TEXT SUMMARIZER USING DEEP LEARNING

PROJECT REPORT

Project Lab (IAI-851)

Bachelor of Technology CSE-AI (I-Nurture)

PROJECT GUIDE:

Mr. Sudhanshu Kumar

SUBMITTED BY:

Anshika Gupta (TCA1960003)

Shivansh Sharma (TCA1960001)

FEBRUARY ,2023

FACULTY OF ENGINEERING & COMPUTING SCIENCES

TEERTHANKER MAHAVEER UNIVERSITY,

MORADABAD

DECLARATION

We hereby declare that this Project Report titled **TEXT SUMMARIZER USING DEEP LEARNING** submitted by us and approved by our project guide, Faculty of Engineering & Computing Sciences. Teerthanker Mahaveer University, Moradabad, is a bonafide work undertaken by us and it is not submitted to any other University or Institution for the award of any degree diploma / certificate or published any time before.

Project ID : **IAI-851**

Student Name: **Shivansh Sharma**

Student Name: **Anshika Gupta**

Project Guide : **Mr. Sudhanshu Kumar**

Table of Contents

Appendix

A: Data Flow Diagram (DFD)

B: Entity Relationship Diagram (ERD)

C: Use Case Diagram (UCD)

D: Screen Shots

1 Project Title

TEXT SUMMARIZER USING DEEP LEARNING

2 Problem Statement

In this new era, where tremendous information is available on the internet, it is most important to provide the improved mechanism to extract the information quickly and most efficiently . It is very difficult for human beings to manually extract the summary of a large documents of text. There are plenty of text material available on the internet.

So there is a problem of searching for relevant documents from the number of documents available, and absorbing relevant information from it.In order to solve the above two problems, the automatic text summarization is very much necessary. Text summarization is the process of identifying the most important meaningful information in a document or set of related documents and compressing them into a shorter version preserving its overall meanings.

Customer reviews can often be long and descriptive. Analyzing these reviews manually, as you can imagine, is really time-consuming. This is where the brilliance of Natural Language Processing can be applied to generate a summary for long reviews. We will be working on a really cool dataset.

Our objective here is to generate a summary for the Amazon Fine Food reviews using the abstraction-based approach we learned about above. You can download the dataset from Kaggle.

To create a text summarizer which summarizes the text or the content of the paragraph in minimum words without changing its meaning. This system is made using NLP and deep learning based model which is branch of machine learning. This text summarizer also summarizes text from the weblinks and also summarizes text from PDF document.

3 Project Description

In the modern Internet age, textual data is ever increasing. Need some way to condense this data while preserving the information and meaning. We need to summarize textual data for that. Text summarization is the process of automatically generating natural language summaries from an input document

while retaining the important points. It would help in easy and fast retrieval of information.

Text summarization is the process of generating short, fluent, and most importantly accurate summary of a respectively longer text document (Brownlee, 2017a). The main idea behind automatic text summarization is to be able to find a short subset of the most essential information from the entire set and present it in a human-readable format. As online textual data grows, automatic text summarization methods have potential to be very helpful because more useful information can be read in a short time.

There are two prominent types of summarization algorithms.

• **Extractive summarization** systems form summaries by copying parts of the source text through some measure of importance and then combine those part/sentences together to render a summary. Importance of sentence is based on linguistic and statistical features.

• **Abstractive summarization** systems generate new phrases, possibly rephrasing or using words that were not in the original text. Naturally abstractive approaches are harder. For perfect abstractive summary, the model has to first truly understand the document and then try to express that understanding in short possibly using new words and phrases. Much harder than extractive. Has complex capabilities like generalization, paraphrasing and incorporating real world knowledge. Majority of the work has traditionally focused on extractive approaches due to the easy of defining hard-coded rules to select important sentences than generate new ones. Also, it promises grammatically correct and

coherent summary. But they often don't summarize long and complex texts well as they are very restrictive.

3.1 Scope of the Work

The goal of automatic text summarization is presenting the source text into a shorter version with semantics. The most important advantage of using a summary is ,it reduces the reading time. Text Summarization methods can be classified into extractive and abstractive summarization. An extractive summarization method consists of selecting important sentences, paragraphs etc. from the original document and concatenating them into shorter form. An Abstractive summarization is an understanding of the main concepts in a document and then express those concepts in clear natural language. There are two different groups of text summarization : indicative and informative. Inductive summarization only represent the main idea of the text to the user. The typical length of this type of summarization is 5 to 10 percent of the main text. On the other hand, the informative summarization systems gives concise information of the main text .The length of informative summary is 20 to 30 percent of the main text .

3.2 Project Modules

Module 1

- **Data Selection**

Data selection is defined as the process of determining the appropriate data type and source, as well as suitable instruments to collect data. Data selection precedes the actual practice of data collection.

