SAVEETHA
INSTITUTE OF MEDICAL AND TECHNICAL SCIENCES
(Declared as Deemed to be University under Section 3 of UGC Act 1956)
Accredited by NAAC A Grade

AUTOMATED TEXT SUMMARIZATION USING NATURAL LANGUAGE PROCESSING

CAPSTONE PROJECT REPORT

DSA0415- FUNDAMENTALS OF DATA SCIENCE FOR BUSINESS DECISION MAKING

Submitted by

M.Gopinadh [192224153]

Department of Artificial Intelligence and Data Science

Guided by

MANGAIYARKARASI K

Course Faculty

Department of Computer Science Engineering

Saveetha School of Engineering

BONAFIDE CERTIFICATE

This is to certify that the project report entitled “TITLE” submitted by M.Gopinadh (192224153), Charan Simha (192124153) to Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Chennai, is a record of bonafide work carried out by him/her under my guidance. The project fulfils the requirements as per the regulations of this institution and in my appraisal meets the required standards for submission.

K. Mangaiyarkarasi
(Course Faculty)
Department of Deep Learning
Saveetha School of Engineering
SIMATS, Chennai – 602 105

ACKNOWLEDGEMENT

This project work would not have been possible without the contribution of many people. It gives me immense pleasure to express my profound gratitude to our Honorable Chancellor **Dr. N. M. Veeraiyan**, Saveetha Institute of Medical and Technical Sciences, for his blessings and for being a source of inspiration. I sincerely thank our Director of Academics **Dr. Deepak Nallaswamy,** SIMATS, for his visionary thoughts and support. I am indebted to extend my gratitude to our Director **Dr. Ramya Deepak,** Saveetha School of Engineering, for facilitating us all the facilities and extended support to gain valuable education and learning experience.

I register my special thanks to **Dr. B. Ramesh,** Principal, Saveetha School of Engineering for the support given to me in the successful conduct of this project. I wish to express my sincere gratitude to my Course faculty **K. Mangaiyarkarasi**, for his inspiring guidance, personal involvement and constant encouragement during the entire course of this work.

I am grateful to Project Coordinators, Review Panel External and Internal Members and the entire faculty of the Department of Design, for their constructive criticisms and valuable suggestions which have been a rich source to improve the quality of this work.

M.Gopinadh.

TABLE OF CONTENTS

INTRODUTION:

Automated text summarization leverages natural language processing (NLP) to condense large volumes of text into shorter, concise summaries while preserving essential information. This enhances information retrieval by allowing users to quickly grasp the core ideas of extensive texts without reading them in full. There are two main types of summarization: extractive and abstractive. Extractive summarization selects important sentences or phrases directly from the original text, combining them to form a summary. Abstractive summarization, on the other hand, generates new sentences that convey the same information as the original text, requiring a deeper understanding and rephrasing of the content.

a. Automated Text Summarization: Enhancing Information Retrieval
b. Utilizes natural language processing to condense(reduce) large text volumes into concise(short and clear) summaries.
c. Extracts essential sentences and phrases for brief overviews.
d. Potentially improves information retrieval and productivity.
e. Offers time-saving benefits for individuals and organizations dealing with large textual data.

Natural Language Processing (NLP) :

a. A branch of artificial intelligence enabling computers to understand, interpret, and generate meaningful human language.
b. Techniques range from basic text processing to advanced machine learning models.
c. Goal: bridge human communication and computer understanding for seamless interactions.

Addresses challenges like ambiguity, context, and semantics.

Overall, automated text summarization is a powerful tool that significantly enhances information retrieval by providing concise, relevant summaries of extensive texts. Its application across various domains underscores its value in today's information-rich world.

PROBLEM STATEMENT:

In today's information-rich world, individuals and organizations are inundated with vast amounts of text data from various sources such as news articles, research papers, legal documents, customer support logs, and medical records. Efficiently extracting and understanding key information from these extensive texts is a significant challenge. Traditional methods of manual summarization are time-consuming, labor-intensive, and prone to human error.

Automated text summarization aims to address these challenges by utilizing natural language processing (NLP) techniques to create concise, relevant summaries of large texts. However, achieving accurate and coherent automated summarization poses several difficulties:

1. **Maintaining Coherence and Relevance**: Ensuring that the generated summaries are not only brief but also coherent and contextually relevant, preserving the essential information of the original text.
2. **Balancing Extractive and Abstractive Methods**: Combining the simplicity and precision of extractive summarization with the depth and fluency of abstractive summarization to produce high-quality summaries.
3. **Handling Diverse Text Types and Domains**: Developing models that can effectively summarize texts from various domains (e.g., news, legal, medical) and of different types (e.g., structured documents, conversational logs).
4. **Scalability and Efficiency**: Creating summarization systems that can process large volumes of text quickly and efficiently, suitable for real-time applications.

