A Project Report on

Automated Text Summarization

Submitted in partial fulfillment of the requirements for the award of the degree of

Bachelor of Engineering

in

Computers

by

Anjali Masur (18102007)
Harshita Jain (18102049)
Kevin Khimasia (18102029)
Sejal Khedekar (18102010)

Under the Guidance of

Dr.Pravin Adivarekar

Department of Computers
NBA Accredited
A.P. Shah Institute of Technology
G.B.Road,Kasarvadavli, Thane(W), Mumbai-400615
UNIVERSITY OF MUMBAI

Academic Year 2021-2022

Approval Sheet

This Project Report entitled ***"Automated Text Summarization"*** Submitted by ***"Anjali Masur"(18102007), "Harshita Jain"(18102049), "Kevin Khimasia" (18102029), "Sejal Khedekar"(18102010)***is approved for the partial fulfillment of the requirement for the award of the degree of ***Bachelor of Engineering*** in ***Computers*** from ***University of Mumbai***.

Dr.Pravin Adivarekar
Guide

Prof. Sachin Malave
Head, Department of Computers

Place:A.P.Shah Institute of Technology, Thane
Date: November 10, 2021

CERTIFICATE

This is to certify that the project entitled ***"Automated Text Summarization"*** submitted by ***"Anjali Masur" (18102007), "Harshita Jain" (18102049), "Kevin Khimasia" (18102029), "Sejal Khedekar" (18102010)*** for the partial fulfillment of the requirement for award of a degree ***Bachelor of Engineering*** in ***Computers***,to the University of Mumbai,is a bonafide work carried out during academic year 2021-2022.

Dr.Pravin Adivarekar
Guide

Prof. Sachin Malave
Head Department of Computers

Dr. Uttam D.Kolekar
Principal

External Examiner(s)

1.

2.

Place:A.P.Shah Institute of Technology, Thane
Date: November 10, 2021

Acknowledgement

We have great pleasure in presenting the report on **Automated Text Summarization.** We take this opportunity to express our sincere thanks towards our guide **Dr.Pravin Adivarekar** & Co-Guide **Co-Guide Name** Department of COMPS, APSIT thane for providing the technical guidelines and suggestions regarding line of work. We would like to express our gratitude towards his constant encouragement, support and guidance through the development of project.

We thank **Prof. Sachin Malave** Head of Department, COMPS, APSIT for his encouragement during progress meeting and providing guidelines to write this report.

We thank **Prof. Amol Kalugade** BE project co-ordinator, Department of COMPS, APSIT for being encouraging throughout the course and for guidance.

We also thank the entire staff of APSIT for their invaluable help rendered during the course of this work. We wish to express our deep gratitude towards all our colleagues of APSIT for their encouragement.

Student Name1: Anjali Masur
Student ID1: 18102007

Student Name2: Harshita Jain
Student ID2: 18102049

Student Name3: Kevin Khimasia
Student ID3: 18102029

Student Name4: Sejal Khedekar
Student ID4: 18102010

Declaration

We declare that this written submission represents our ideas in our own words and where others' ideas or words have been included, We have adequately cited and referenced the original sources. We also declare that We have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in our submission. We understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

(Signature)

(Anjali Masur, 18102007)
(Harshita Jain, 18102049)
(Kevin Khimasia, 18102029)
(Sejal Khedekar, 18102010)

Date: November 10, 2021

Abstract

Text summarization approaches can be split into two groups: extractive summarization and abstractive summarization. The objective of extractive and abstractive summarization is to produce a generalized summary, which conveys information in a precise way that generally requires advanced language generation and compression techniques. The extractive model aims at selecting sentences from the passage or text and picking up the main, essential or relevant information from it as it is without any modification. Thus, it has some limitations. To overcome this and come up with a more concise summary, abstractive text summarization comes into the picture. Abstractive text summarization has a neural network called RNN and CNN which has many layers and shortens the summary generated in extractive text summarization.

Contents

List of Figures

List of Abbreviations

ML:	Machine Learning
NLP:	Natural Language Processing
NLTK:	Natural Language Tool Kit
RNN:	Recurrent Neural Networks
CNN:	Convolutional Neural Networks
LSTM:	Long-Short Term Memory

Chapter 1

Introduction

As of late, there has been a blast in the measure of text data from an assortment of sources. This volume of text is a priceless source of information and knowledge, which should be effectively summarized to be useful. In this problem, the main objective is to automate text summarization. This expanding availability of documents has demanded exhaustive research in automatic text summarization. Because of increasing information on the internet, these kinds of research are gaining more and more attention among the researchers. The whole concept is to reduce or minimize the valuable information present in the documents.This is commonly used by several websites and applications to create news feed and article summaries. It has become very essential for us due to our busy schedules. We prefer short summaries with all the important points over reading a whole report and summarizing it ourselves. So, several attempts had been made to automate the summarizing process.

Chapter 2

Project Concept

2.1 Objectives

- To generate a summary of a text/paragraph by providing the input in the form of a paragraph or a file.
- To generate bullet/key points of a paragraph that would cover only the important points and won't be as long as a summary.
- Summarization of a text/paragraph would be achieved by extractive and abstractive approaches.

