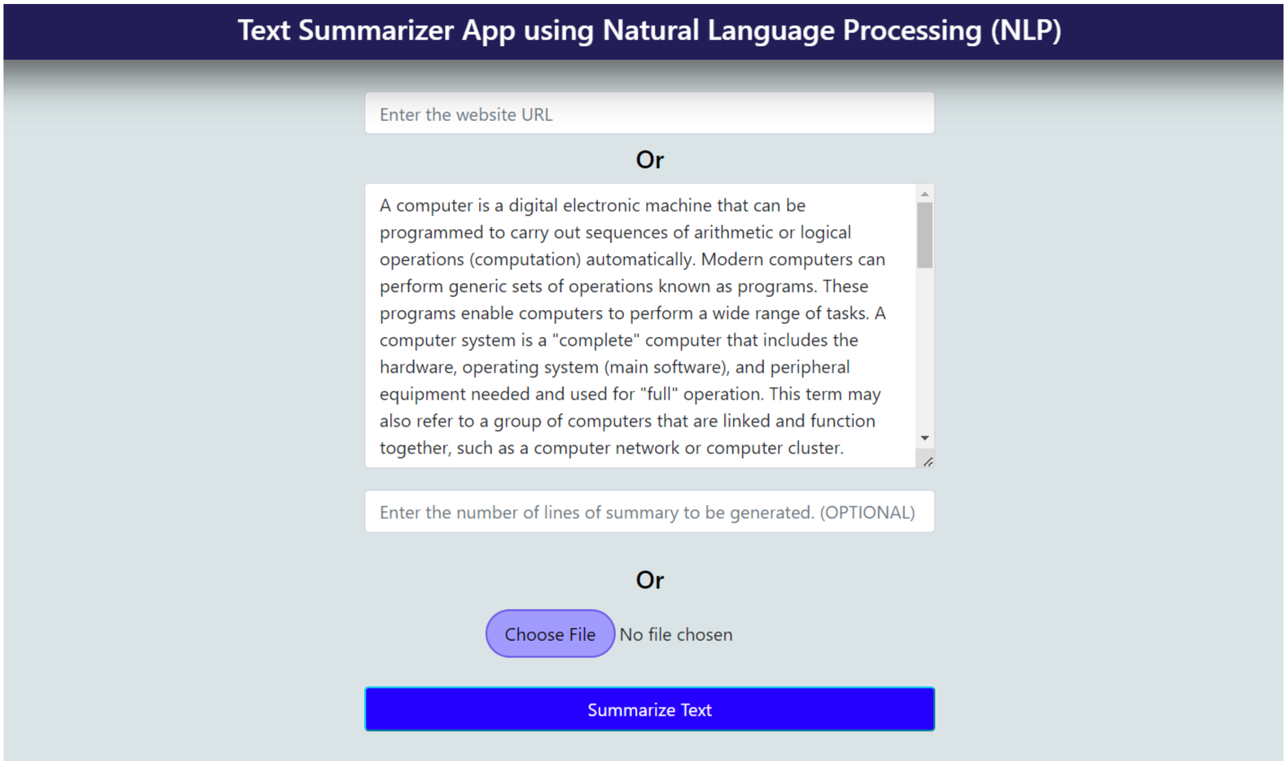
Software interfaces: We are using Django framework for creating this web application. We are using python programming language for backend programming. Python uses NLTK library for natural language processing to solve the problem

3.5 Experiment and Results for Validation and Verification

GUI of project:


Text Summarizer App using Natural Language Processing (NLP)
Enter the website URL
Or
Enter the text
Enter the number of lines of summary to be generated. (OPTIONAL)
Or
Choose File
No file chosen
Summarize Text

