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

Teerthanker Mahaveer University, Moradabad

Faculty of Engineering and Computing Sciences

Table of Contents

1.Project Title

TEXT SUMMARIZER USING DEEP LEARNING

2.Domain

Data science is the field of study that combines domain expertise, programming skills, and knowledge of mathematics and statistics to extract meaningful insights from data. Data science practitioners apply machine learning algorithms to numbers, text, images, video, audio, and more to produce artificial intelligence (AI) systems to perform tasks that ordinarily require human intelligence. In turn, these systems generate insights which analysts and business users can translate into tangible business value.

Natural Language Processing

Natural language processing (NLP) is the ability of a computer program to understand human language as it is spoken and written -- referred to as natural language. It is a component of artificial intelligence (AI).

NLP has existed for more than 50 years and has roots in the field of linguistics. It has a variety of real-world applications in a number of fields, including medical research, search engines and business intelligence.

NLP enables computers to understand natural language as humans do. Whether the language is spoken or written, natural language processing uses artificial intelligence to take real-world input, process it, and make sense of it in a way a computer can understand. Just as humans have different sensors -- such as ears to hear and eyes to see -- computers have programs to read and microphones to collect audio. And just as humans have a brain to process that input, computers have a program to process their respective inputs. At some point in processing, the input is converted to code that the computer can understand.

There are two main phases to natural language processing: data preprocessing and algorithm development.

Data preprocessing involves preparing and "cleaning" text data for machines to be able to analyze it. preprocessing puts data in workable form and highlights

features in the text that an algorithm can work with. There are several ways this can be done, including:

- **Tokenization**. This is when text is broken down into smaller units to work with.
- **Stop word removal**. This is when common words are removed from text so unique words that offer the most information about the text remain.
- **Lemmatization and stemming**. This is when words are reduced to their root forms to process.
- **Part-of-speech tagging**. This is when words are marked based on the part-of speech they are -- such as nouns, verbs and adjectives.

Once the data has been preprocessed, an algorithm is developed to process it. There are many different natural language processing algorithms, but two main types are commonly used:

- **Rules-based system**. This system uses carefully designed linguistic rules. This approach was used early on in the development of natural language processing, and is still used.
- **Machine learning-based system**. Machine learning algorithms use statistical methods. They learn to perform tasks based on training data they are fed, and adjust their methods as more data is processed. Using a combination of machine learning, deep learning and neural networks, natural language processing algorithms hone their own rules through repeated processing and learning.

3. 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.

Objectives

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

• Automatic summarization algorithms are less biased than human summarizers.

• 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 text documents they are able to process.

4. 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.

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.

