SUMMER INTERNSHIP II

Text Summarization using Natural Language Processing

Bachelor of Technology

Computer Science and Engineering

Submitted By

Karka Rohan

A70405221107

Drashya Sodha

A70405221225

Tanvish Renghe

A70405221111

Yeduru Jagadish Reddy

A70405221125

AMITY SCHOOL OF ENGINEERING AND TECHNOLOGY

AMITY UNIVERSITY MUMBAI

Year 2024-25

AMITY UNIVERSITY MUMBAI

Established vide Maharashtra Act No. 13 of 2014 of Government of Maharashtra and recognized under section 2(f) of UGC Act 1956

CERTIFICATE

This is to certify that Karka Rohan A70405221107, Drashya Sodha A70405221225, Tanvish Renghe A70405221111, Yeduru Jagadish Reddy A7040522112 of B. Tech Computer Science and Engineering semester VII has satisfactorily completed the Summer Internship II prescribed by Amity School of Engineering & Technology, Amity University Mumbai during academic year 2024-2025.

Stamp/seal of the Institute **Name and Signature of Examiner**

Acknowledgement

We express our sincere gratitude to our mentor/guide **Dr. Rajiv Iyer and Dr. Jeyavel J.** for their able guidance, continuous support and cooperation throughout our project, without which the present work would not have been possible.

We are grateful to all of those with whom we have had the pleasure to work on this and other related projects. We would especially like to thank **Dr. Rajiv Iyer and Dr. Jeyavel J.,** as our teachers and mentor; they have taught us more than we could ever give them the credit for here. They have shown us, by giving examples, what a good mentor (and person) should be.

Abstract

Automatic text summarization is basically summarizing of the given paragraph using natural language processing and machine learning. There has been an explosion in the amount of text data from a variety of sources. This volume of text is an invaluable source of information and knowledge which needs to be effectively summarized to be useful. In this review, the main approaches to automatic text summarization are described. We review the different processes for summarization and describe the effectiveness and shortcomings of the different methods. Two types will be used i.e.-extractive approach and abstractive approach. The basic idea behind summarization is finding the subset of the data which contains the information of all the set. There is a great need to reduce unnecessary data. It is very difficult to summarize the document manually so there is the great need for automatic methods. Approaches have been proposed inspired by the application of deep learning methods for automatic machine translation, specifically by framing the problem of text summarization as a sequence-to-sequence learning problem.

In this era everything is digitalized we can find a large amount of digital data for different purposes on the internet and relatively it's very hard to summarize this data manually. Automatic Text Summarization (ATS) is the subsequent big one that could simply summarize the source data and give us a short version that could preserve the content and the overall meaning. While the concept of ATS started long back in 1950's, this field is still struggling to give the best and efficient summaries. ATS proceeds towards 2 methods, Extractive and Abstractive Summarization. The Extractive and Abstractive methods had a process to improve text summarization technique. Text Summarization is implemented with NLP due to packages and methods in Python. Different approaches are present for summarizing the text and having few algorithms with which we can implement it. Text Rank is what to extractive text summarization and it is an unsupervised learning. Text Rank algorithm also uses undirected graphs, weighted graphs. keyword extraction, sentence extraction. So, in this paper, a model is made to get better result in text summarization with Genism library in NLP. This method improves the overall meaning of the phrase and the person reading it can understand it in a better way.

Keywords: Text summarization, NLP, Extractive, Abstractive.

List of Figures

Figure No	Page No.
Figure 1	8

Table of Contents

Sr.No	Contents	Page No.
	Acknowledgments	iii
	Abstract	iv
	List of figures	v
1.	Introduction	1
	Purpose of selecting a company/workplace	9
	Industry / Company overview	10
	Description of the project	11
	Salient contributions of the project	12
	Scope of assigned work	13
2.	Review of Literature/ Theoretical Background	14
3	Description of work	17
4.	Results	18
5.	Conclusions	19
6.	References	20

1. Introduction

Text summarization is a key application of Natural Language Processing (NLP) that focuses on automatically generating concise summaries from large volumes of text while retaining essential information. With the growing influx of data—such as news articles, research papers, legal documents, and social media posts—summarization techniques enable efficient consumption of information, saving time and effort for users.