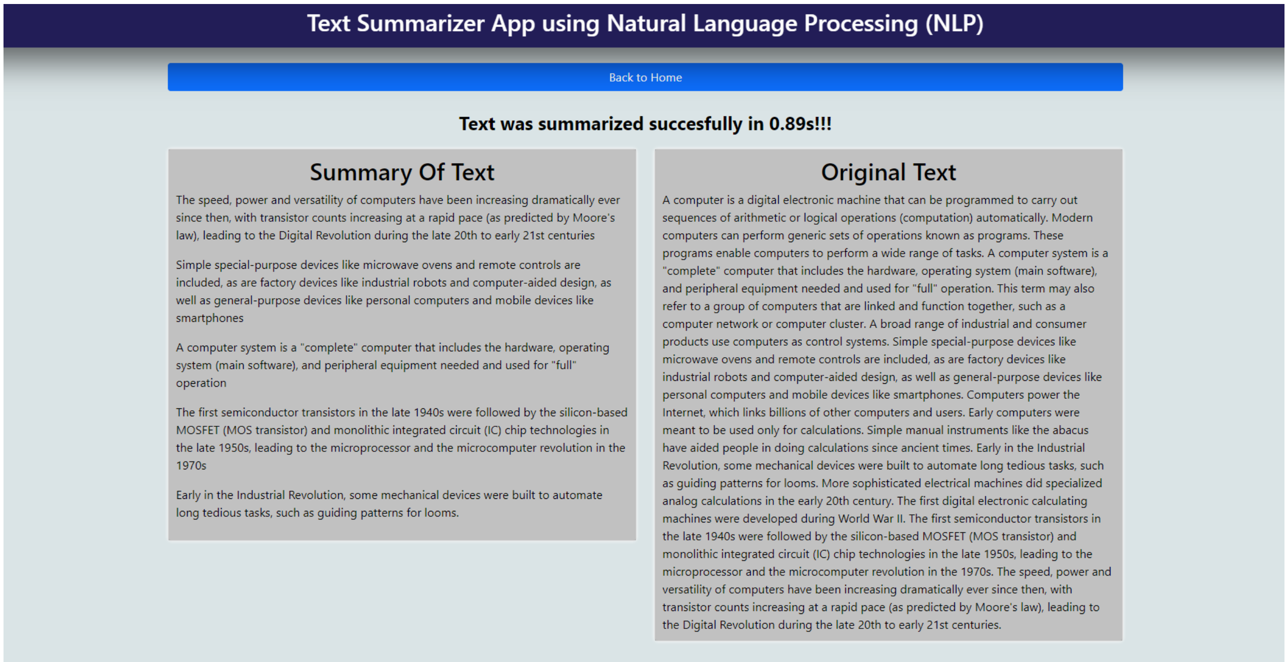
Input as Text:



A computer is a digital electronic machine that can be programmed to carry out sequences of arithmetic or logical operations (computation) automatically. Modern computers can perform generic sets of operations known as programs. These programs enable computers to perform a wide range of tasks. A computer system is a "complete" computer that includes the hardware, operating system (main software), and peripheral equipment needed and used for "full" operation. This term may also refer to a group of computers that are linked and function together, such as a computer network or computer cluster.

Enter the number of lines of summary to be generated. (OPTIONAL)
Or
Choose File
No file chosen
Summarize Text

Summarized Output :



Text Summarizer App using Natural Language Processing (NLP)

Text was summarized succesfully in 0.89s!!!

Summary Of Text

The speed, power and versatility of computers have been increasing dramatically ever since then, with transistor counts increasing at a rapid pace (as predicted by Moore's law), leading to the Digital Revolution during the late 20th to early 21st centuries

Simple special-purpose devices like microwave ovens and remote controls are included, as are factory devices like industrial robots and computer-aided design, as well as general-purpose devices like personal computers and mobile devices like smartphones

A computer system is a "complete" computer that includes the hardware, operating system (main software), and peripheral equipment needed and used for "full" operation