Addressing these issues requires advancements in NLP algorithms and models to enhance the performance of automated summarization systems. Effective solutions will significantly improve information retrieval, aiding users in quickly understanding large amounts of text data and making informed decisions.

DATASET ANALYSIS:

Analyzing datasets is crucial for developing effective automated text summarization models. This involves understanding the structure, content, and characteristics of the data, as well as evaluating the performance of summarization algorithms. Below is an outline for conducting a thorough dataset analysis for automated text summarization:

1. Data Collection

- **Source Identification**: Identify sources of text data relevant to the summarization task. Common sources include news articles, scientific papers, legal documents, customer support logs, and medical records.
- **Dataset Examples**: Popular datasets for summarization include CNN/Daily Mail, XSum, PubMed, ArXiv, DUC, and AMI Meeting Corpus.

2. Data Preprocessing

- **Text Cleaning**: Remove noise such as HTML tags, special characters, and irrelevant metadata.
- **Tokenization**: Split text into tokens (words, phrases, or sentences) to facilitate further processing.
- **Normalization**: Convert text to a consistent format (e.g., lowercase conversion, stemming, and lemmatization).

3. Exploratory Data Analysis (EDA)

- **Data Distribution**: Analyze the distribution of document lengths and summary lengths to understand variability and identify any outliers.
- **Word Frequency**: Calculate word frequencies to identify common terms and important keywords within the dataset.

- **Sentence Structure**: Examine the structure of sentences in the text to understand the complexity and style of writing.

ENVIROMENT SETUP:

Setting up a proper environment is crucial for developing and deploying automated text summarization models. This involves selecting the right hardware, installing necessary software and libraries, and configuring the development environment. Below is a comprehensive guide to setting up an environment for automated text summarization.

1. Hardware Requirements

- **Processor**: A multi-core CPU is essential for running data preprocessing and training tasks efficiently.
- **Memory**: At least 16 GB of RAM is recommended to handle large datasets and model training.
- **Storage**: Ensure sufficient storage (SSD preferred) for datasets, models, and intermediate files. A minimum of 100 GB is suggested.
- **GPU**: For training advanced neural network models, a CUDA-compatible GPU (such as NVIDIA) with at least 8 GB VRAM is highly recommended.

2. Software Requirements

Operating System

- **OS**: Linux (Ubuntu recommended) or macOS. Windows can also be used but might require additional setup for certain libraries.

Development Environment

- **Python**: Version 3.7 or higher. Python is the primary language used for NLP and machine learning tasks.

- **Jupyter Notebook**: Useful for interactive development and experimentation.