There are two primary types of summarization techniques: extractive and abstractive. Extractive summarization identifies and selects the most relevant sentences or phrases from the original text, stitching them together to create a summary. On the other hand, abstractive summarization goes beyond simply extracting sentences; it generates new sentences by paraphrasing or rephrasing the content to express the core idea more succinctly, mimicking human summarization.

Recent advances in NLP, driven by machine learning and deep learning models, have significantly improved the performance of text summarization systems. Algorithms such as transformers (e.g., BERT, GPT, and T5) allow for high-quality abstractive summaries by capturing deeper contextual meaning. Moreover, hybrid approaches that combine extractive and abstractive methods are gaining popularity for achieving better results.

Text summarization has numerous practical applications, including automatic news aggregation, email summarization, legal document analysis, and academic paper highlights. As the need for digesting information efficiently continues to grow, NLP-powered summarization tools are becoming increasingly integral across industries.

The whole idea of automatic text summarization is to collect the necessary and crisp points from a large amount of data. There is a lot of information that is available on the internet, and it also keeps growing every day and having to collect the main data from it becomes hard since it takes a lot of time. The use of automatic text summarization makes it easier for the users to collect important data from huge amounts of information. Some of the graph base ranking algorithms are Text Rank [1], Hyperlinked Induced Topic Search [2], Positional Power Function [3] and so on. In this paper we are going to implement the Text Rank algorithm. Noting down the important points manually from large amount of data can be a very stressful job. So, automatic text summarization takes out the crucial words and sends them back in a way that the readers find easy. This

automatic text summarization is a small piece of NLP which cuts down the information and sends to the readers. It also arranges the information and sends back the sentences that are useful to create a crisp summary.

The words that occur the most no of times are considered the most worth. The topmost words are also arranged and then a summary is created. The Extractive methodology chooses the principal significant lines from the information text and utilizes them to think of the outline. The abstractive methodology addresses the information text in a type then generates the outline with the desired output of words and sentences that disagree from the first text sentences. Extractive systems extract vital text units (such as sentences, paragraphs etc.) within the input document. The theoretic approach is practically identical to the way that human summarizers 1st perceives the most ideas of a document, so generate new sentences that aren't seen within the original document. The general design of an associate ATS system has the subsequent tasks: Pre-processing, Processing, Post-Processing. Text summarization is in this field as a conclusion that monitors are needed to grasp what humans have composed and generated human-readable outputs. Human language technology also can be viewed as a study of computer science (AI). Therefore, several existing AI algorithms and strategies, as well as neural network models, are used for finding human language technology connected issues. With the present analysis, researchers typically believe 2 kinds of approaches for text summarization as shown in Figure 1 extractive summarization and abstractive summarization.

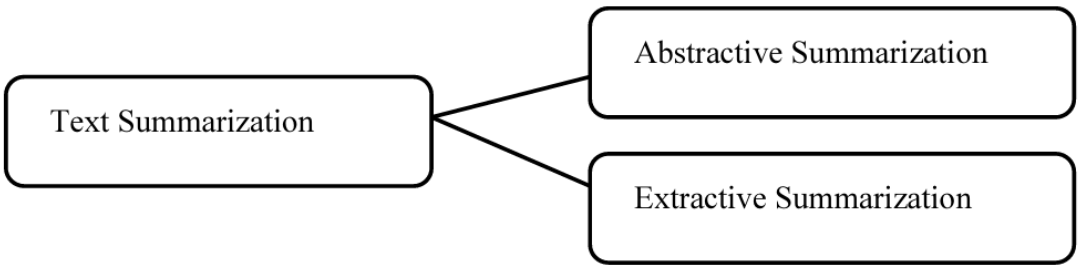


Figure 1. Types of Summarizations

1.1 Purpose of selecting a Project

Text summarization using NLP is a powerful tool aimed at solving the problem of information overload in today's digital world. With the exponential growth of online content, such as news articles, research papers, emails, and social media posts, it has become difficult for individuals and organizations to process large volumes of text efficiently. Summarization helps by extracting or generating concise versions of lengthy content, allowing users to grasp the key points quickly and make informed decisions without reading the entire text. This improves productivity and enhances user experience, especially in environments where time and attention are limited.