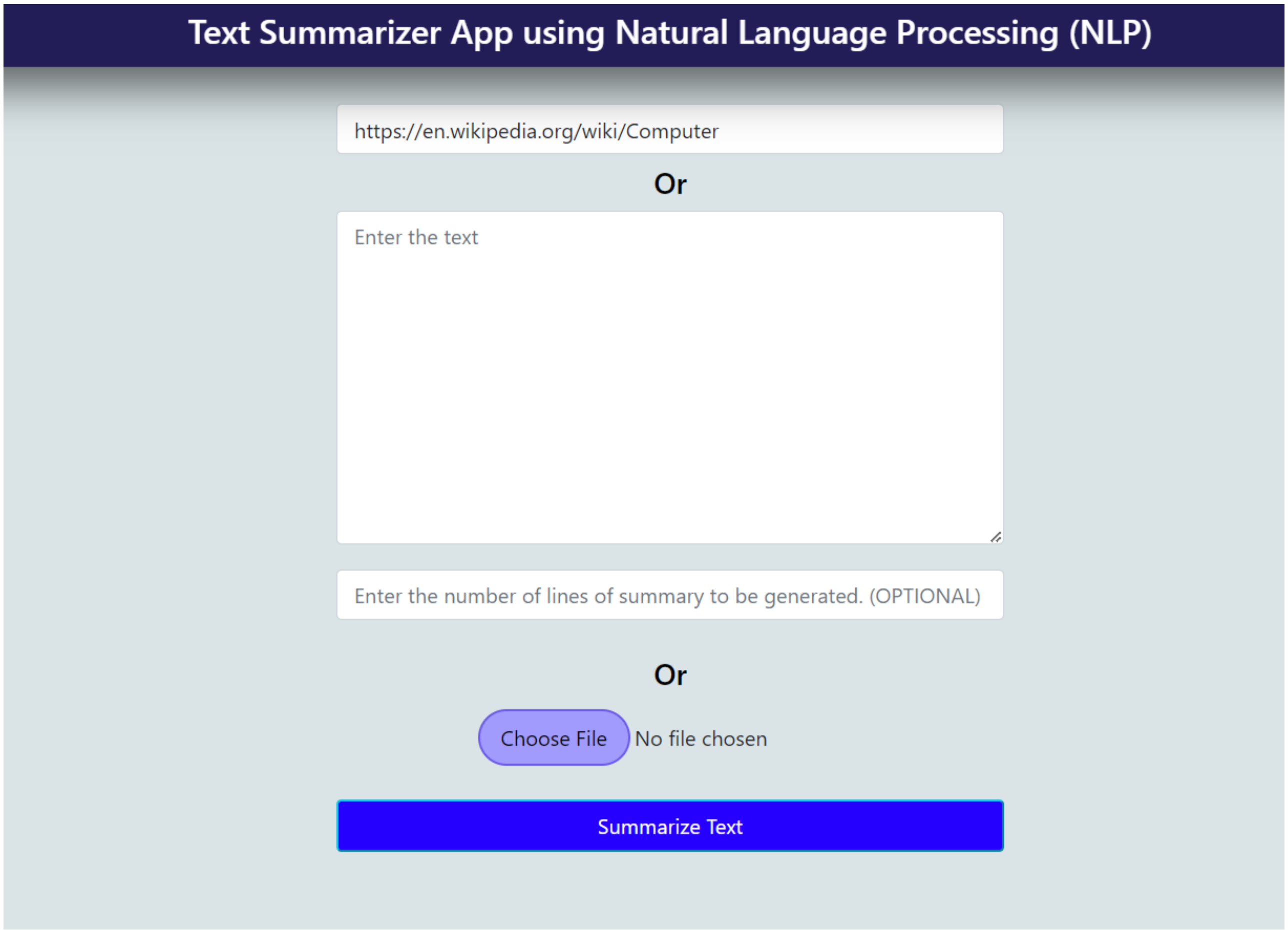
The first semiconductor transistors in the late 1940s were followed by the silicon-based MOSFET (MOS transistor) and monolithic integrated circuit (IC) chip technologies in the late 1950s, leading to the microprocessor and the microcomputer revolution in the 1970s

Early in the Industrial Revolution, some mechanical devices were built to automate long tedious tasks, such as guiding patterns for looms.

