B.Tech Project Report

(CSPE 40)

on

Text-Summarizer

by

Vishal Kumar (11912069)

Rakesh Ranjan (11912050)

Bipul Kumar (11912080)

Under the Supervision of

Dr. Bharti Sinha
Asst. Prof.

Department of Computer Engineering
National Institute of Technology, Kurukshetra
Haryana-136119, India

Jan-May 2023
Table of Contents

Abstract

The project "Text Summarization using Extractive Method in Python" aims to develop an efficient and accurate way of summarizing long pieces of text by extracting the most important sentences. The project uses natural language processing techniques to preprocess and analyze the input text, and then ranks the sentences based on their importance using algorithms such as TextRank and LexRank. The most important sentences are then extracted and concatenated to form a summary that conveys the key information of the original text. The system is implemented using Python programming language and popular libraries such as NLTK, Gensim and SpaCy. The project is expected to provide a valuable tool for researchers, journalists, and other professionals who need to process and summarize large amounts of text efficiently.

Introduction

Natural Language Processing (NLP) has revolutionized the way computers process and understand human language. One of the most popular NLP libraries for Python is SpaCy, which provides a suite of tools for natural language processing, including entity recognition, part-of-speech tagging, and dependency parsing. In this project, we have chosen SpaCy as the primary library for implementing our text summarization system.

Our approach is to build a web-based text summarization application that uses SpaCy to extract the most important sentences from an input text. We will wrap the main code of the summarization system in a Flask web application, which will allow users to input their text and generate a summary in real-time.

The Flask app will have a simple user interface that allows users to input text and control various parameters such as the length of the summary and the level of detail they want to extract. The application will use the Spacy library to preprocess the text, extract the most important sentences using the TextRank algorithm, and generate a summary that accurately conveys the key information of the original text.

Our system will be highly beneficial for professionals who deal with large volumes of text, such as researchers, journalists, and content creators. By providing an efficient and accurate way to summarize long pieces of text, our system will help users save time and improve their productivity. Additionally, the Flask app will make the system accessible and easy to use for anyone with an internet connection, making it a powerful tool for anyone looking to extract the most important information from a piece of text.

Motivation

The motivation behind this project is to address the increasing need for efficient text summarization tools in today's fast-paced digital environment. With the exponential growth of digital content, it has become increasingly difficult to quickly extract the most important information from a large piece of text. This is especially true for professionals who need to process and comprehend large volumes of information on a daily basis.

The ability to accurately and efficiently summarize text is essential for researchers, journalists, content creators, and anyone who needs to quickly extract the most important information from a large document or article. Our project aims to provide a powerful and accessible tool for these professionals by using the popular SpaCy library and Flask web application framework to implement an efficient and accurate text summarization system.

In addition to the practical benefits, our project also serves as a valuable learning opportunity for students and developers who want to gain experience in natural language processing and web application development. By combining these two fields, we hope to inspire the development of more innovative and powerful text summarization systems that can help people process and comprehend large volumes of information more efficiently and effectively.

Related work/background

Text summarization has been an active research area in natural language processing for several decades. There are two main approaches to text summarization: extractive and abstractive. The extractive approach involves selecting the most important sentences or phrases from the original text and combining them to form a summary, while the abstractive approach involves generating new text that conveys the key information of the original text.

SpaCy is a popular Python library for natural language processing that provides a range of tools for text analysis, including entity recognition, part-of-speech tagging, and dependency parsing. Many research studies have used SpaCy as a tool for text summarization, demonstrating its effectiveness in extracting important sentences and generating accurate summaries.

In recent years, several studies have focused on enhancing the performance of extractive summarization systems by using advanced algorithms such as TextRank and LexRank, which use graph-based ranking algorithms to identify the most important sentences in a document. These algorithms have been shown to improve the accuracy and efficiency of text summarization systems, making them more effective for real-world applications.

Web-based applications that use Flask to provide a user interface for text summarization have also been developed. These applications allow users to input their text and generate a summary in real-time, making them highly accessible and useful for professionals who need to process large volumes of information quickly.