From a research perspective, the project provides an opportunity to explore key NLP concepts such as Natural Language Understanding (NLU), sentence ranking, semantic similarity, and advanced deep learning models like transformers. It also offers hands-on experience with pre-trained language models, such as BERT, GPT, and T5, which are widely used for building state-of-the-art summarization systems. Implementing text summarization involves working on multiple NLP tasks, including tokenization, sentence segmentation, keyword extraction, and sentiment analysis, making it a well-rounded academic project.

The real-world applications of text summarization span across multiple domains. For example, news aggregators provide concise snippets of articles, helping users stay updated in minimal time. Similarly, legal and financial sectors benefit from summarizing long reports or contracts, making critical information more accessible. Summarization is also integrated into customer service chatbots to provide brief answers from large knowledge bases. These applications demonstrate how summarization improves accessibility and ensures that users can focus on essential information without cognitive overload.

Overall, selecting text summarization using NLP as a project offers not only practical value but also technical depth. It bridges the gap between research and industry by tackling a real-world problem with cutting-edge technology, making it a highly relevant and rewarding project for students and professionals alike.

1.2 Project Overview

The Text Summarization using NLP project focuses on developing a system that can automatically generate concise and coherent summaries from long pieces of text. Summarization can be broadly classified into two types: extractive summarization and abstractive summarization. Extractive methods select key sentences or phrases directly from the original text, while abstractive methods generate new sentences that capture the meaning of the input text, much like a human-written summary. This project explores these approaches to reduce the reading load and highlight critical information effectively.

The system utilizes various NLP techniques and models to process text and identify the most important elements. For extractive summarization, algorithms like Text Rank or sentence-ranking models can be employed. For abstractive summarization, advanced deep learning techniques such as transformers (e.g., BERT, GPT, or T5) are used to generate human-like summaries. The project pipeline typically involves several steps, including text preprocessing, tokenization, sentence embedding, and relevance scoring. This ensures that the generated summaries retain key information while maintaining coherence and grammatical accuracy.

The application areas for text summarization are extensive and include industries such as media, education, healthcare, and finance. News aggregators use it to generate brief headlines or summaries of trending stories. In academia, it can summarize research papers to help students and scholars quickly grasp key findings. Similarly, businesses use summarization for customer feedback analysis, while legal professionals benefit from tools that summarize lengthy contracts and documents. This versatility makes text summarization an essential tool for improving information accessibility across different sectors.

Overall, this project offers both technical and practical value by providing hands-on experience with NLP models and solving real-world problems. It showcases the power of AI to enhance productivity and decision-making, making it highly relevant for today's information-driven society. Students and professionals working on this project gain insights into the inner workings of NLP, from data preprocessing to model training and evaluation, preparing them for advanced applications in the field of artificial intelligence.

1.3 Description of the project

Text summarization using NLP involves developing algorithms to automatically generate concise and meaningful summaries from larger text inputs. The primary objective is to retain the most relevant information while reducing redundancy, making it easier for users to grasp key ideas without reading the entire document. This task plays a vital role in addressing information overload by efficiently condensing long articles, reports, or research papers into shorter, coherent summaries that capture the main points.

There are two main types of summarizations: extractive and abstractive. Extractive summarization selects key sentences or phrases directly from the input text based on their importance or relevance, often using algorithms like TextRank or machine learning models for sentence ranking. In contrast, abstractive summarization involves generating new sentences that convey the core meaning of the text, like how a human would summarize. Abstractive methods often employ deep learning models such as transformers (e.g., BERT, GPT, or T5) to understand the context and rephrase content fluently.