Original Text

A computer is a digital electronic machine that can be programmed to carry out sequences of arithmetic or logical operations (computation) automatically. Modern computers can perform generic sets of operations known as programs. These programs enable computers to perform a wide range of tasks. A computer system is a "complete" computer that includes the hardware, operating system (main software), and peripheral equipment needed and used for "full" operation. This term may also refer to a group of computers that are linked and function together, such as a computer network or computer cluster. A broad range of industrial and consumer products use computers as control systems. Simple special-purpose devices like microwave ovens and remote controls are included, as are factory devices like industrial robots and computer-aided design, as well as general-purpose devices like personal computers and mobile devices like smartphones. Computers power the Internet, which links billions of other computers and users. Early computers were meant to be used only for calculations. Simple manual instruments like the abacus have aided people in doing calculations since ancient times. Early in the Industrial Revolution, some mechanical devices were built to automate long tedious tasks, such as guiding patterns for looms. More sophisticated electrical machines did specialized analog calculations in the early 20th century. The first digital electronic calculating machines were developed during World War II. The first semiconductor transistors in the late 1940s were followed by the silicon-based MOSFET (MOS transistor) and monolithic integrated circuit (IC) chip technologies in the late 1950s, leading to the microprocessor and the microcomputer revolution in the 1970s. The speed, power and versatility of computers have been increasing dramatically ever since then, with transistor counts increasing at a rapid pace (as predicted by Moore's law), leading to the Digital Revolution during the late 20th to early 21st centuries.

Input as weblink:



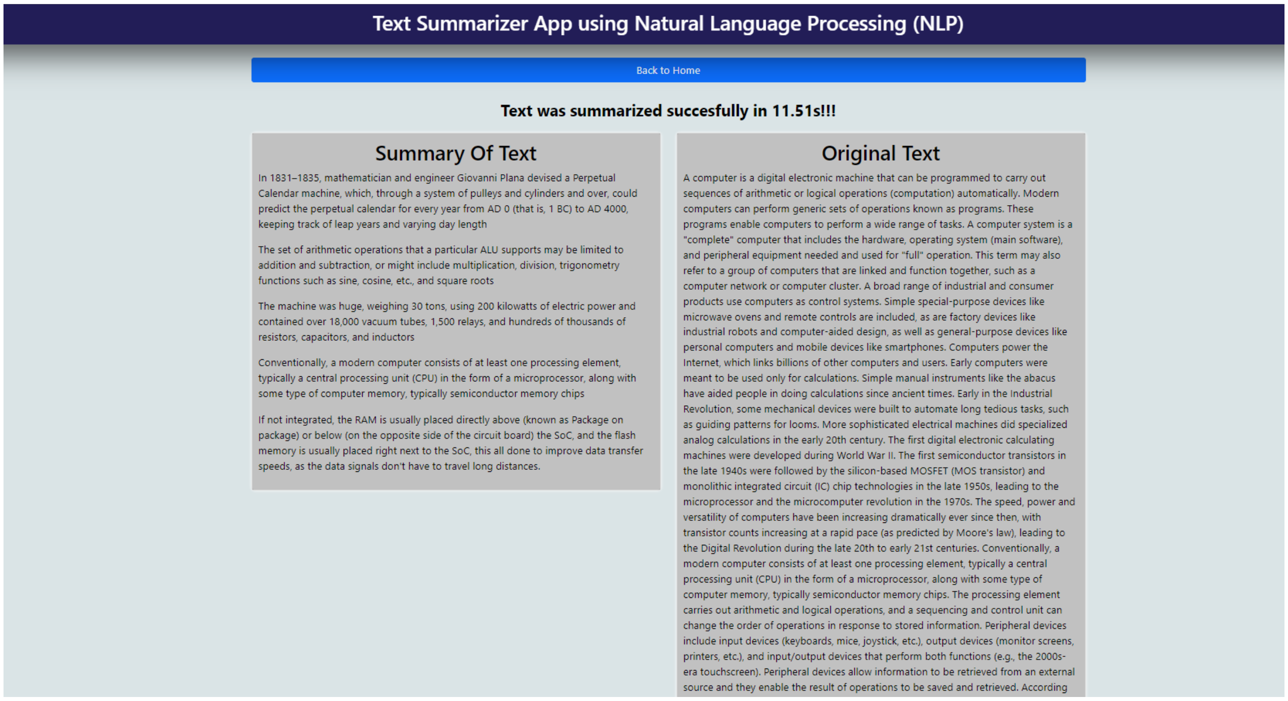
Text Summarizer App using Natural Language Processing (NLP)