2.2 Literature Review

Title : Text summarization using neural networks Authors : Mr Anish Jadhav, Mr Rajat Jain, Mr Steve Fernandes, Mrs Sana Shaikh Year of publication : 2019
Extractive Text Summarization is the method of extracting content from the document and combining it to form a text smaller in size. This ensures that only the words having relevance in the document are selected for the summarization. Whereas, Abstractive Text Summarization is capable of depicting information by creating new sentences. It can be divided into Structured and Semantic approaches, each of which can be subdivided into subcategories based on various methods. The methods are:
•Tree-based approach
•Ontology-based approach
•Rule-based approach
•Graph-based approach

Title : Extractive text summarization using sentence ranking Authors : Mrs J.N.Madhuri, Mr Ganesh Kumar Year of publication : 2019
Automatic Text summarization is the technique to identify the most useful and necessary information in a text. It has two approaches 1) Abstractive text summarization and 2) Extractive text summarization. An extractive text summarization means an important information or sentence are extracted from the given text file or original document. In this paper, a novel statistical method to perform an extractive text summarization on single

document is demonstrated. The method extraction of sentences, which gives the idea of the input text in a short form, is presented.

Title : An overview on extractive text summarization Authors : Mr Shohreh Rad Ramini, Mr Ali Toofanzahdeh Mozhdehi Year of publications : 2017
Text summarization is the process of automatically creating and condensing form of a given document and preserving its information content source into a shorter version with overall meaning. According to difference requirements summary with respect to input text, established summarization systems should be created and classified based on the type of input text. In this study, at first, the topic of text mining and its relationship with text summarization are considered.Then a review has been done on some of the summarization approaches and their important parameters for extracting predominant sentences.

2.3 Problem Definition

The need for text summarization is continuously increasing as today's world is getting flooded with a growing number of articles and links to choose from with the expansion of the internet. Human beings tend to read the whole document to develop an understanding of it and generate a summary by keeping the main points in mind. It is getting extremely difficult to obtain the required information from this pool of words and sentences in a short period. Going through all the documents, articles, and different forms of information to manually summarize is extremely time-consuming and exhausting for humans. Summarization helps in saving valuable time and conveys the main essence from which the reader can decide if they want to dig deeper.

2.4 Scope

The aim of this project is to achieve automation of generating a summary for the given set of data by generating a summarized text of fixed word length by extractive summarization techniques. The model designed in the project will be trained such that it will choose important words and sentences from the input text and arrange them to formulate meaningful sentences. We will be implementing abstractive summarization as well where the ambiguity of sentences in the summary will be reduced as this approach generates a summary by framing new sentences that serve the purpose. The existing summarization tools have a restriction on the word length for input text so we will be working on this aspect, and try to remove such barriers.

2.5 Technology Stack

- Python 3.8
- Pandas
- Numpy
- NLP

- NLTK
- RNN
- LSTM
- MySQL
- HTML, Php

2.6 Benefits for Society

- Summaries reduce reading time.
- When researching documents, summaries make the selection process easier.
- Automatic summarization improves the effectiveness of indexing.
- 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 texts they are able to process.

Chapter 3

Project Design

3.1 Proposed System

In this project we will be using abstractive approach to summarize the text. We would create our own model and integrate it with the website to provide user-friendly interface. Users can generate the summary by adding text or by uploading files. Non-registered users can summarize the text with specific word and usage limit. Whereas, registered users would have the benefit of summarizing the text without any usage limit and their summary would be saved in the database for a week so that they can revisit and access their summarized text.

3.2 Design (Flow of Modules) & Class Diagram

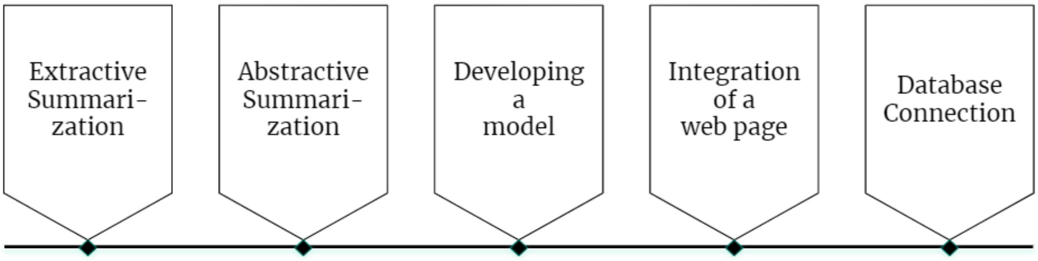
Extractive Summari- zation
Abstractive Summari- zation
Developing a model
Integration of a web page
Database Connection

Figure 3.1: Flow of Modules

As shown in the Fig. 3.1, both extractive and abstractive summarization approaches will be implemented in this project, and in the order represented in the figure. AAbstractive summarization will be used for actual text summarization whereas extractive text summarization will be used for generation of notes. A ML model will be designed on successful completion of abstractive summarization. To provide User Interface (UI) we will integrate a web page with the ML model and add database connectivity to store the user credentials and their activity on the website.