Text summarization has numerous applications across various domains, including news aggregation, legal and financial report analysis, and customer service chatbots. For instance, it can be used to summarize news articles for quick consumption, generate research paper overviews, or condense product reviews for e-commerce platforms. By making critical information accessible in a fraction of the time, NLP-based summarization enhances productivity and improves user experience in a wide range of fields.

1.4 Salient contributions of the project

One of the primary contributions of text summarization using NLP is its ability to enhance information accessibility. In an era of overwhelming data, it helps users quickly understand large volumes of text by distilling essential points into concise summaries. This significantly improves productivity in fields like journalism, research, and education, where staying updated or reviewing lengthy materials can be time-consuming. News aggregators, for example, use summarization to generate brief headlines and summaries, while academic tools assist students by summarizing long research articles.

Another key contribution is in improving decision-making processes across various industries. Financial analysts, legal professionals, and business executives benefit from automatic summarization systems that can condense reports, contracts, or meeting minutes into actionable insights. This ensures that critical information is not missed, even when time is limited. Additionally, organizations use NLP-based summarization tools to analyze customer feedback or social media posts, providing quick insights into customer sentiment and emerging trends.

Finally, NLP-based summarization contributes to the development of advanced AI systems and applications by pushing the boundaries of natural language understanding and generation. Research into abstractive summarization has led to significant improvements in deep learning models like transformers, which are also applied in machine translation, chatbots, and content generation. Moreover, the integration of summarization into virtual assistants and chat interfaces enhances user interaction by delivering quick, relevant responses, making the technology a critical enabler for future advancements in artificial intelligence.

1.5 Scope of assigned work

The scope of work for a text summarization project using NLP encompasses research, design, development, and evaluation of algorithms capable of generating concise summaries from large text inputs. The project begins with a thorough study of the two primary approaches: extractive summarization, which selects key sentences from the input text, and abstractive summarization, which generates new sentences that convey the essence of the original content. This phase involves exploring algorithms like TextRank for extractive methods and deep learning-based models such as BERT, GPT, or T5 for abstractive summarization.

Another essential part of the work involves data preprocessing and feature engineering to ensure the input text is structured and cleaned for accurate processing. Tokenization, sentence segmentation, stop-word removal, and part-of-speech tagging are some preprocessing tasks necessary for both extractive and abstractive techniques. For deep learning-based approaches, the project may involve fine-tuning pre-trained models on domain-specific datasets to improve performance, such as summarizing legal documents or scientific articles. Additionally, model optimization techniques will be used to balance accuracy with computational efficiency.

The scope also includes evaluation of the summarization system using performance metrics such as ROUGE (Recall-Oriented Understudy for Gisting Evaluation) scores, BLEU (Bilingual Evaluation Understudy) scores, and human assessment for coherence and readability. Once the system is developed, the project may explore its deployment in real-world applications, such as integrating it into news aggregators, chatbots, or mobile apps. The final deliverables typically include a working prototype, documentation of algorithms and datasets used, and an analysis of results, ensuring the project aligns with both academic and practical objectives.

2. Theoretical Background

Natural Language Processing (NLP) is the intersection of Computer Science, Linguistics and Machine Learning that is involved with the interaction between computers and humans in natural language.

NLP is a way toward empowering PCs to comprehend and deliver human dialect. Uses of NLP systems are utilized in separating text, machine interpretation and Voice Agents like Alexa and Siri. NLP is one of the fields that profited from the advanced methodologies in Machine Adapting, particularly from Profound Learning strategies.

Regular Dialect Preparing method utilizes the characteristic dialect toolbox for making the principle arrange in python tasks to work with human dialect data. This is simpler to-use by giving the interfaces to at least one than 40 corpora and dictionary resources, for portrayal, for part passages sentences and to get the words in its unique frame Marking, parsing, and glossary thinking for current reasoning quality basic dialect dealing with libraries, and for dynamic discourse. The NLTK will utilize a colossal instrument area and will provide some help for individuals with the whole basic dialect taking care of system. This will assist individuals with "part sentences from sections to part up words, seeing the syntactic segments of those words, denoting the fundamental subjects, doing this it serves to your machine by acknowledging the main thing to the substance.