- **Data Cleaning**

Data cleaning is one of the important parts of machine learning. It plays a significant part in building a model. It surely isn't the fanciest part of machine learning and at the same time, there aren't any hidden tricks or secrets to uncover. However, the success or failure of a project relies on proper data cleaning.

Module 2

- **Feature Selection**

Feature Selection is the method of reducing the input variable to your model by using only relevant data and getting rid of noise in data.

It is the process of automatically choosing relevant features for your machine learning model based on the type of problem you are trying to solve. We do this by including or excluding important features without changing them. It helps in cutting down the noise in our data and reducing the size of our input data.

- **Model Training**

A machine learning training model is a process in which a machine learning (ML) algorithm is fed with sufficient training data to learn from.

ML models can be trained to benefit manufacturing processes in several ways. The ability of ML models to process large volumes of data can help manufacturers identify anomalies and test correlations while searching for patterns across the data feed. It can equip manufacturers with predictive maintenance capabilities and minimize planned and unplanned downtime.

Module 3

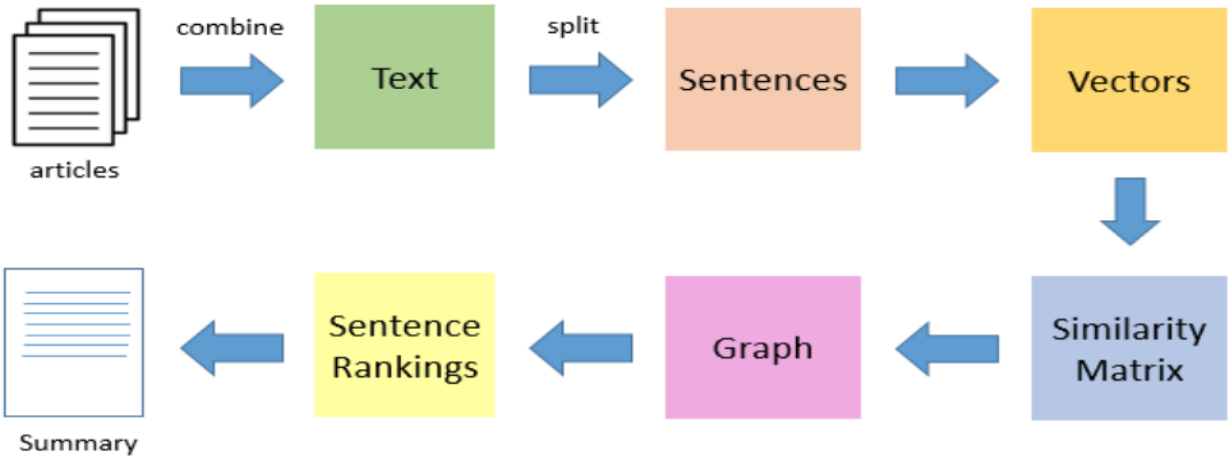
- **Model Evaluation**

Model evaluation is the process of using different evaluation metrics to understand a machine learning model's performance, as well as its strengths and weaknesses. Model evaluation is important to assess the efficacy of a model during initial research phases, and it also plays a role in model monitoring.