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

Or

Enter the text
Enter the number of lines of summary to be generated. (OPTIONAL)
Or
Choose File
No file chosen
Summarize Text

Summarized Output :



Text was summarized succesfully in 11.51s!!!

Summary Of Text

In 1831–1835, mathematician and engineer Giovanni Plana devised a Perpetual Calendar machine, which, through a system of pulleys and cylinders and over, could predict the perpetual calendar for every year from AD 0 (that is, 1 BC) to AD 4000, keeping track of leap years and varying day length

The set of arithmetic operations that a particular ALU supports may be limited to addition and subtraction, or might include multiplication, division, trigonometry functions such as sine, cosine, etc., and square roots

The machine was huge, weighing 30 tons, using 200 kilowatts of electric power and contained over 18,000 vacuum tubes, 1,500 relays, and hundreds of thousands of resistors, capacitors, and inductors

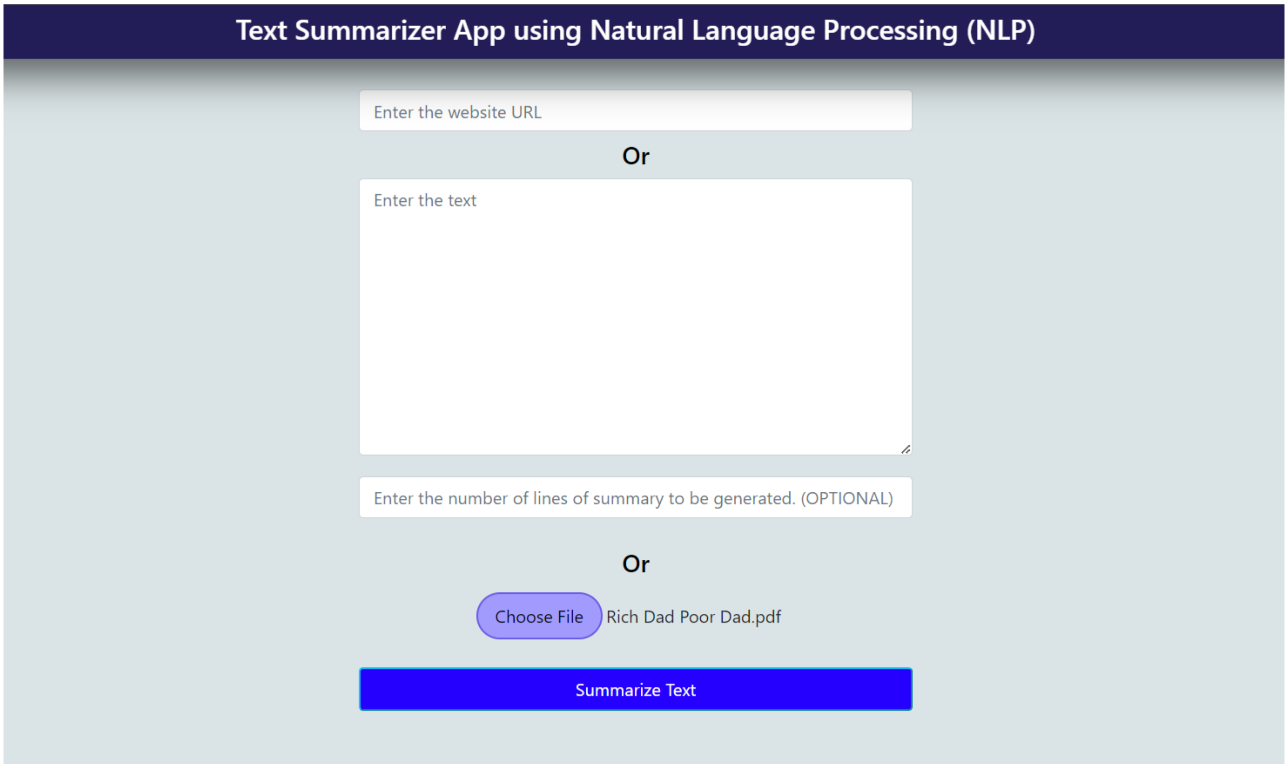
Conventionally, a modern computer consists of at least one processing element, typically a central processing unit (CPU) in the form of a microprocessor, along with some type of computer memory, typically semiconductor memory chips