The paper by G. Vijay Kumar and colleagues discusses the development and challenges of Automatic Text Summarization (ATS) using Natural Language Processing (NLP). With the exponential growth of digital information, manual summarization has become infeasible. ATS is presented as a solution to efficiently condense information while maintaining key points and meaning. Historical Context and Current Trends in ATS.

The paper acknowledges that the idea of ATS dates to the 1950s, but despite technological advances, summarization remains a complex task. Traditional systems have struggled to balance brevity with coherence and relevance. The authors emphasize the two main summarization techniques:

1. **Extractive-Summarization:**
   This approach extracts key sentences or phrases from the original text. Algorithms like TextRank, inspired by Google's PageRank, generate summaries by modelling the text as a graph and identifying the most connected or influential sentences. TextRank works by creating undirected, weighted graphs to represent sentence relationships, performing keyword and sentence extraction to produce summaries.

2. **Abstractive-Summarization:**
   This more advanced technique generates new sentences that express the core ideas from the original text, similar to human summarization. However, the paper acknowledges the increased complexity of generating grammatically correct and meaningful summaries compared to extractive approaches.

   Technological Implementation and Methodology:

   The research leverages Python NLP libraries, particularly Gensim, for implementing extractive summarization. Gensim offers powerful tools for natural language processing, including TextRank-based summarization. The authors highlight the unsupervised nature of TextRank, which does not require labeled datasets, making it suitable for processing diverse datasets.

   The paper also emphasizes how NLP preprocessing techniques, such as tokenization, stop-word removal, and stemming, are essential to improve summarization accuracy. These steps ensure the algorithm focuses on the most relevant parts of the text while discarding redundant or irrelevant information.

   The literature review in this paper by Prudhvi, K., Bharath Chowdary, A., Subba Rami Reddy, P., and Lakshmi Prasanna, P. likely examines the evolution of text summarization techniques within the field of Natural Language Processing (NLP), placing particular emphasis on the extractive and abstractive approaches. Below is an expected outline of the main themes that might be covered:

   1. Extractive Text Summarization

   The authors probably highlight the popularity of extractive summarization techniques, where important sentences or phrases are identified and extracted verbatim from the text.

- Algorithms: Common methods like TextRank (inspired by PageRank), LexRank, and TF-IDF (Term Frequency-Inverse Document Frequency) are likely discussed.
- Challenges: The literature might underline challenges such as disjointed summaries, where extracted sentences lack coherence since they are not rephrased.

2. Abstractive Text Summarization

The paper likely covers advancements in abstractive summarization, where new sentences are generated to capture the meaning of the original content.

- Neural Network Models: The review might discuss the increasing use of deep learning architectures like Seq2Seq models (Sequence-to-Sequence) and transformers (e.g., BERT and GPT) to generate abstractive summaries.
- Limitations: While abstractive summarization models are more natural and flexible, they face challenges such as maintaining grammatical correctness and accurately representing the source text.