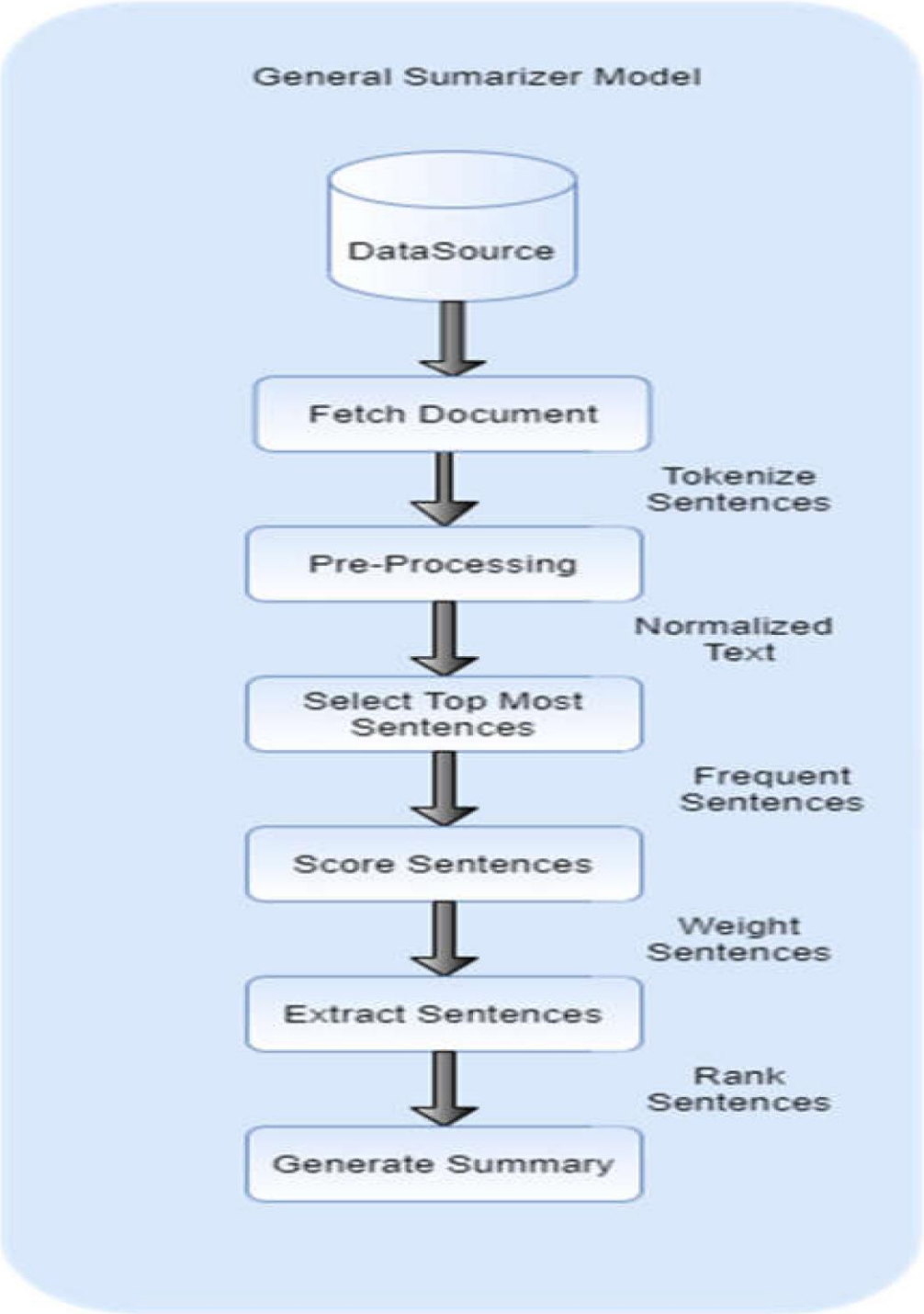
To understand if your model(s) is working well with new data, you can leverage a number of evaluation metrics.

1. **Accuracy**
2. **Precision**
3. **Confusion Matrix**
4. **Log-Loss**
5. **AUC(Area under Curve)**

3.3 Context Diagram (High Level)



4 Implementation Methodology


General Sumarizer Model
DataSource
Fetch Document
Tokenize Sentences
Pre-Processing
Normalized Text
Select Top Most Sentences
Frequent Sentences
Score Sentences
Weight Sentences
Extract Sentences
Rank Sentences
Generate Summary

5 Technologies to be used

5.1 Software Platform

- Python 3.6.8
- PyCharm (IDE)
- Operating System (Windows 7,8,10,11)

5.2 Hardware Platform

- RAM – Minimum 4GB
- Hard Disk – Minimum 32GB
- Processor i-5, AMD 3 upwards

5.3 Tools, if any

- Frontend
  - Flask
  - HTML/CSS

- Backend
  - Pandas
  - Numpy
  - Scikit-Learn
  - Matplotlib
  - GridSearch CV
  - Linear Regression
  - Decision Tree Regressor

6 Advantages of this Project

- Summarizing reduces perusing time
- While investigating reports, outlines make the determination procedure simpler
- Summarization improves the adequacy of ordering
- Summarization calculations are less one-sided than human summarizers
- Personalized summaries are useful in question-answering systems as they provide personalized information
- Utilizing programmed or Summarization frameworks empower business theoretical administrations to build the number of content archives they can process
- Text summarization can help users quickly grasp the essence and context of the data, and identify the most important or interesting aspects.
- This can improve the efficiency and productivity of data analysis, as well as the communication and presentation of the results.

6.1 Possible current uses of summarization:

1. People need to learn much from texts. But they tend to want to spend less time while doing this.

2. It aims to solve this problem by supplying them the summaries of the text from which they want to gain information.

3. Goals of this project are that these summaries will be as important as possible in the aspect of the texts' intention.

4. The user will be eligible to select the summary length.

5. Supplying the user, a smooth and clear interface.

6. Configuring a fast replying server system

7 Future Scope and further enhancement of the Project

The future study is to build a robust, domain and language independent extractive text summarization that works well with multi-documents. Similarly, because the

quality evaluation of the summary is done manually by experienced assessors, it is highly subjective. There are specific quality assessment criteria, such as grammaticality and coherence, but the results are different when two experts evaluate the same summary.

The project is wide in scope, this project looks at single document summarization - the area of multi document summarization is not covered. Also, the summaries produced are largely extracts of the document being summarized, rather than newly generated abstracts. The parameters used are optimal for news articles, although that can be changed easily.