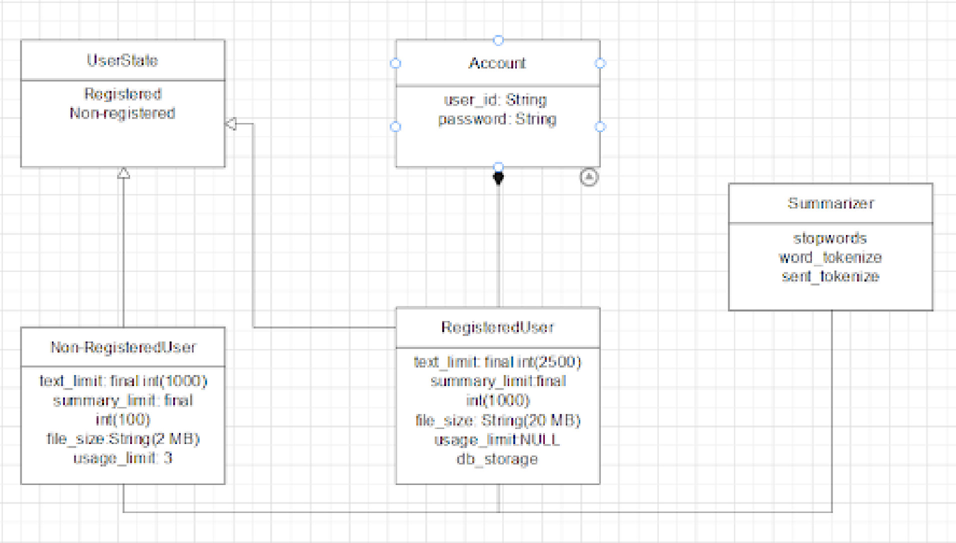

Figure 3.2: Class Diagram Model

3.3 Modules

3.3.1 Module 1: User State (Web Page)

Whenever the application will be accessed, login status will be checked. If login hasn't been done then the user will be able to access only few of the features of the application. Upon logging in, the registered user can access additional features, helpful for summarization of text.

3.3.2 Module 2: Non-Registered User

If any user hasn't logged in or hasn't registered, they will be able to use the application but with few restrictions. A limit will be set in terms of number of words for adding text for which summarization is required. (word limit: 1000) Similarly a limit for summarized text will also be applied. (word limit: 100) For non-registered users, there will also be a usage limit i.e. they can use the tool only thrice after which they will have to register if the wish to continue using the application.

3.3.3 Module 3: Account

If users decides upon creating an account they have to go through registering activity. A SQL database will be connected to the web page which will store the user's credentials upon generation of an account. Then the user can login through their credentials and access the account. With the help of an account, the user can track their previous work easily.

3.3.4 Module 4: Registered User

Once logging activity is done user can access additional features and summarize more text. The limit for input text and summarized text will be increased. There will be no limit to the number of times a user can use the tool for summarizing text passages. We also wish to add the feature of storing the summarized text for 7 days after which it will be erased from the database. The user can revisit the application and access the summarized text generated within last 7 days.

3.4 References

- Title of the paper : Text summarization using neural networks Authors : Mr Anish Jadhav, Mr Rajat Jain, Mr Steve Fernandes, Mrs Sana Shaikh Year of publication : 2018
- Title : Extractive text summarization using sentence ranking Authors : Mrs J.N.Madhuri, Mr Ganesh Kumar Year of publication : 2019
- Title : An overview on extractive text summarization Authors : Mr Shohreh Rad Ramini, Mr Ali Toofanzahdeh Mozhdehi Year of publications : 2017
- https://medium.com/analytics-vidhya/simple-text-summarization-using-nltk-eedc36ebaaf8
- https://stackabuse.com/text-summarization-with-nltk-in-python/
- https://towardsdatascience.com/a-quick-introduction-to-text-summarization-in-machine-learning-3d27ccf18a9f
- https://www.analyticsvidhya.com/blog/2018/11/introduction-text-summarization-textrank-python/

Chapter 4

Planning for Next Semester

- Developing a model for abstractive summarization using RNN and LSTM networks which will be the basis of text summarization.
- The extractive summarization achieved currently will be used for notes generation.
- Once we complete the model, we will integrate it with a web page that would act as the UI for this project.
- Additionally, we will establish database connection to store the activity of users.

Bibliography

[1] Anish Jadhav, Rajat Jain, Steve Fernandes, Sana Shaikh,“Text summarization using neural networks”,IEEE International Conference on Advances in Computing, Communication and Control (ICAC3), 2019.

[2] J.N.Madhuri, R.Ganesh Kumar,“Extractive text summarization using sentence ranking”, IEEE International Conference on Data Science and Communication (IconDSC), 2019.

[3] Shohreh Rad Ramini, Ali Toofanzahdeh Mozhdehi, “An overview on extractive text summarization”, IEEE International Conference on Knowledge-Based Engineering and Innovation (KBEI), 2017.