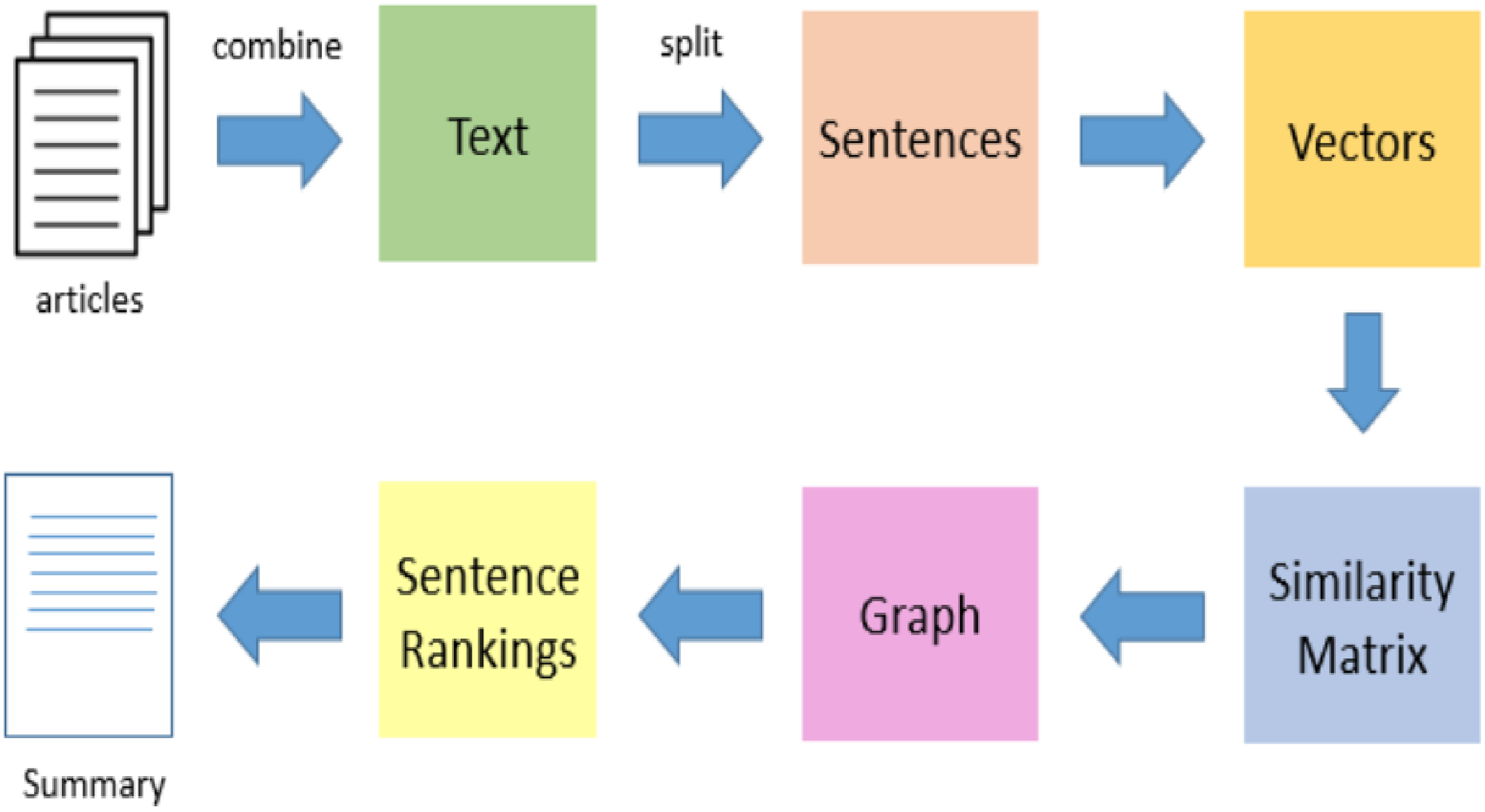
ARCHITECTURE DIAGRAM:

► Data Ingestion:

- **Sources**: Various text data sources such as news articles, research papers, legal documents, customer support logs, and medical records.
- **Ingestion Pipeline**: Mechanisms to collect and import data into the system.

► Data Preprocessing:

