TEXT SUMMARIZER

Data Science Project

Introduction

In the digital age, the exponential growth of textual data necessitates efficient methods to condense information without losing its essence. Text summarization addresses this need by producing concise versions of longer documents.

This project focuses on two primary approaches:

- **Extractive summarization:** Selects and compiles existing sentences from the source text.
- **Abstractive summarization:** Generates novel sentences that encapsulate the main

Objectives

- To develop an efficient text summarization system that can generate concise summaries of long news articles while retaining the essential information and context.
- To implement both extractive and abstractive summarization techniques, enabling a comparative analysis of their effectiveness on real-world datasets like BBC News Summary.
- To visualize and analyze the dataset and summaries, including article length distribution, sentence counts, and summary compression ratio.

02

Dataset Description

Dataset link : https://www.kaggle.com/datasets/pariza/bbc-news-summary

It consists of 2,225 news articles collected from the BBC website. Each article is accompanied by a human-written summary

Categories

- Business
- Entertainment
- Politics
- Sport
- Tech

Key Features

- Title – The headline of the news article
- Article Text – Full body of the news content
- Category – Topical classification of the article
- Summary – Manually written concise summary for each article

Preprocessing Techniques Used

Sentence Segmentation

Splitting articles into individual sentences with careful punctuation handling

Tokenization

Tokenizing text using the Nltk tokenizer for abstractive summarization

Text Cleaning

- Lowercasing of the characters in the sentences
- Removal of newline characters and extra whitespace.

Abbrevations

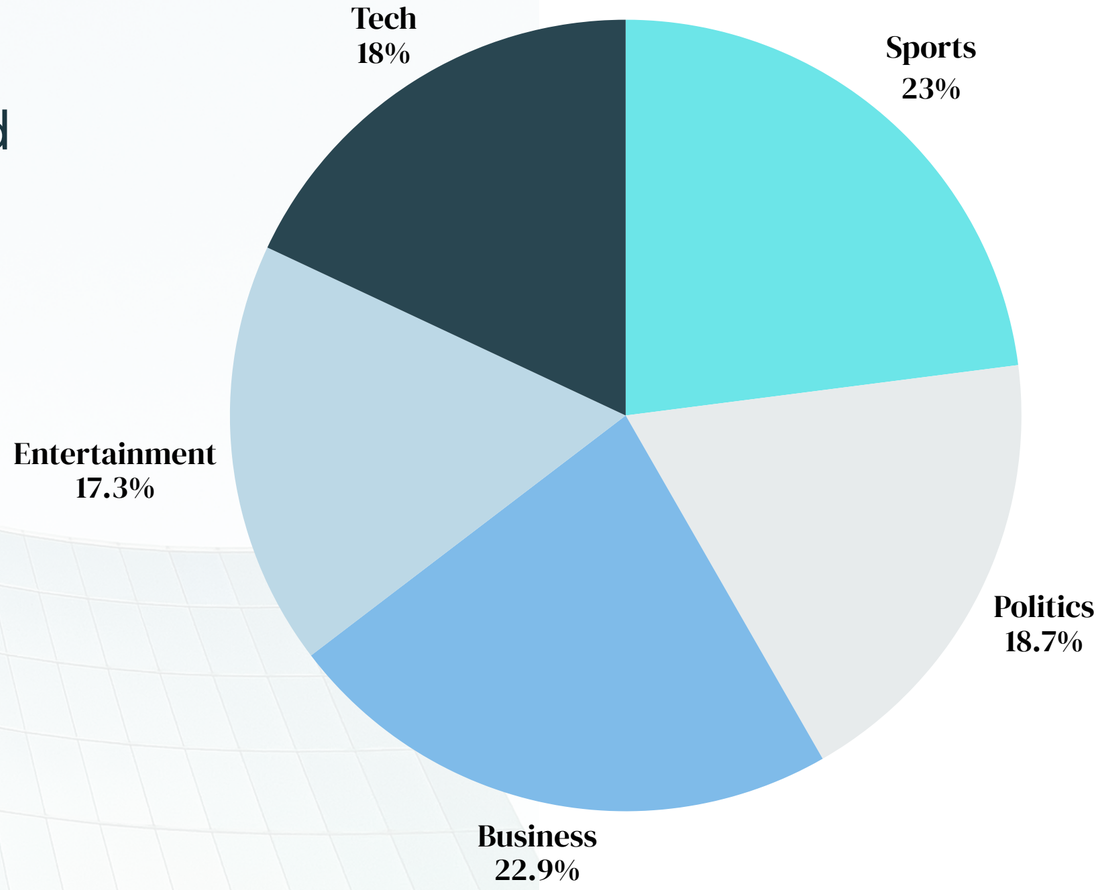
Standardization of abbreviations.

