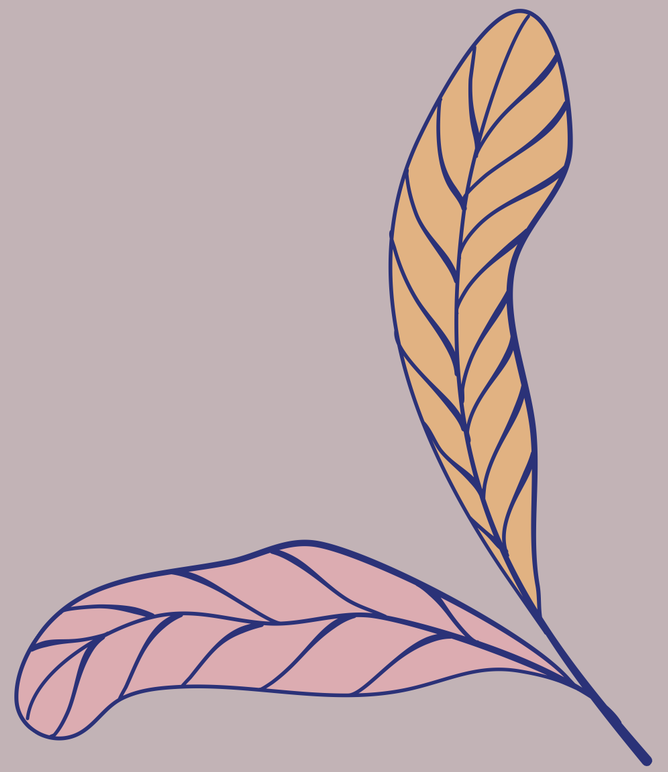
Streamlining News Analysis NLP-Driven Text Summarization

Jenny Johnson
Department of Computer Applications
Amal Jyothi College of Engineering,
Kanjirappally, Kottayam

Mr Ajith G S
Department of Computer Applications
Amal Jyothi College of Engineering,
Kanjirappally, Kottayam

CONTENTS

1 Abstract
2 Introduction
3 Literature Survey
4 Methodology
5 Implementation
6 Conclusion
7 References



Abstract

Leveraging state-of-the-art **NLP models**, the proposed system can generate concise yet comprehensive summaries of news articles, enabling users to quickly grasp the key points and insights without having to sift through lengthy source materials. Users can customize their news sources, select specific topics of interest, and generate summaries from both web-based news articles and uploaded documents. The results of this study show that NLP-driven text summarization can make news analysis faster and more accurate.

Introduction

In the digital age, the **exponential growth of online news** and information has **revolutionized** the way individuals and organizations **consume and process news content**. Traditional methods of news consumption, such as reading full-length articles or relying on human-curated summaries, have become increasingly **time-consuming and inefficient**, failing to keep pace with the sheer volume and velocity of news generation. To address this challenge, this research paper explores the application of **Natural Language Processing** (NLP) techniques, specifically text summarization, as a means of streamlining the news analysis process.

Introduction

- The primary objective of this study is to develop a **user-friendly and versatile system** that can leverage advanced NLP models to automatically generate **concise and informative** summaries of news articles
- The proposed system is designed to be **adaptable and customizable**, allowing users to select their preferred news sources, topics of interest, and even upload their own documents for summarization.
- By integrating both extractive and abstractive summarization techniques, the system aims to provide users with a tailored and efficient news analysis.

Literature Survey

- Mishra et al. [1] **“News text Analysis using Text Summarization and Sentiment Analysis based on NLP.”** They demonstrate how the combination of these techniques can provide a more comprehensive understanding of news content, helping users not only grasp the key information but also gauge the overall **sentiment and tone of the articles**.

Literature Survey

- Gupta and Patel [2] “**Method Of Text Summarization Using Lsa And Sentence Based Topic Modelling With Bert**.”, incorporate **BERT** (Bidirectional Encoder Representations from Transformers), a powerful transformer-based language model, into their **sentence-based** topic modeling approach to improve the quality and coherence of the generated summaries.

Literature Survey

- Ramani et al. [3] “**An Explorative Study on Extractive Text Summarization through k-means, LSA, and TextRank**”. TextRank is a graph-based algorithm that ranks sentences based on their relative importance within the text, similar to the PageRank algorithm used by search engines.

Literature Survey

- Yan and Zhou [4] “**A Text Structure-based Extractive And Abstractive Summarization Method**”. Their approach leverages the hierarchical structure of the input text to identify key sentences and generate new, concise statements to form the final summary.

Literature Survey

- [5] Majid Ramezani “**Unsupervised Broadcast News Summarization; a Comparative Study on Maximal Marginal Relevance** (**MMR**) **and Latent Semantic Analysis** (**LSA**)”.**MMR** is an extractive summarization technique that aims to balance the relevance and diversity of selected sentences, ensuring the summary covers the most important information without being redundant.