If not integrated, the RAM is usually placed directly above (known as Package on package) or below (on the opposite side of the circuit board) the SoC, and the flash memory is usually placed right next to the SoC, this all done to improve data transfer speeds, as the data signals don't have to travel long distances.

Original Text

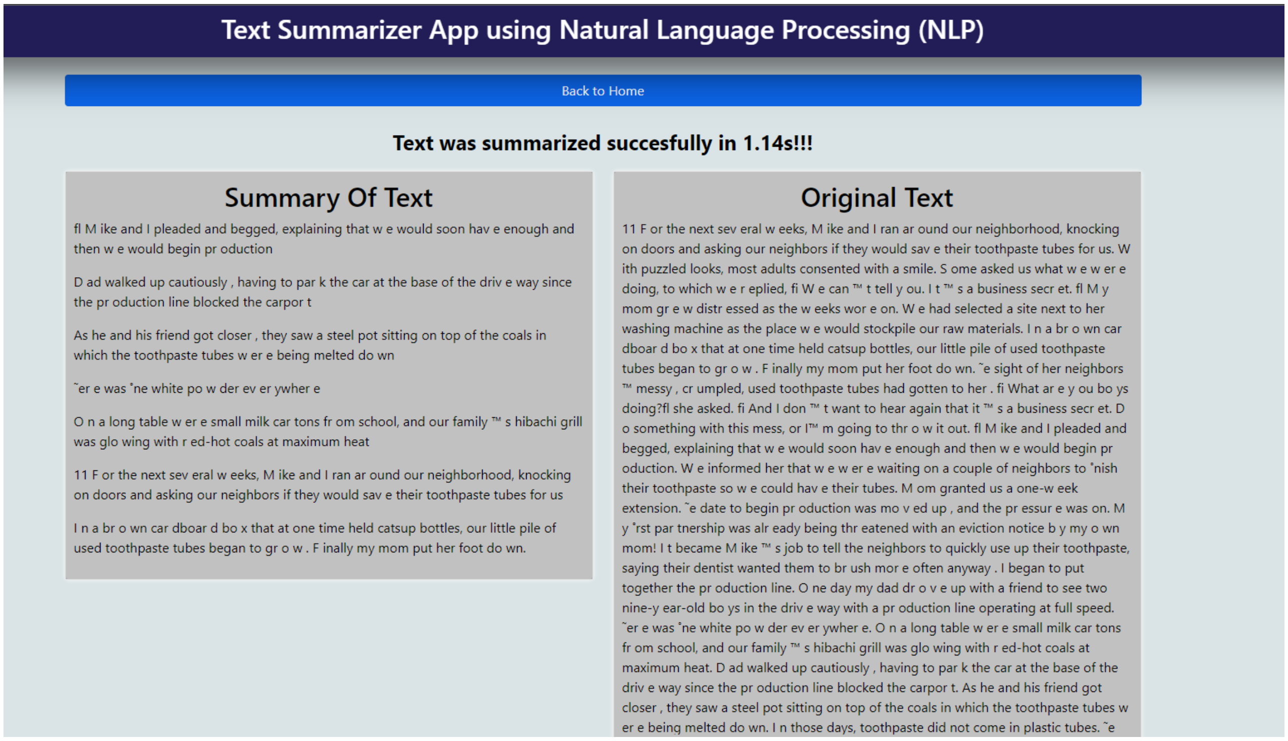
A computer is a digital electronic machine that can be programmed to carry out sequences of arithmetic or logical operations (computation) automatically. Modern computers can perform generic sets of operations known as programs. These programs enable computers to perform a wide range of tasks. A computer system is a "complete" computer that includes the hardware, operating system (main software), and peripheral equipment needed and used for "full" operation. This term may also refer to a group of computers that are linked and function together, such as a computer network or computer cluster. A broad range of industrial and consumer products use computers as control systems. Simple special-purpose devices like microwave ovens and remote controls are included, as are factory devices like industrial robots and computer-aided design, as well as general-purpose devices like personal computers and mobile devices like smartphones. Computers power the Internet, which links billions of other computers and users. Early computers were meant to be used only for calculations. Simple manual instruments like the abacus have aided people in doing calculations since ancient times. Early in the Industrial Revolution, some mechanical devices were built to automate long tedious tasks, such as guiding patterns for looms. More sophisticated electrical machines did specialized analog calculations in the early 20th century. The first digital electronic calculating machines were developed during World War II. The first semiconductor transistors in the late 1940s were followed by the silicon-based MOSFET (MOS transistor) and monolithic integrated circuit (IC) chip technologies in the late 1950s, leading to the microprocessor and the microcomputer revolution in the 1970s. The speed, power and versatility of computers have been increasing dramatically ever since then, with transistor counts increasing at a rapid pace (as predicted by Moore's law), leading to the Digital Revolution during the late 20th to early 21st centuries. Conventionally, a modern computer consists of at least one processing element, typically a central processing unit (CPU) in the form of a microprocessor, along with some type of computer memory, typically semiconductor memory chips. The processing element carries out arithmetic and logical operations, and a sequencing and control unit can change the order of operations in response to stored information. Peripheral devices include input devices (keyboards, mice, joystick, etc.), output devices (monitor screens, printers, etc.), and input/output devices that perform both functions (e.g., the 2000s-era touchscreen). Peripheral devices allow information to be retrieved from an external source and they enable the result of operations to be saved and retrieved. According