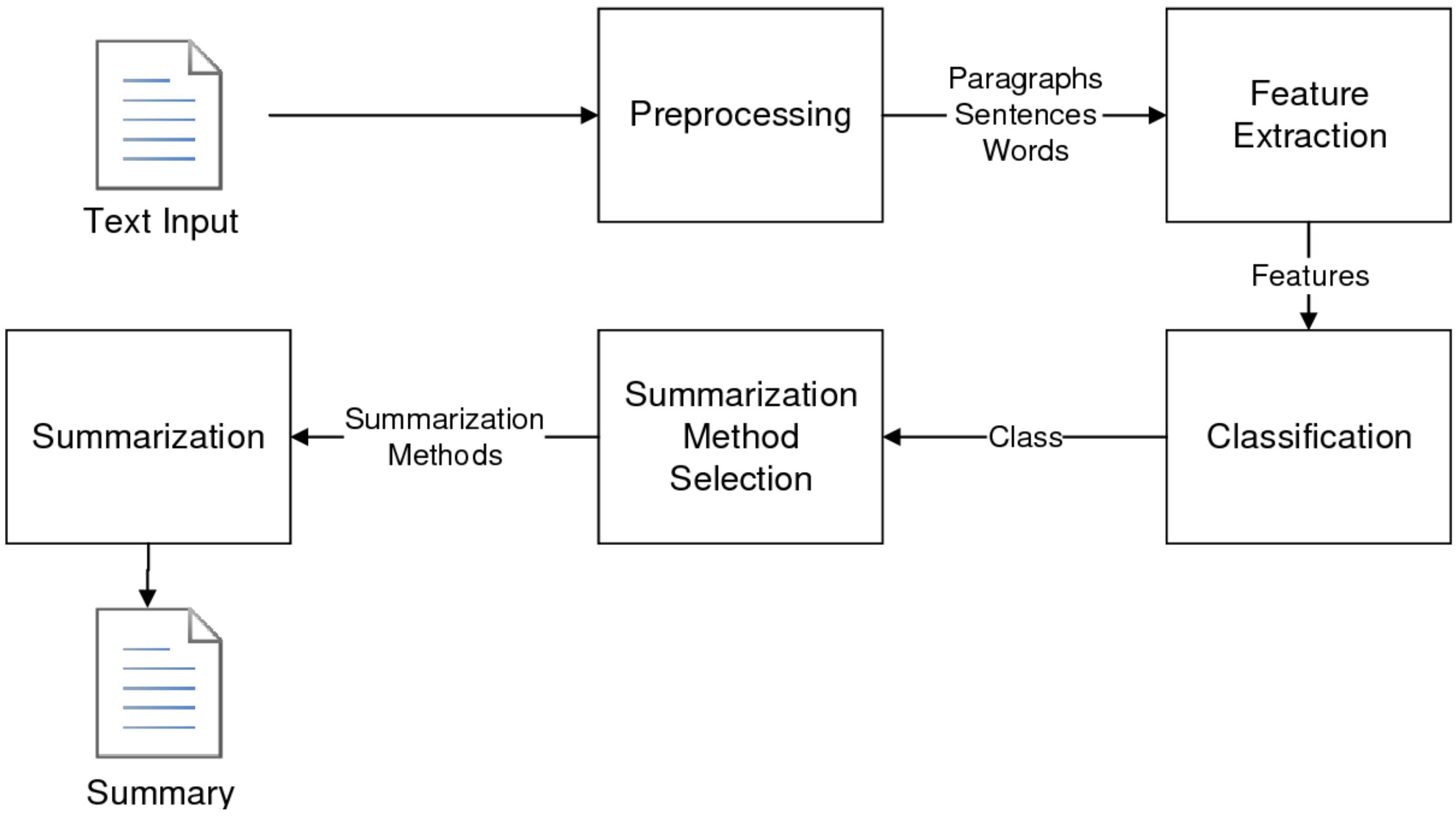
- **Text Cleaning**: Removing noise, special characters, and irrelevant metadata.
- **Tokenization**: Splitting text into tokens (words, sentences).
- **Normalization**: Lowercasing, stemming, lemmatization, and removing stop words.



CLASS DIAGRAM:

- **DataIngestion**: Handles collecting and importing text data from various sources.
- **Preprocessing**: Manages text cleaning, tokenization, and normalization tasks.
- **FeatureExtraction**: Extracts relevant features from the text data for use in summarization models.
- **SummarizationModel**: Represents a generic summarization model. It can be extended into specific models like ExtractiveModel and AbstractiveModel.
- **ModelTraining**: Responsible for training summarization models using labeled datasets.
- **Evaluation**: Assesses the performance of the summarization models using various metrics.
- **SummarizationEngine**: Applies trained models to generate summaries from new texts and handles post-processing.
- **Database**: Stores raw text, processed data, summaries, and model outputs.

- **UserInterface**: Provides a web interface and API for user interaction.



CODE SKELETON:

```
from gensim.summarization import summarize

# Input text
text = """
Natural language processing (NLP) is a field of artificial intelligence that deals with the interaction between computers and humans using natural language. It aims to enable computers to understand, interpret, and generate human language in a way that is both meaningful and useful.

One of the key tasks in NLP is text summarization, which involves condensing a piece of text while retaining its key information. There are two main approaches to text summarization: extractive and abstractive summarization.
```

Extractive summarization involves selecting important sentences or phrases from the original text and arranging them to create a summary. Abstractive summarization, on the other hand, involves generating a summary that may contain new phrases or sentences not present in the original text.

Gensim is a popular Python library for NLP tasks, including text summarization. It provides easy-to-use functions for both extractive and abstractive summarization.

Let's use Gensim to generate a summary for this text.
"""

# Generate the summary
summary = summarize(text)

# Print the summary
print(summary)

RESULT:

INPUT:

The education system in India has undergone significant transformations over the years, yet it grapples with various challenges. With a vast and diverse population, India's education system caters to a wide spectrum of learners, from rural areas to urban centers. Despite efforts to enhance accessibility and quality, disparities persist, especially in terms of infrastructure, teacher quality, and learning outcomes between regions and socioeconomic groups. The government plays a pivotal role in formulating policies and implementing programs to address these issues. The introduction of initiatives like the Right to Education Act aimed to ensure free and compulsory education for all children aged 6 to 14, but its full implementation remains a challenge. Additionally, the emphasis on rote memorization and

standardized testing often overshadows critical thinking and creativity in the curriculum. However, India's education system also boasts prestigious institutions like the Indian Institutes of Technology (IITs) and Indian Institutes of Management (IIMs), which contribute significantly to research and innovation globally. Moving forward, efforts to bridge educational gaps, enhance pedagogical methods, and promote inclusive learning environments are crucial for fostering holistic development and meeting the diverse needs of India's burgeoning population.

