Report On

Text Summarization using TextRank

Submitted in partial fulfillment of the requirements of the Course project in Semester VII of fourth year Artificial Intelligence and Data Science

by

Yash Patil (Roll No. 17)

Shubhamkar Patra (Roll No. 35)

Chetan Sapkal (Roll No. 37)

Supervisor

Dr. Tatwadarshi P. N

University of Mumbai

Vidyavardhini's College of Engineering & Technology

Department of Artificial Intelligence and Data Science

(2023-24)

Vidyavardhini's College of Engineering & Technology
Department of Artificial Intelligence and Data Science

CERTIFICATE

This is to certify that the project entitled “Text Summarization Using TextRank” is a bonafide work of “ Yash Patil (Roll No. 18), Shubhamkar Patra (Roll No. 35), Chetan Sapkal (Roll No. 37)” submitted to the University of Mumbai in partial fulfillment of the requirement for the Course project in Semester VII of fourth year Artificial Intelligence and Data Science engineering.

Supervisor

Dr. Tatwadarshi P. N.

Dr. Tatwadarshi P. N.
Head of Department

Table of Contents

Chapter No		Title	Page No.
1		**Abstract**	**1**
2		**Introduction**	**2**
	2.1	Introduction	
	2.2	Problem Statement	
	2.3	Objective	
3		**Proposed System**	**3**
	3.1	Introduction	
	3.2	Details of Hardware and Software	
	3.3	Results	
	3.4	Conclusion	

Chapter 1: Abstract

This project focuses on the development of a comprehensive text summarization system, leveraging a suite of Natural Language Processing (NLP) techniques. The central aim is to automate the process of creating concise summaries of extensive text documents while retaining the most vital content. In addition to summarization, the system offers opportunities for information extraction and content condensation. The project explores techniques such as text preprocessing, the calculation of sentence vectors, creating a similarity matrix, sentence ranking via the PageRank algorithm, and the coherent generation of a summary.

Chapter 2: Introduction

2.1 Introduction

The project embarks on a journey of empowering NLP and text summarization. It starts with the necessary setup and data preparation, including the import of fundamental Python libraries like NumPy, Pandas, NLTK, and Scikit-learn. These libraries are the backbone of the project, facilitating data manipulation and NLP tasks. Additionally, the code acquires pre-trained GloVe word embeddings, critical for converting textual data into vector representations. Furthermore, it downloads NLTK's stopwords dataset, pivotal in the process of text preprocessing. Input text data, often structured as a CSV file, is read and then meticulously processed. The preprocessing not only encompasses the removal of stopwords and punctuation but also involves stemming and lemmatization to ensure that the text's linguistic richness is preserved.

2.2 Problem Statement

The project is a response to the ever-present challenge of handling large volumes of text and information. In an era inundated with data, the need for automated text summarization remains a pressing issue. The goal is to provide a solution that allows users to efficiently generate concise versions of text documents. Text summarization serves as a vital component of applications like information retrieval, search engines, and content summarization in the news and publishing industry.

2.3 Objectives

- Text Preprocessing: The project undertakes the crucial task of preprocessing the input text, which includes not only removing stopwords and punctuation but also addressing issues such as stemming and lemmatization to ensure the context and semantics of the text are accurately represented.
- Sentence Vectorization: Each sentence undergoes a transformation into numerical vectors through pre-trained word embeddings such as GloVe. This not only quantifies the meaning of sentences but also allows for more nuanced context analysis.
- Similarity Matrix: The cosine similarity between sentence vectors is

computed to create a similarity matrix. This matrix lays the foundation for understanding relationships between sentences and their relative importance.

- Sentence Ranking: The PageRank algorithm, famously used in search engines, is applied to rank the sentences. PageRank takes into account the structural relationships within the document, providing a nuanced perspective on sentence importance.
- Summary Generation: The top-ranked sentences are assembled to form a coherent and meaningful summary, which can range from a brief overview to an in-depth summary based on user needs.

Chapter 3: Proposed System

3.1 Introduction

The proposed system leverages Natural Language Processing (NLP) to automate the creation of concise and high-quality text summaries from extensive documents. It involves thorough text preprocessing, linguistic soundness, and advanced semantic analysis through word embeddings like GloVe. The system employs cosine similarity and the PageRank algorithm to rank sentences based on context. It delivers coherent summaries catering to various needs, addressing information overload and knowledge extraction challenges. This system signifies the transformative potential of NLP in reshaping text data management and knowledge extraction.

3.2 Details of Hardware and Software

- Python
- Google Colab

3.3 Results

3.4 Conclusion

In conclusion, the project provides a glimpse into the world of automated text summarization powered by NLP. It introduces an efficient and effective way to extract key information from extensive textual documents. The quality of the summary depends on various factors, including the quality of word embeddings, the specific domain of application, and the extent of customization. The project serves as a starting point for text summarization and can be further refined and adapted for various real-world applications.

This project is a testament to the potential of NLP and its capacity to transform the way we interact with and manage textual data. Automated text summarization represents a critical component of modern information retrieval systems, knowledge management, and content condensation, and this project lays the foundation for the future development of such systems.

References

[1] Dazhi Yang_ and Allan N. Zhang Singapore Institute of Manufacturing Technology "Title of the paper Performing literature review using text mining, Part III: Summarizing articles using Text Rank".

[2] Ali Toofanzadeh Mozhdehi, Mohamad Abdolahi and Shohreh Rad Rahimi title " Overview of extractive text summarization" .

[3] Wengen Li and Jiabao Zhao School of management and engineering title "Text Rank algorithm by exploiting Wikipedia for short text keywords extraction."

[4] Sonya Rapinta Manalu, Willy School of Computer Science title "Stop Words in Review Summarization Using Text Rank ".

[5] Blog Vidhya Analytics https://www.analyticsvidhya.com/blog/2018/11/introduction-text-summarization-textrank-python/