SCTR's Pune Institute of Computer Technology
Dhankawadi, Pune

A PROJECT REPORT ON

- Develop document summarization system.

SUBMITTED BY

Vinayak Jamadar (41137)
Sanket Jhavar (41138)

Under the guidance of

Prof. U. S. Pawar

DEPARTMENT OF COMPUTER ENGINEERING
Academic Year 2023-24

DEPARTMENT OF COMPUTER ENGINEERING

SCTR's Pune Institute of Computer Technology Dhankawadi, Pune Maharashtra 411043

CERTIFICATE

This is to certify that the SPPU Curriculum-based Mini Project
- Develop document summarization system.

Submitted by

Vinayak Jamadar (41137)
Sanket Jhavar (41138)

has satisfactorily completed the curriculum-based Mini Project under the guidance of Prof. U. S. Pawar towards the partial fulfillment of the final year Computer Engineering Semester VII,
Academic Year 2023-24 of Savitribai Phule Pune University.

Date:

Place: PUNE **Name & Sign of Project Guide:**

Acknowledgment

It gives me great pleasure to present the mini project on - Develop document summarization system.

First of all, I would like to take this opportunity to thank my guide Prof. U. S. Pawar for giving me all the help and guidance needed. I am grateful for his kind support and valuable suggestions that proved to be beneficial in the overall completion of this project.

I am thankful to our Head of the Computer Engineering Department, Dr. G. V. Kale, for her indispensable support and suggestions throughout the internship work. I would also genuinely like to express my gratitude to the CC, Prof. Samadhan Jadhav, for his constant guidance.

Finally, I would like to thank my mentor, Prof. U. S. Pawar for her constant help and support during the overall process.

Title:
- Develop document summarization system.

Problem Statement:
Developing a document summarization system that can automatically generate concise and coherent summaries from diverse document types.

Objective:
- Developing a document summarization system that automatically generates concise and coherent summaries from various document types.
- Preserving the most important content while efficiently condensing lengthy documents, addressing the challenge of information overload in the digital age

Theory:

Text summarization:
Text summarization is the process of distilling the most important information from a source text.

Steps to do text summarization:
- Text cleaning
- Sentence tokenization
- Word tokenization
- Word-frequency table
- Summarization

Text cleaning:
- Text cleaning is the initial step in the text summarization process. It involves preprocessing the raw text data to remove any irrelevant or unnecessary information.
- Common text cleaning tasks include removing special characters, punctuation, and numbers, as well as converting text to lowercase to ensure consistency.
- This step helps improve the quality of the text data and makes it more suitable for subsequent processing.

Sentence tokenization:

- Sentence tokenization is the process of splitting the cleaned text into individual sentences. This is crucial because summarization often involves selecting and condensing entire sentences.
- Tokenization typically relies on linguistic rules and algorithms to identify sentence boundaries, such as periods, question marks, and exclamation points.
- Breaking the text into sentences enables the summarization algorithm to work on a sentence-level basis.

Word tokenization:

- After sentence tokenization, the text is further broken down into individual words or tokens. This is important for analyzing the content at a granular level.
- Word tokenization splits a sentence into its constituent words, allowing for the analysis of word frequencies and relationships between words in the text.
- Common techniques for word tokenization include using spaces as word separators and handling contractions and possessives appropriately.

Word-frequency table:

- The word-frequency table is a data structure that records the frequency of each word in the text.
- For each unique word in the text, the table stores the number of times it appears. This table is instrumental in identifying important words or phrases that occur frequently and could be potential candidates for inclusion in the summary.
- Word frequencies help in selecting significant content for the summary and identifying keywords.

Summarization:

- The summarization step is where the actual document summarization takes place. There are two main approaches to text summarization: extractive and abstractive.
- In **extractive summarization**, sentences or passages from the original text are selected and combined to create a summary. This selection is often based on the importance of sentences, which can be determined using various methods, such as sentence scores derived from word frequencies or other metrics.
- In **abstractive summarization**, the system generates a summary that may not necessarily use the same words as the original text. It aims to create a concise, coherent summary by rephrasing and restructuring the content.
- Summarization algorithms use the information gathered from the previous steps, such as sentence and word tokenization and the word-frequency table, to construct a meaningful summary of the document.

Outcome:

1. **Selected Sentences**: The primary outcome of the extractive summarization approach is a summary composed of a set of selected sentences from the original document. These sentences are directly extracted from the source text.
2. **Preservation of Original Wording**: The selected sentences retain the original wording from the source document, ensuring that the summary remains faithful to the language used in the original text.
3. **Relevance and Importance**: The selected sentences are chosen based on their relevance and importance to the content of the document. This ensures that the summary captures the core ideas and crucial information from the source.
4. **Reduced Length**: The outcome is significantly shorter in length compared to the original document, making it more concise and easier for readers to quickly grasp the key points.
5. **Objective Representation**: Extractive summarization provides an objective representation of the source material by relying on statistical and linguistic features to determine sentence importance.
6. **Efficient Information Retrieval**: Users can efficiently retrieve essential information from the document without the need to read the entire text, making it a valuable tool for information triage.

Result:

Original text:

```
In [170…  print(text)
```

```
Maria Sharapova has basically no friends as tennis players on the WTA Tour. The Russian player has no problems in openly speak
ing about it and in a recent interview she said: 'I don't really hide any feelings too much.
I think everyone knows this is my job here. When I'm on the courts or when I'm on the court playing, I'm a competitor and I wa
nt to beat every single person whether they're in the locker room or across the net.
So I'm not the one to strike up a conversation about the weather and know that in the next few minutes I have to go and try to
win a tennis match.
I'm a pretty competitive girl. I say my hellos, but I'm not sending any players flowers as well. Uhm, I'm not really friendly
or close to many players.
I have not a lot of friends away from the courts.' When she said she is not really close to a lot of players, is that somethin
g strategic that she is doing? Is it different on the men's tour than the women's tour? 'No, not at all.
I think just because you're in the same sport doesn't mean that you have to be friends with everyone just because you're categ
orized, you're a tennis player, so you're going to get along with tennis players.
I think every person has different interests. I have friends that have completely different jobs and interests, and I've met t
hem in very different parts of my life.
I think everyone just thinks because we're tennis players we should be the greatest of friends. But ultimately tennis is just
a very small part of what we do.
There are so many other things that we're interested in, that we do.'
```

Summary:

```
In [171…  print(summary)
```

```
Maria Sharapova has basically no friends as tennis players on the WTA Tour. So I'm not the one to strike up a conversation abo
ut the weather and know that in the next few minutes I have to go and try to win a tennis match.
 I think just because you're in the same sport doesn't mean that you have to be friends with everyone just because you're cate
gorized, you're a tennis player, so you're going to get along with tennis players.
 I have friends that have completely different jobs and interests, and I've met them in very different parts of my life.
 I think everyone just thinks because we're tennis players we should be the greatest of friends.
```

Conclusion:

Our extractive document summarization system successfully condenses documents, preserving key content while providing a concise and efficient summary.

Through a series of well-defined steps including text cleaning, sentence tokenization, word tokenization, the creation of a word-frequency table, and the summarization process, we have successfully created a document summarization system that employs an extractive approach, resulting in a summary composed of selected sentences from the source document.

This approach offers an objective and valuable solution for information retrieval, research, and decision-making, demonstrating its potential in addressing the challenges of information overload in today's digital landscape.