OUTPUT:

However, India's education system also boasts prestigious institutions like the Indian Institutes of Technology (IITs) and Indian Institutes of Management (IIMs), which contribute significantly to research and innovation globally. With a vast and diverse population, India's education system caters to a wide spectrum of learners, from rural areas to urban centers. Despite efforts to enhance accessibility and quality, disparities persist, especially in terms of infrastructure, teacher quality, and learning outcomes between regions and socioeconomic groups. The education system in India has undergone significant transformations over the years, yet it grapples with various challenges. Additionally, the emphasis on rote memorization and standardized testing often overshadows critical thinking and creativity in the curriculum.

CONCLUSION:

- Automated text summary is a useful tool for reducing lengthy literary texts into succinct, educational summaries
- Uses natural language processing and algorithms to extract key information.
- Preserves original content while expediting information consumption.
- Useful in content curation, research, and news aggregation.
- Despite challenges like context sensitivity and coherence, it's becoming more accurate due to machine learning advancements.
- Essential for managing growing textual data in the digital age.

- Automated text summarization using NLP simplifies the process of extracting essential information from large volumes of text.
- This technology employs algorithms to identify key phrases, sentences, and concepts within a document.
- NLP-based summarization improves information retrieval efficiency and saves time for readers.
- It enables users to quickly grasp the main points of a document without having to read it entirely.
- Researchers can utilize automated summarization to sift through vast amounts of literature and extract relevant insights.
- Students benefit from NLP-based summaries for studying complex subjects and preparing for exams.
- News agencies can use this technology to generate concise summaries of breaking news stories for quick dissemination.
- Businesses leverage text summarization to analyze customer feedback, reviews, and market trends.
- Legal professionals use NLP-based summaries to extract key arguments and evidence from legal documents.
- Automated summarization enhances the accessibility of information for individuals with visual impairments.

REFERENCES:

- [1] Jaccard similarity and Jaccard distance in Python, https://pyshark.com/jaccard-similarity-and-jaccard-distance-in-python, Retrieved on 12/8/2021
- [2]A. Kogilavani, Dr.P.Balasubramani, “Clustering And Feature Specific Sentence Extraction Based Summarization of Multiple Documents”, International journal of computer science & information Technology, vol.2, no.4, Aug. 2010.

- [3] “Multi-document summarization”, Wikipedia, the free encyclopedia, 2012.
- [4] Abualigah, L., Bashabsheh, M.Q., Alabool, H., Shehab, M. (2020), Text Summarization: A Brief Review. In: Abd Elaziz, M., Al-qaness, M., Ewees, A., Dahou, A. (eds) Recent Advances in NLP: The Case of Arabic Language. Studies in Computational Intelligence, vol 874. Springer,
- [5]Nenkova, Ani, Kathleen McKeown "A survey of text summarization techniques." In Mining text data, pp. 43-76. Springer, Boston, MA, 2012.
- [6] Haque, Majharul, Suraiya Pervin, and Zerina Begum. "Literature review of automatic multiple documents text summarization." International Journal of Innovation and Applied Studies 3, no. 1 (2013), 121-129.
- [7] Widyassari, Adhika Pramita, Supriadi Rustad, Guruh Fajar Shidik, Edi Noersasongko, Abdul Syukur, and AffandyAffandy, "Review of automatic text summarization techniques & methods", Journal of King Saud University-Computer and Information Sciences (2020).
- [8]Allahyari, Mehdi, SeyedaminPouriyeh, Mehdi Assefi, Saeid Safaei, Elizabeth D. Trippe, Juan B. Gutierrez, and Krys Kochut, "Text summarization techniques: a brief survey." arXiv preprint arXiv:1707.02268 (2017).
- [9] Wang, Mengqian, Manhua Wang, Fei Yu, Yue Yang, Jennifer Walker and Javed Mostafa, "A systematic review of automatic text summarization for biomedical literature and EHRs", Journal of the American Medical Informatics Association 28, no. 10 (2021): 2287-2297.
- [10] Jani, D., Patel, N., Yadav, H., Suthar, S., Patel, S. (2022). A Concise Review on Automatic Text Summarization. In: Nayak, J., Behera, H., Naik, B., Vimal, S., Pelusi, D.