The model we built for abstractive summarization did a good job on generating humanreadable sentences from given inputs. However, it did not always generate summaries capturing all the important information in the input documents. To solve this problem, based on our research, we propose adding a custom layer to the model that performs attention mechanism (Lopyrev, 2015).

The attention mechanism has been proved to be useful in tasks like abstractive summarization. Lastly, we suggest using larger datasets to train the models. Researchers in the past have trained their text summarization models on millions of documents to achieve good results (Nallapati, Zhou, Santos, Gulçehre, & Xiang 2016). Whereas, due to limited resources, the largest dataset we used only had about twenty thousand articles. If these changes can be applied, we think that the performance of the model may improve.

8 Project Repository Location

S#	Project Artifacts (softcopy)	Location (GitHub links)
1.	Project Synopsis Report (Final Version)	https://github.com/anshikagupta0308/College_Projects/tree/main/Text%
2.	Project Requirement	https://github.com/anshikagupta0308/College_Projects/tree/main/Text%

S#	Project Artifacts (softcopy)	Location (GitHub links)
	specifications	
3.	Project Report (Final Version)	https://github.com/anshikagupta0308/College_Projects/tree/main/Text%
4.	Project Source Code (final version) with executable	https://github.com/anshikagupta0308/College_Projects/tree/main/Text%

9 Conclusion

Automatic text summarization is an old challenge but the current research direction diverts towards emerging trends in biomedicine, product review, education domains, emails and blogs. Automated summarization is an important area in NLP (Natural Language Processing) research. It consists of automatically creating a summary of one or more texts. The purpose of extractive document summarization is to automatically select a number of indicative sentences, passages, or paragraphs from the original document .Text summarization approaches based on Neural Network, Graph Theoretic, Fuzzy and Cluster have,

to an extent, succeeded in making an effective summary of a document.Both extractive and abstractive methods have been researched. Most summarization techniques are based on extractive methods. Abstractive method is similar to summaries made by humans. Abstractive summarization as of now requires heavy machinery for language generation and is difficult to replicate into the domain specific areas.

Text summarization is an interesting machine learning field that is increasingly gaining attraction. As research in this area continues, we can expect to see breakthroughs that will assist in fluently and accurately shortening long text documents. Hereby, We can say we have successfully completed text summarization using NLP as per problem statement with efficiency. By this project we have solved the problem by the summaries of the text to gain information. We have tried our best to make these summaries as important as possible in the aspect of text intention. We can add various features to our web applications like we can take input of almost any text format like(.doc and .docx,.rtf) by uploading it directly in our input box for text summarization. We can also integrate features like the voice text acceptance for the text summarization. Example, someone reads out loud the text paragraph from the newspaper or passage from novel which is difficult to understand and needs to be summarized. We have certain limitation while dealing with punctuation marks and spaces so in future we will try to make it as proper as possible.

We have learned all the basics of Extractive and Abstractive Method of automatic text summarization and tried to implement extractive one. We have made a basic

automatic text summarizer using nltk library using python and it is working on small documents. We have used extractive approach to do text summarization.

10 References

1. Goularte, Fábio & Nassar, Silvia & Fileto, Renato & Saggion, Horacio. (2018). A Text Summarization Method based on Fuzzy Rules and applicable to Automated Assessment. Expert Systems with Applications. 115. 10.1016/j.eswa.2018.07.047.
2. Jo, Duke Taeho. (2017). K nearest neighbor for text summarization using feature similarity. 1-5. 10.1109/ICCCCEE.2017.7866705.
3. Anand, Deepa & Wagh, Rupali. (2019). Effective Deep Learning Approaches for Summarization of Legal Texts. Journal of King Saud University - Computer and Information Sciences. 10.1016/j.jksuci.2019.11.015.
4. P. Krishnaveni and S. R. Balasundaram, "Automatic text summarization by local scoring and ranking for improving coherence," 2017 International Conference on Computing Methodologies and Communication (ICCMC), Erode, 2017, pp. 59-64. doi: 10.1109/ICCMC.2017.8282539