3. Description of work

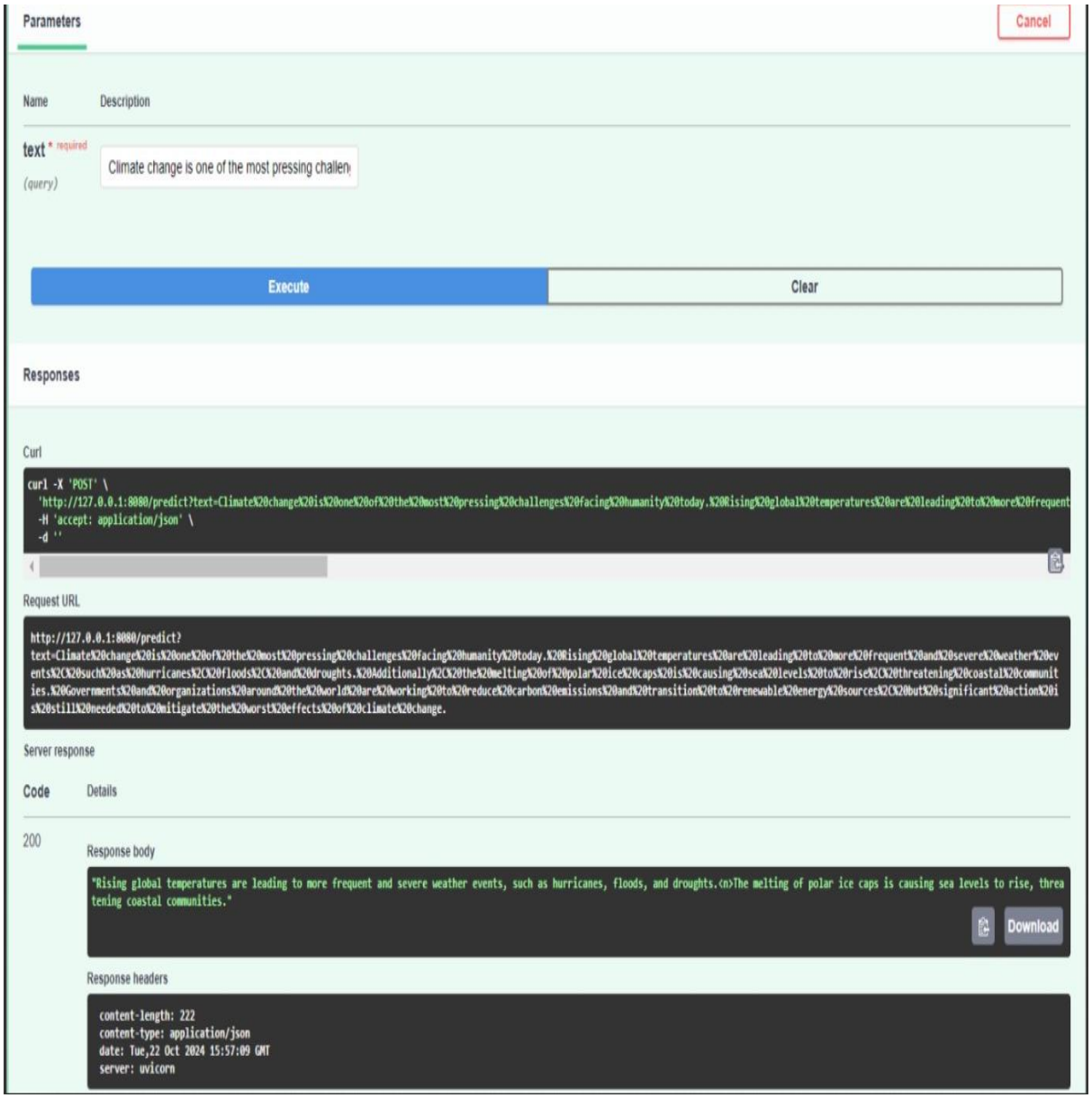
The process of building a text summarization system using Natural Language Processing **(**NLP**)** involves several stages, starting with pre-processing of the input text. Tokenization, stop word removal, and sentence segmentation help in cleaning the data, allowing the model to focus on meaningful patterns. Depending on the method—extractive or abstractive—the system identifies key information. Extractive summarization selects essential sentences directly from the text, using algorithms like TextRank or BERT embeddings, while abstractive summarization relies on advanced models such as T5 or GPT to generate new sentences that capture the text's core meaning.

Fast API, a modern web framework for building APIs in Python, plays a crucial role in deploying the NLP-based summarization system as a service. FastAPI allows developers to create RESTful endpoints for the summarization model with minimal overhead. For instance, the workflow may involve uploading text data via an API endpoint, which the FastAPI backend processes by passing it through the summarization model. The API then returns a summarized version in real-time, making the solution accessible to users via web applications or other platforms.

One advantage of using Fast API is its asynchronous capabilities, which enable handling multiple requests efficiently. This is particularly beneficial when dealing with large text data or multiple concurrent users, ensuring low latency. Fast API also integrates well with pre-trained NLP models via libraries like Hugging Face Transformers or SpaCy, making it easier to load and use these models without extensive development effort. Additionally, caching mechanisms such as Redis or Fast Api's dependency injection system can optimize response times for frequently requested summaries.