Methodology

The proposed system comprises the following key components:

- **Data Collection**: The system incorporates **web scrapers that automatically fetch** news articles from a diverse range of online sources (e.g., Google News, BBC News etc.), and specialized industry-specific platforms. These **web scrapers** can be **configured to retrieve** articles based on **user-specified topics, categories, or keywords.**

Methodology

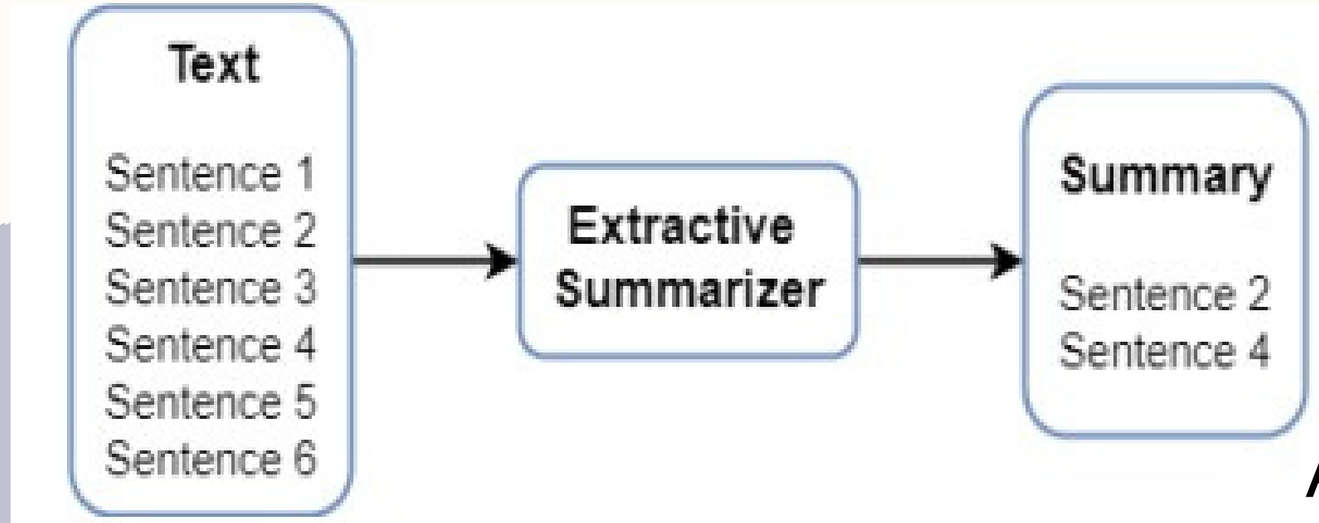
- **Text Extraction**: The system employs tools like the **Newspaper3k** library to extract the relevant text content from the HTML pages, ensuring that only the primary article content is captured and processed. For uploaded documents, the **system supports** a variety of file formats, including **PDF, DOCX, and TXT**, and utilizes appropriate text extraction techniques, such as using the **PyPDF2 and python-docx** libraries

Methodology

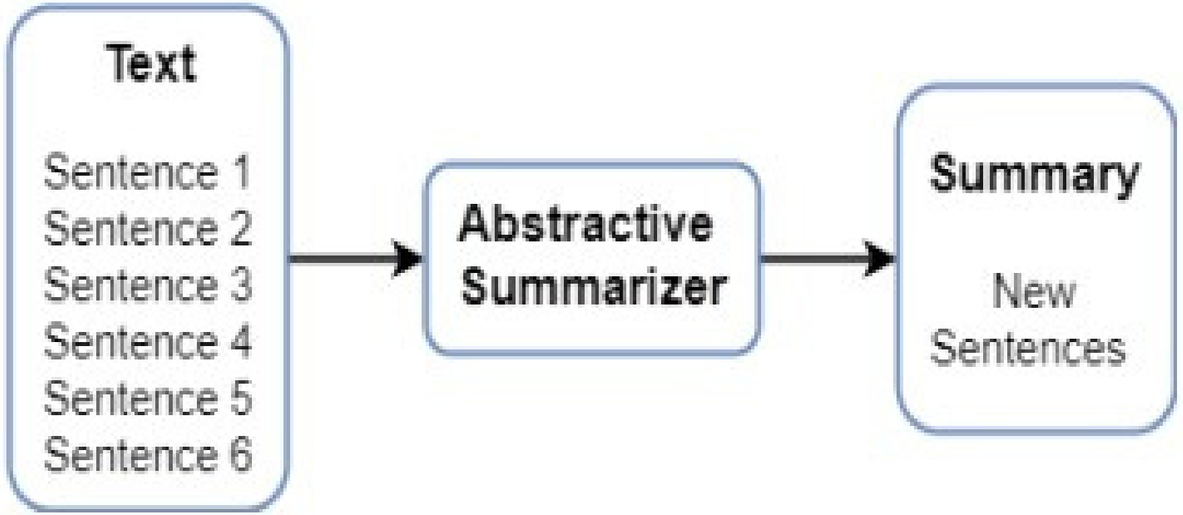
- **Text Summarization**: The core of the system is the text summarization module, which leverages state-of-the-art **NLP models** to generate concise and informative summaries of the extracted news content. This research explores the use of both **extractive and abstractive** summarization techniques.