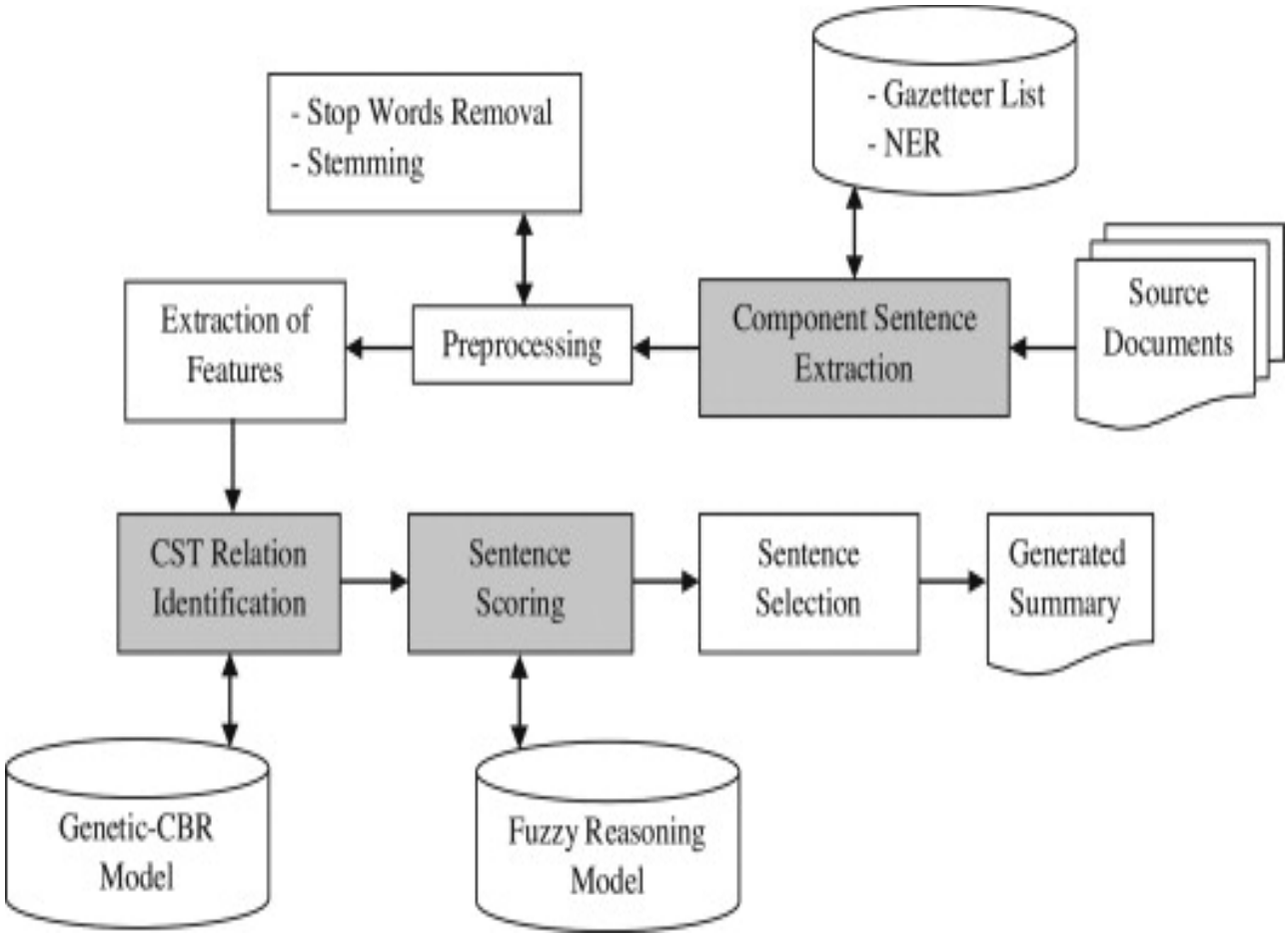
5. J. Chen and H. Zhuge, "Extractive Text-Image Summarization Using Multi-Modal RNN," 2018 14th International Conference on Semantics, Knowledge and Grids (SKG), Guangzhou, China, 2018, pp. 245-248. doi: 10.1109/SKG.2018.00033
6. Valverde Tohalino, Jorge & Amancio, Diego. (2017). Extractive Multidocument Summarization Using Multilayer Networks. Physica A: Statistical Mechanics and its Applications. 503. 10.1016/j.physa.2018.03.013.
7. J. N. Madhuri and R. Ganesh Kumar, "Extractive Text Summarization Using Sentence Ranking," 2019 International Conference on Data Science and Communication (IconDSC), Bangalore, India, 2019, pp.1- 3.doi: 10.1109/IconDSC.2019.8817040
8. N. S. Shirwandkar and S. Kulkarni, "Extractive Text Summarization Using Deep Learning," 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), Pune, India, 2018, pp. 1-5. doi: 10.1109/ICCUBEA.2018.8697465
9. Azar, Mahmood & Hamey, Len. (2016). Text Summarization Using Unsupervised Deep Learning. Expert Systems with Applications. 68. 10.1016/j.eswa.2016.10.017.
10. Jain, Aditya & Bhatia, Divij & Thakur, Manish. (2017). Extractive Text Summarization Using Word Vector Embedding. 51-55. 10.1109/MLDS.2017.12.
11. K.Selvani Deepthi, Dr.Radhika Y(2015),“Extractive Text Summarization using Modified Weighing and Sentence Symmetric Feature Methods”, in an International Journal of Modern Education and Computer Science, ISSN: 2075-0161 Volume 7 No-10 pp: 33-39, October 2015.
12. Ahmad T. Al-Taani. “,Automatic Text Summarization Approaches” International Conference on Infocom Technologies and Unmanned Systems (ICTUS'2017)
13. Neelima Bhatia, ArunimaJaiswal, “Automatic Text Summarization: Single and Multiple Summarizations ”, International Journal of Computer Applications