4.1 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

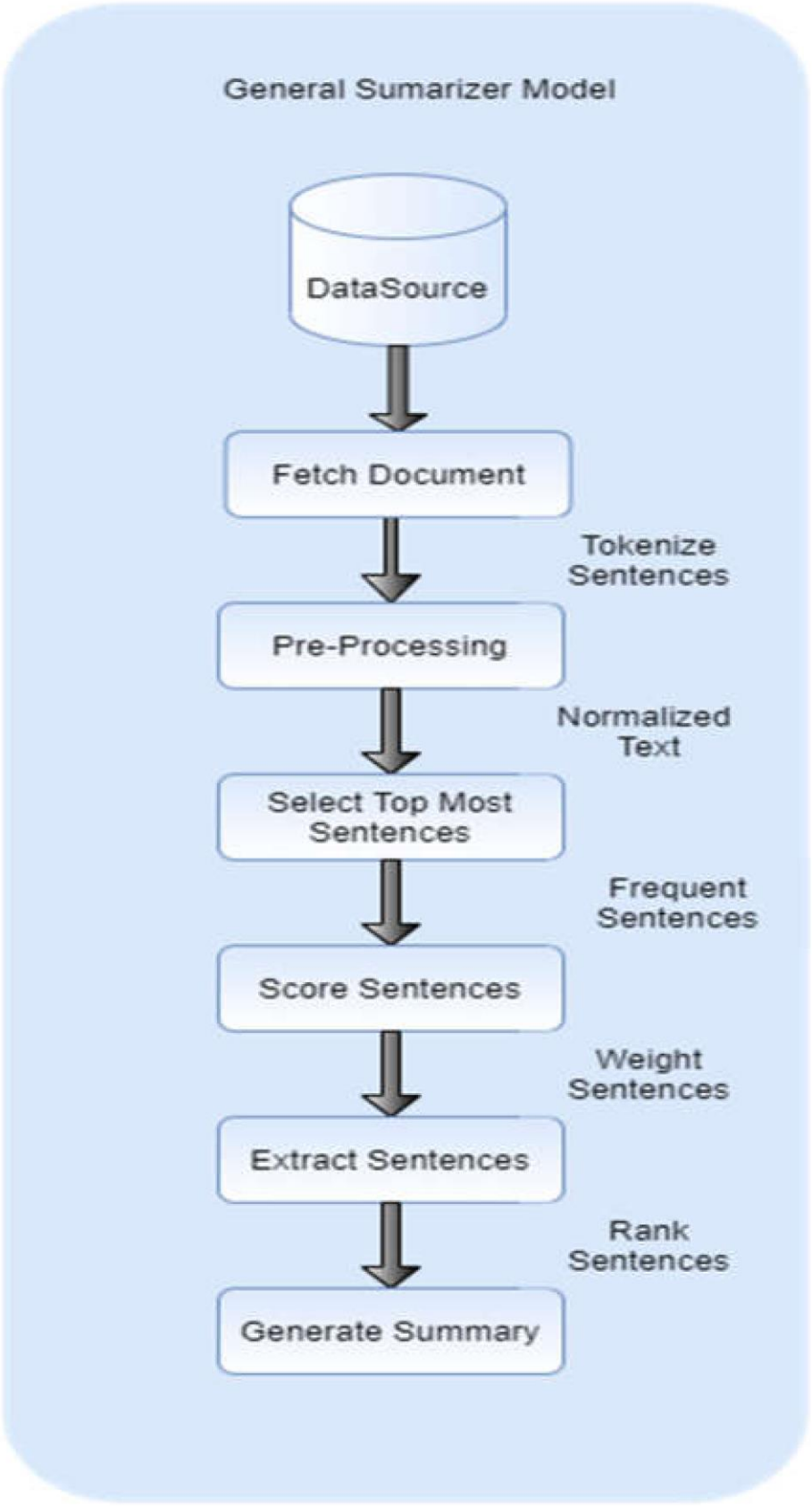
- **Model Evaluation**

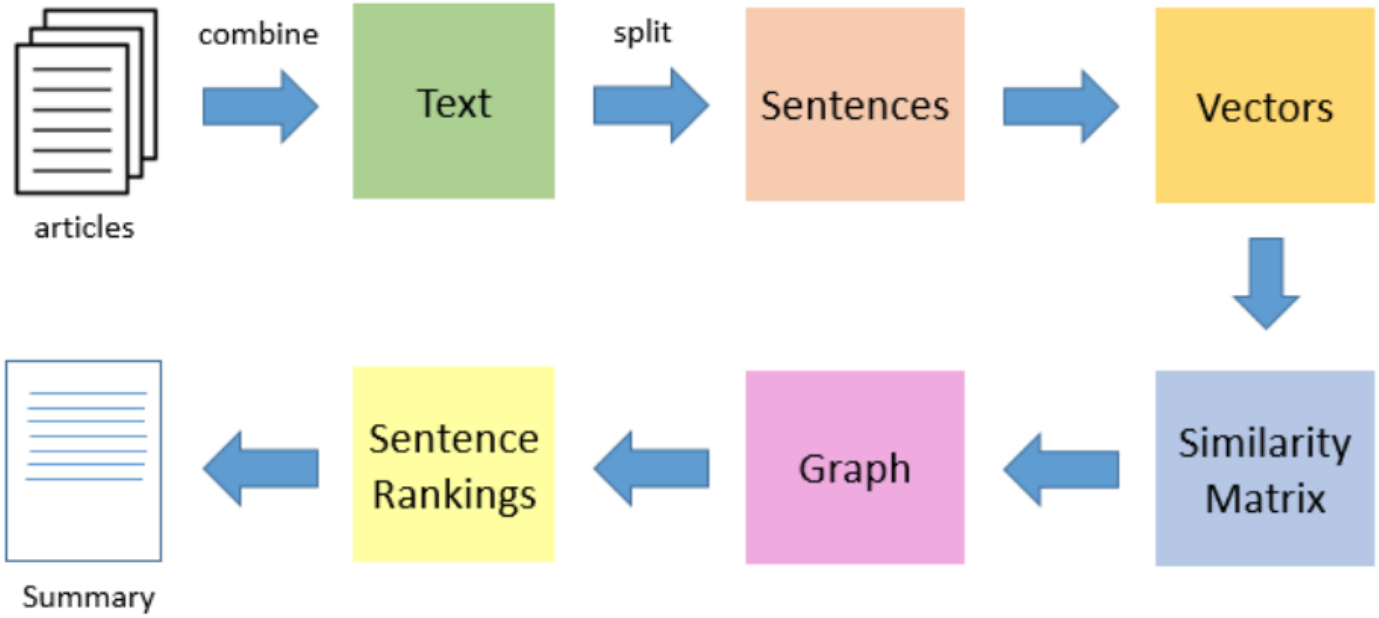
Model evaluation is the process of using different evaluation metrics to understand a machine learning model's performance, as well as its strengths and weaknesses. Model evaluation is important to assess the efficacy of a model during initial research phases, and it also plays a role in model monitoring.

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)**

5.Implementation Methodology





6.Technologies to be used

6.1.Software Platform

Python 3.6.8

Jupyter Notebook from Anaconda 3 (IDE)

Operating System (Windows 7,8,10,11)

6.2.Hardware Platform

RAM – Minimum 4GB

Hard Disk – Minimum 32GB

Processor i-5, AMD 3 upwards

6.3. Tools / Libraries

Pandas

Numpy

Scikit-Learn

Matplotlib

GridSearch CV

Lasso Regression

Linear Regression

Decision Tree Regressor

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.

8.Team details

Project Name & ID	Course Name	Student ID	Student Name	Role	Signature
TEXT SUMMARIZER USING DEEP LEARNING	Project Lab (IAI-851)	TCA1960001	Shivansh Sharma	Analysis.	
		TCA1960003	Anshika Gupta	Implementation	

9.Conclusion

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.

10.References

- https://www.analyticsvidhya.com/blog/2019/06/comprehensive-guide-text-summarization-using-deep-learning-python/
- https://data-flair.training/blogs/machine-learning-text-summarization/
- https://www.kdnuggets.com/2019/01/approaches-text-summarization-overview.html
- https://www.google.com/url?sa=i&url=https%3%2F%2Fwww.topcoder.com%2Fthrive%2Farticles%2Ftext-summarization in nlp&psig=AOvVaw2Bjnb9Ups2KLcLZEM5PtfT&ust=1677141229227000&source=images&cd=vfe&ved=0CBAQjRxqFwoTCPCoy4vcqP0CFQAAAAAdAAAAABAE
- https://www.kaggle.com/datasets/snap/amazon-fine-food-reviews