Input as PDF:


Text Summarizer App using Natural Language Processing (NLP)
Enter the website URL
Or
Enter the text
Enter the number of lines of summary to be generated. (OPTIONAL)
Or
Choose File
Summarize Text

Rich Dad Poor Dad.pdf

Summarized Output :



Text Summarizer App using Natural Language Processing (NLP)

Text was summarized succesfully in 1.14s!!!

Summary Of Text

fl M ike and I pleaded and begged, explaining that w e would soon hav e enough and then w e would begin pr oduction

D ad walked up cautiously , having to par k the car at the base of the driv e way since the pr oduction line blocked the carpor t

As he and his friend got closer , they saw a steel pot sitting on top of the coals in which the toothpaste tubes w er e being melted do wn

˜er e was ˚ne white po w der ev er ywher e

O n a long table w er e small milk car tons fr om school, and our family ™ s hibachi grill was glo wing with r ed-hot coals at maximum heat

11 F or the next sev eral w eeks, M ike and I ran ar ound our neighborhood, knocking on doors and asking our neighbors if they would sav e their toothpaste tubes for us

I n a br o wn car dboar d bo x that at one time held catsup bottles, our little pile of used toothpaste tubes began to gr o w . F inally my mom put her foot do wn.

Original Text

11 F or the next sev eral w eeks, M ike and I ran ar ound our neighborhood, knocking on doors and asking our neighbors if they would sav e their toothpaste tubes for us. W ith puzzled looks, most adults consented with a smile. S ome asked us what w e w er e doing, to which w e r eplied, fi W e can ™ t tell y ou. I t ™ s a business secr et. fl M y mom gr e w distr essed as the w eeks wor e on. W e had selected a site next to her washing machine as the place w e would stockpile our raw materials. I n a br o wn car dboar d bo x that at one time held catsup bottles, our little pile of used toothpaste tubes began to gr o w . F inally my mom put her foot do wn. ˜e sight of her neighbors ™ messy , cr umpled, used toothpaste tubes had gotten to her . fi What ar e y ou bo ys doing?fl she asked. fi And I don ™ t want to hear again that it ™ s a business secr et. D o something with this mess, or I™ m going to thr o w it out. fl M ike and I pleaded and begged, explaining that w e would soon hav e enough and then w e would begin pr oduction. W e informed her that w e w er e waiting on a couple of neighbors to ˚nish their toothpaste so w e could hav e their tubes. M om granted us a one-w eek extension. ˜e date to begin pr oduction was mo v ed up , and the pr essur e was on. M y ˚rst par tnership was alr eady being thr eatened with an eviction notice b y my o wn mom! I t became M ike ™ s job to tell the neighbors to quickly use up their toothpaste, saying their dentist wanted them to br ush mor e often anyway . I began to put together the pr oduction line. O ne day my dad dr o v e up with a friend to see two nine-y ear-old bo ys in the driv e way with a pr oduction line operating at full speed. ˜er e was ˚ne white po w der ev er ywher e. O n a long table w er e small milk car tons fr om school, and our family ™ s hibachi grill was glo wing with r ed-hot coals at maximum heat. D ad walked up cautiously , having to par k the car at the base of the driv e way since the pr oduction line blocked the carpor t. As he and his friend got closer , they saw a steel pot sitting on top of the coals in which the toothpaste tubes w er e being melted do wn. I n those days, toothpaste did not come in plastic tubes. ˜e