14. Mehdi Allahyari, SeyedaminPouriyeh, Mehdi Assefi, SaeidSafaei, Elizabeth D. Trippe, Juan B. Gutierrez, KrysKochut, " Text Summarization Techniques: A Brief Survey", (IJACSA) International Journal of Advanced Computer Science and Applications
15. Pankaj Gupta, Ritu Tiwari and NirmalRobert,"Sentiment Analysis and Text Summarization of Online Reviews: A Survey"InternationalConzatiference on Communication and Signal Processing,August 2013
16. Vishalgupta,Gurpreet Singh Lehal,"A Survey of Text Summarization Extractive Techniques."JOURNAL OF EMERGING TECHNOLOGIES IN WEB INTELLIGENCE, VOL. 2, NO. 3, AUGUST 2010
17. Jiwei Tan, XiaojunWan,Jianguo Xiao Institute of Computer Science and Technology,Peking University "Abstractive document summarization with a GraphBased attentional neural model. "
18. SeonggiRyang, Graduate school of Information science and technology, University of Tokyo Takeshi Abekawa, National institute of informatics "Framework of automatic text summarization using Reinforcement learning" 48
19. Tianshi, YaserKeneshloo, Narenramakrishnan, Chandan K. Reddy, Senior member, IEEE " Neural Abstractive text summarization with sequence-to - sequence models"
20. Josef Steinberger, KarelJežek, "Using latent Semantic analysis In Text Summarization and Summary Evaluation", Department of Computer Science and Engineering, UniverzitníCZ-306 14 Plzeň.
21. Sumitha C., Dr. A. Jaya, Amal Ganesh, "A study on Abstract Summarization Techniques in Indian Languages", Elsevier Proceeding of Computer Science, No. 87, pp.25-31, 2016.
22. Dipanjan Das, Andre F.T. Martins, "A Survey on Automatic Text Summarization",Language Technologies Inst

Annexure A

Data Flow Diagram (DFD)

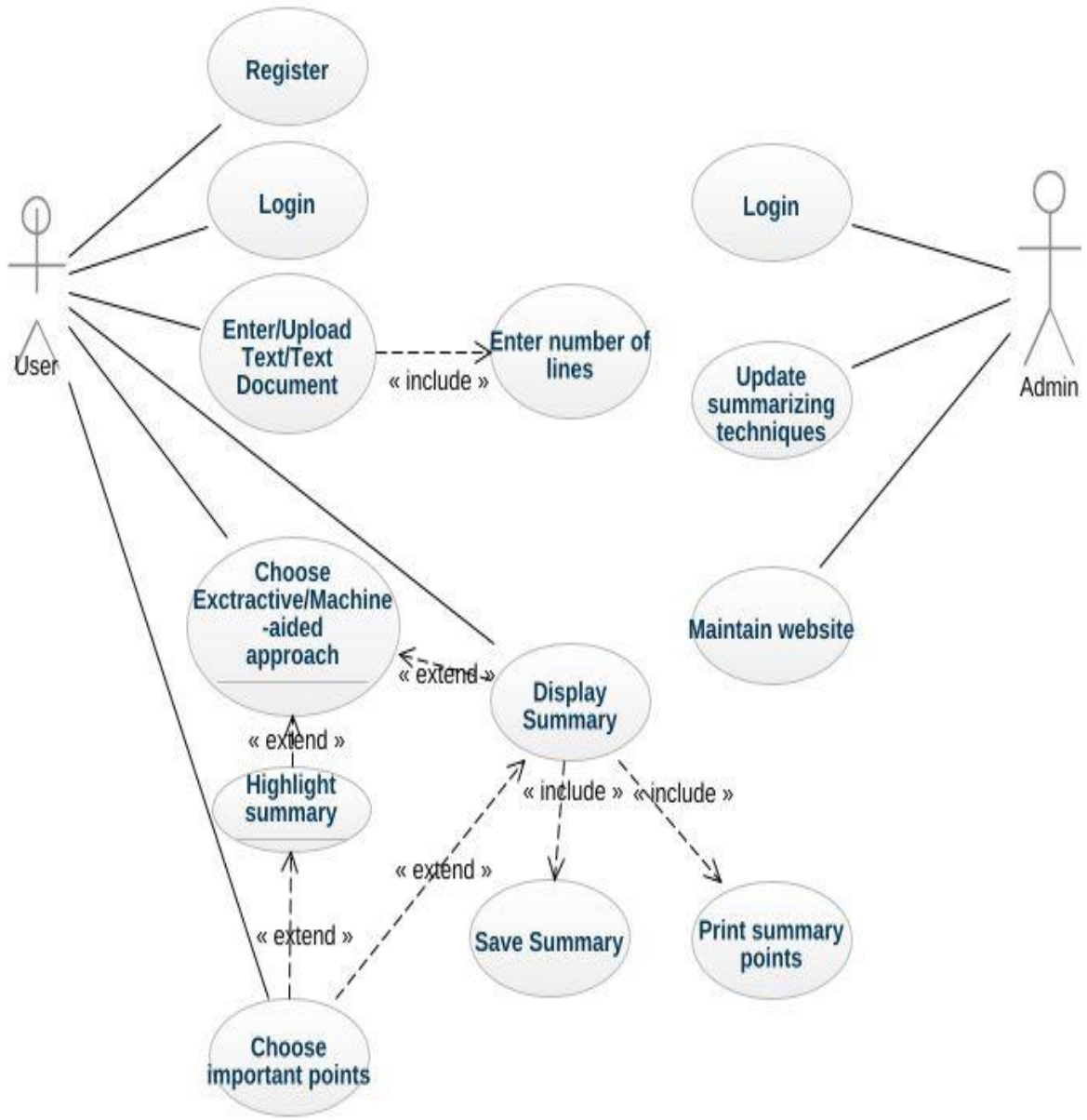
(Mandatory)