Methodology

Extractive Summarization



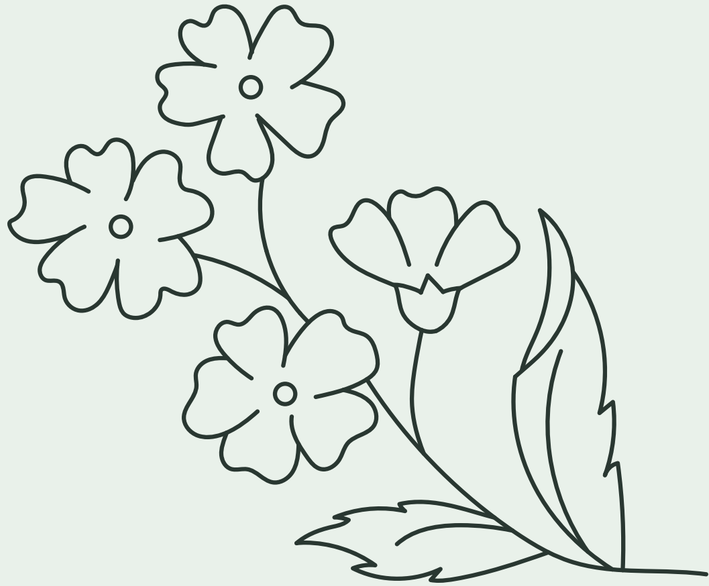
Abstractive Summarization



Implementation

- **Web Scraping**: The system utilizes the **BeautifulSoup** library to fetch news articles from a wide range of online sources

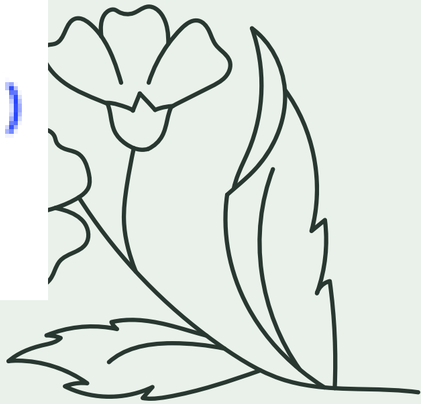
```
def fetch_news_search_topic(topic):
    site = 'https://news.google.com/rss/search?q={}'.format(topic)
    op = urlopen(site)
    rd = op.read()
    op.close()
    sp_page = Soup(rd, 'xml')
    news_list = sp_page.find_all('item')
    return news_list


def fetch_top_news():
    site = 'https://news.google.com/news/rss'
    op = urlopen(site)
    rd = op.read()
    op.close()
    sp_page = Soup(rd, 'xml')
    news_list = sp_page.find_all('item')
    return news_list
```



Implementation

- **Text Extraction:** Newspaper3k library is used to extract the text from pages. For uploaded documents like , PDF, DOCX, and TXT, text extraction techniques, such as using the PyPDF2 and python-docx libraries are used

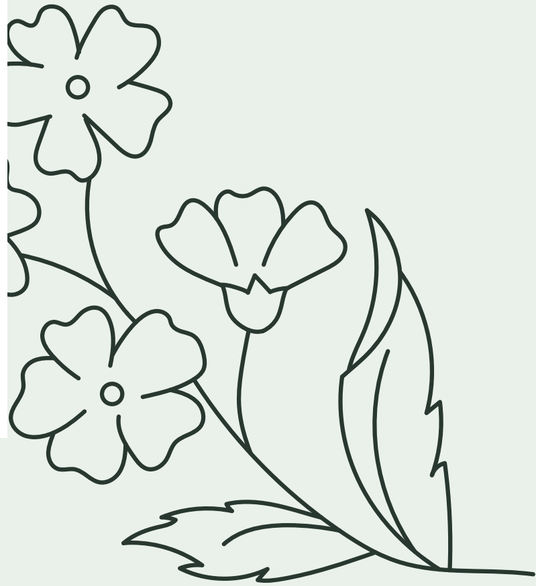
```
def extract_text_from_pdf(uploaded_file):
    pdf_reader = PdfReader(uploaded_file)
    text = ""
    for page_number in range(len(pdf_reader.pages)):
        text += pdf_reader.pages[page_number].extract_text()
    return text
```



Implementation

- **Text Summarization**: The core of the system is the text summarization module, which leverages advanced NLP models.For extractive summarization, the system employs the **Latent Semantic Analysis** (LSA) algorithm.