3.6 CONCLUSION AND FUTURE WORK

Text summarization is an interesting machine learning field that is increasingly gaining attraction. As research in this area continues, we can expect to see breakthroughs that will assist in fluently and accurately shortening long text documents.

Hereby, We can say we have successfully completed text summarization using NLP as per problem statement with efficiency. By this project we have solved the problem by the summaries of the text to gain information. We have tried our best to make these summaries as important as possible in the aspect of text intention

We can add various features to our web applications like we can take input of almost any text format like(.doc and .docx,.rtf) by uploading it directly in our input box for text summarization

We can also integrate features like the voice text acceptance for the text summarization. Example, someone reads out loud the text paragraph from the newspaper or passage from novel which is difficult to understand and needs to be summarized.

We have certain limitation while dealing with punctuation marks and spaces so in future we will try to make it as proper as possible.

Reference

- Journals / Conference Papers:
  - [1] Sinha, Aakash, Abhishek Yadav, and Akshay Gahlot. "Extractive text summarization using neural networks." arXiv preprint arXiv:1802.10137 (2018).
  - [2] Peter J. Liu, and Christopher D. Manning. "Get to the point: Summarization with pointer-generator networks." *arXiv preprint arXiv:1704.04368* (2017).
  - [3] Liu, Linqing, et al. "Generative adversarial network for abstractive text summarization." *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 32. No. 1. 2018.
  - [4] Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." *Journal of machine Learning research* 3.Jan (2003): 993-1022.

- Weblinks:
  - [5] Comprehensive Guide to Text Summarization using Deep Learning in Python. - https://www.analyticsvidhya.com/blog/2019/06/comprehensive-guide-text-summarization-using-deep-learning-python/
  - [6] Text Summarization using Machine Learning. = https://data-flair.training/blogs/machine-learning-text-summarization/
  - [7] Approaches to Text Summarization: An Overview. = https://www.kdnuggets.com/2019/01/approaches-text-summarization-overview.html

- [8] “Text summarization approaches for NLP ” - https://www.machinelearningplus.com/nlp/text-summarization-approaches-nlp-example/
- [9] “Beautiful Soup: Build a Web Scraper with Python” accessed on 07 October 2021 - https://realpython.com/beautiful-soup-web-scraper-python/
- [10] “Text Summarization Using Cosine Similarity and Clustering Approach” - https://easychair.org/publications/preprint/p2gV