- Stop Words Removal
- Stemming
- Gazetteer List
- NER
Extraction of Features
Preprocessing
Component Sentence Extraction
Source Documents
CST Relation Identification
Sentence Scoring
Sentence Selection
Generated Summary
Genetic-CBR Model
Fuzzy Reasoning Model

Annexure B

Entity-Relationship Diagram (ERD)

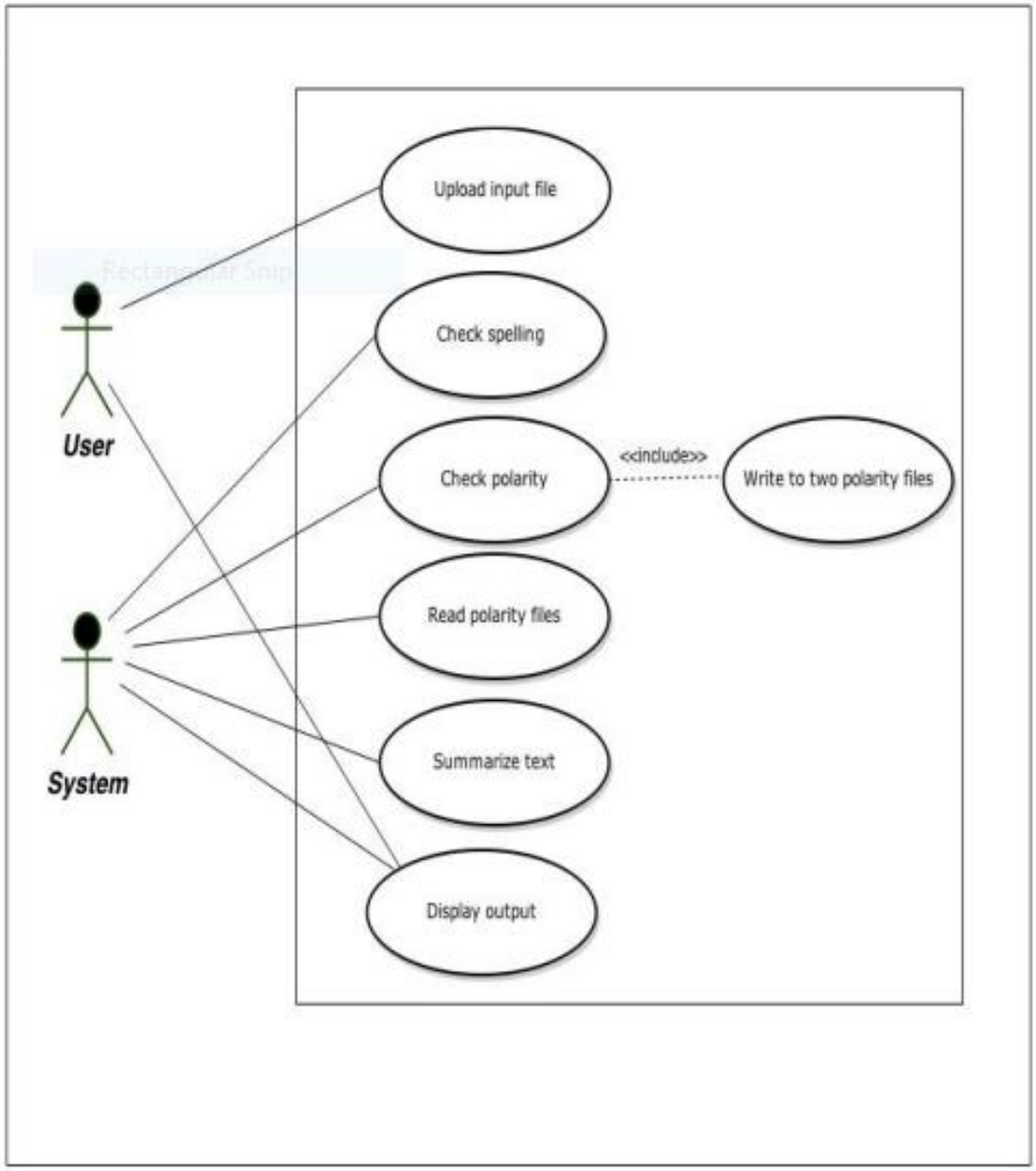
(Mandatory)