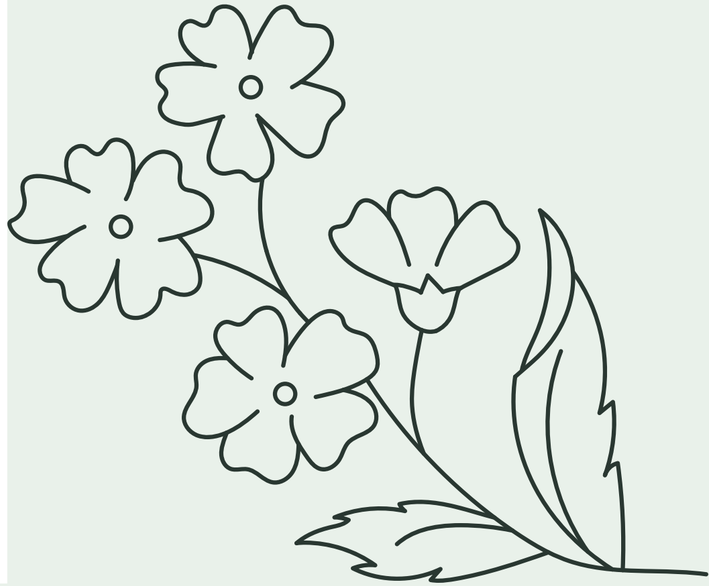
```
def summarize_text(text):
    sentences = sent_tokenize(text)
    plaintext = ' '.join(sentences)
    parser = PlaintextParser.from_string(plaintext, Tokenizer("english"))
    summarizer = LsaSummarizer()
    summary = summarizer(parser.document, 5)
    summarized_text = ' '.join([str(sentence) for sentence in summary])
    return summarized_text
```



Implementation

- **User Interface**: The system is built using the Streamlit framework, providing a user-friendly and responsive web-based interface.

```
def display_news(list_of_news, news_quantity):
    c = 0
    for news in list_of_news:
        c += 1
        st.write('**({}) {}**'.format(c, news.title.text))
        news_data = Article(news.link.text)
        try:
            news_data.download()
            news_data.parse()
            news_data.nlp()
        except Exception as e:
            st.error(e)
        with st.expander(news.title.text):
            st.markdown(
                '''<h6 style='text-align: justify;'>{}"</h6>'''.format(news_data.summary),
                unsafe_allow_html=True)
            st.markdown("[Read more at {}...]({})".format(news.source.text, news.link.text))
        st.success("Published Date: " + news.pubDate.text)
```



Conclusion

The successful implementation and rigorous evaluation of the system demonstrate the efficacy of NLP-driven text summarization in **enhancing** the **efficiency** and **effectiveness** of **news analysis**. This research has the potential to significantly impact diverse domains, from business intelligence and market research to public policy decision-making and crisis response, by streamlining the process of consuming and understanding large volumes of news and information within a short time.

Future scope

- **Multimodal Summarization**: Integration of visual and audio information into the summarization process to provide a more comprehensive and engaging news analysis experience.
- **Multilingual Support**: Extending the system's capabilities to handle news content in multiple languages.
- **Personalization and Recommendation**: Developing advanced mechanisms to personalize the news selection and summarization based on user preferences, browsing history etc.

Reference

[1] Abir Mishra, Akshat Sahay, M Anjusha Pandey & Siddharth Swarup Routaray (2023) “News text Analysis using Text Summarization and Sentiment Analysis based on NLP.”

https://ieeexplore.ieee.org/document/10127895

[2] Hritvik Gupta & Mayank Patel (2021) “Method Of Text Summarization Using Lsa And Sentence Based Topic Modelling With Bert.”

https://ieeexplore.ieee.org/document/9395976

Reference

[3] Kasarapu Ramani, K. Bhavana, A. Akshaya, K. Sai Harshita, C. R. Thoran Kumar & M. Srikanth. (2023) “An Explorative Study on Extractive Text Summarization through k-means, LSA, and TextRank.”

https://ieeexplore.ieee.org/document/10134303

[4] Jing Yan & Shihua Zhou (2022) “A Text Structure-based Extractive And Abstractive Summarization Method.”

https://ieeexplore.ieee.org/document/9778497.

Reference

[5] Majid Ramezani, Mohammad-Salar Shahryari, Amir-Reza Feizi-Derakhshi & Mohammad-Reza Feizi-Derakhshi (2023) “Unsupervised Broadcast News Summarization; a Comparative Study on Maximal Marginal Relevance (MMR) and Latent Semantic Analysis (LSA).”

https://ieeexplore.ieee.org/document/10105403

Thank You