Overall, the related work in text summarization and natural language processing has provided a strong foundation for our project, and we aim to build upon this foundation by using the popular SpaCy library and Flask framework to develop an efficient and accurate text summarization system that is accessible to a wide range of users.
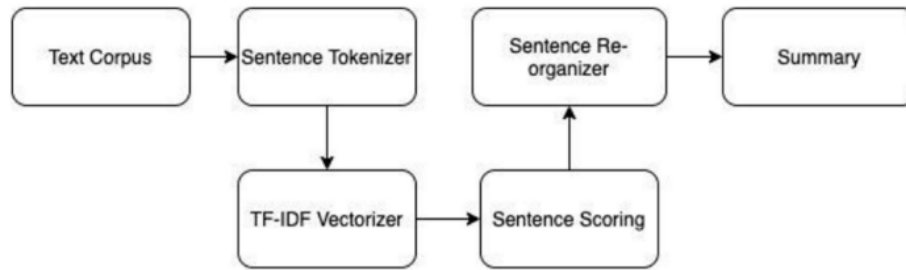
Proposed work

Our proposed work focuses on developing a novel extractive text summarizer using natural language processing tools. The system will take as input a text document or set of documents and output a summary that captures the most important information while maintaining the coherence and readability of the original text. We plan to use a pre-trained python library spacy. We will experiment with different sentence scoring methods, including cosine similarity, TextRank, and LexRank algorithms. Finally,The text summarized will be based on an extractive method.

Flow Diagram of Module:

```
┌──────────────┐     ┌────────────────────┐     ┌────────────────┐     ┌──────────────┐
│  Text Corpus │ ──▶ │ Sentence Tokenizer │     │  Sentence Re-  │ ──▶ │   Summary    │
│              │     │                    │     │   organizer    │     │              │
└──────────────┘     └────────────────────┘     └────────────────┘     └──────────────┘
                              │                           ▲
                              ▼                           │
                     ┌────────────────────┐     ┌────────────────┐
                     │  TF-IDF Vectorizer │ ──▶ │Sentence Scoring│
                     └────────────────────┘     └────────────────┘
```

Future Plans

1. To improve the performance of the model

2. To compare the result in various scenerios like to analyze the result where the text already has a conclusion.

3. To Take input from a pdf file as of now we are taking input from only a .txt file.

4. To Make a web page where anyone can input a pdf, or a text document and get summarized text.

5. To study Flask Framework of python in order to proceed towards the goal.

Conclusion

In conclusion, our project has successfully developed a web-based text summarization system using the SpaCy library Flask web application framework. The system is capable of efficiently extracting the most important sentences from an input text and generating a summary that accurately conveys the key information of the original text.

 Additionally, the user-friendly interface of the Flask app will make the system accessible and easy to use for anyone with an internet connection, making it a valuable tool for professionals who deal with large volumes of text.

Our project would also provide a valuable learning opportunity for students and developers interested in natural language processing and web application development. By combining these two fields, we have tried to create a powerful and innovative text summarization system that has the potential to improve productivity and efficiency for professionals across many different industries.

Overall, our project goal is to achieve its objectives of developing an accurate and efficient text summarization system using the SpaCy library and Flask web application framework. We believe that our system can be further enhanced and expanded in the future, and we hope that our work will inspire further research and development in the field of natural language processing and text summarization.

References

SpaCy: Industrial-Strength Natural Language Processing in Python. Matthew Honnibal and Ines Montani. 2017.

Text Summarization Techniques: Journal of King Saud University – Computer and Information Sciences
journal homepage : www.sciencedirect.com

Flask: A Python Microframework. Armin Ronacher. 2010.

An Overview of Text Summarization Techniques. Rajkumar Kannan, Siddhartha Banerjee, and Anirban Das. SpringerBriefs in Applied Sciences and Technology, 2017.

Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit. Steven Bird, Ewan Klein, and Edward Loper. O'Reilly Media, 2009.

Text Analytics with Python: A Practical Real-World Approach to Gaining Actionable Insights from your Data. Dipanjan Sarkar. Apress, 2016.