Annexure C

Use-Case Diagram (UCD)

(Optional)



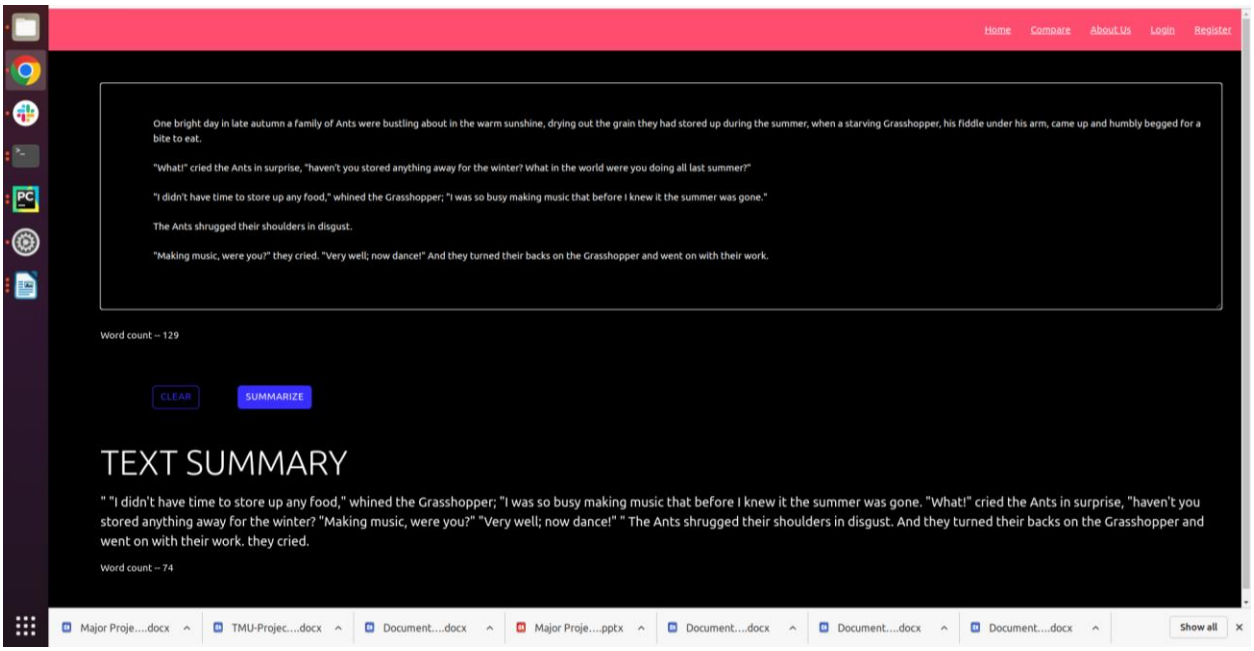
Annexure D

Screen Shots

Login Page:

Register Page:

Home Page:


Home
Compare
About Us
Login
Register
One bright day in late autumn a family of Ants were bustling about in the warm sunshine, drying out the grain they had stored up during the summer, when a starving Grasshopper, his fiddle under his arm, came up and humbly begged for a bite to eat.
"What!" cried the Ants in surprise, "haven't you stored anything away for the winter? What in the world were you doing all last summer?"
"I didn't have time to store up any food," whined the Grasshopper; "I was so busy making music that before I knew it the summer was gone."
The Ants shrugged their shoulders in disgust.
"Making music, were you?" they cried. "Very well; now dance!" And they turned their backs on the Grasshopper and went on with their work.
Word count -- 129
CLEAR
SUMMARIZE
TEXT SUMMARY
" "I didn't have time to store up any food," whined the Grasshopper; "I was so busy making music that before I knew it the summer was gone. "What!" cried the Ants in surprise, "haven't you stored anything away for the winter? "Making music, were you?" "Very well; now dance!" " The Ants shrugged their shoulders in disgust. And they turned their backs on the Grasshopper and went on with their work. they cried.
Word count -- 74
Show all