This combination of NLP and FastAPI opens various practical applications, including automated news summarization, legal document processing, and customer service email summarization. By exposing the summarization model as a fast, scalable API, businesses can integrate it into their workflows, providing efficient information retrieval to end users.

4. Results


Parameters
Cancel
Name
Description
text * required
(query)
Climate change is one of the most pressing challen
Execute
Clear
Responses
Curl
curl -X 'POST' \
'http://127.0.0.1:8080/predict?text=Climate%20change%20is%20one%20of%20the%20most%20pressing%20challenges%20facing%20humanity%20today.%20Rising%20global%20temperatures%20are%20leading%20to%20more%20frequent
-H 'accept: application/json' \
-d ''
Request URL
http://127.0.0.1:8080/predict?
text=Climate%20change%20is%20one%20of%20the%20most%20pressing%20challenges%20facing%20humanity%20today.%20Rising%20global%20temperatures%20are%20leading%20to%20more%20frequent%20and%20severe%20weather%20events%2C%20such%20as%20hurricanes%2C%20floods%2C%20and%20droughts.%20Additionally%2C%20the%20melting%20of%20polar%20ice%20caps%20is%20causing%20sea%20levels%20to%20rise%2C%20threatening%20coastal%20communities.%20Governments%20and%20organizations%20around%20the%20world%20are%20working%20to%20reduce%20carbon%20emissions%20and%20transition%20to%20renewable%20energy%20sources%2C%20but%20significant%20action%20is%20still%20needed%20to%20mitigate%20the%20worst%20effects%20of%20climate%20change.
Server response
Code
Details
200
Response body
"Rising global temperatures are leading to more frequent and severe weather events, such as hurricanes, floods, and droughts.<n>The melting of polar ice caps is causing sea levels to rise, threatening coastal communities."
Download
Response headers
content-length: 222
content-type: application/json
date: Tue,22 Oct 2024 15:57:09 GMT
server: uvicorn

Conclusion

As with time internet is growing at a very fast rate and with-it data and information is also increasing. it will be going to be difficult for human to summarize large amount of data. Thus, there is a need of automatic text summarization because of this huge amount of data. Until now, we have read multiple papers regarding text summarization, natural language processing and lesk algorithms. There are multiple automatic text summarizers with great capabilities and giving good results. 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 an extractive approach to do text summarization. We have successfully implemented a state-of-the-art model for abstractive sentence summarization to recurrent neural network architecture.

The model is a simplified version of the encoder-decoder framework for machine translation. The model is trained on the Amazon-fine-food-review corpus to generate summaries of reviews based on the first line of each review. There are a few limitations of the model which can be improved in further work. First limitation is that it sometimes generates repeated words in the summary, the other problem is it takes too much time to generate a summary if the input text size is large enough, the other issue is that for large text input it sometimes misinterprets the context and generates exactly opposite context summary.

References

[1] Kumar, G. V., Yadav, A., Vishnupriya, B., Lahari, M. N., Smriti, J., & Reddy, D. S. (2021). *Text Summarizing Using NLP*. In *Recent Trends in Intensive Computing* (pp. 60–67). IOS Press. https://doi.org/10.3233/APC210179

[2] Prudhvi, K., Bharath Chowdary, A., Subba Rami Reddy, P., Lakshmi Prasanna, P. (2021). Text Summarization Using Natural Language Processing. In: Satapathy, S., Bhateja, V., Janakiramaiah, B., Chen, YW. (eds) Intelligent System Design. Advances in Intelligent Systems and Computing, vol 1171. Springer, Singapore. https://doi.org/10.1007/978-981-15-5400-1_54

[3] B. N, D. Kumari, B. N, M. N, S. K. P and S. R. A, "Text Summarization using NLP Technique," *2022 International Conference on Distributed Computing, VLSI, Electrical Circuits and Robotics (DISCOVER)*, Shivamogga, India, 2022, pp. 30-35, doi: 10.1109/DISCOVER55800.2022.9974823.