Exploratory Data Analysis (EDA)

Distribution of Articles by Category

The dataset consists of 2,225 articles distributed across five categories:

- Sport: 511 articles
- Politics: 417 articles
- Business: 510 articles
- Entertainment: 386 articles
- Tech: 401 articles



Article Length Statistics

- Mean article length: 384.04 words
- Median article length: 332 words
- Minimum article length: 89 words
- Maximum article length: 4,432 words

Summary Statistics

- Mean summary length: 165.17 words
- Median summary length: 142 words
- Minimum summary length: 38 words
- Maximum summary length: 2,073 words

Algorithms/Techniques/Model Used

Extractive Summarization

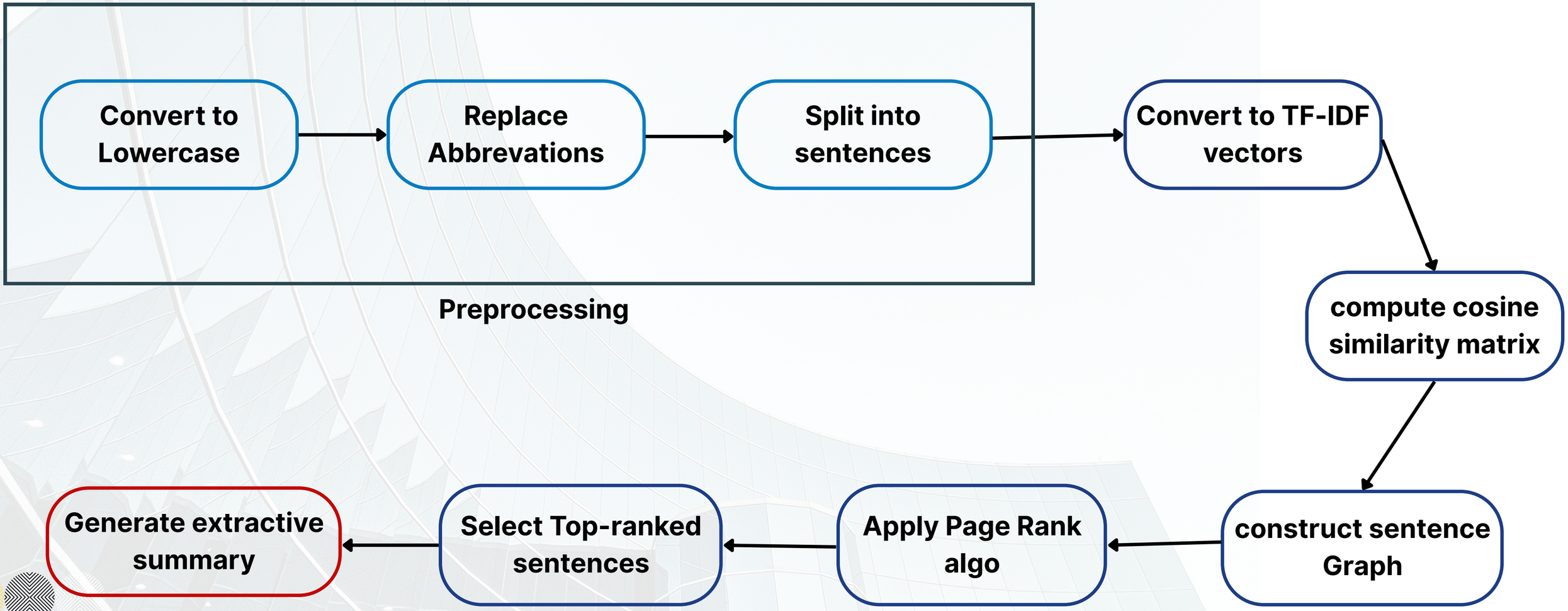
- **TF-IDF Vectorization:** Converts sentences into numerical vectors based on term importance.
- **Cosine Similarity**: Measures similarity between sentence pairs to form a similarity matrix.
- **Similarity Matrix Construction:** Builds a weighted matrix representing sentence connections.
- **Sentence Graph Formation:** Creates a graph where sentences are nodes and edges are weighted by similarity.
- **PageRank Algorithm:** Ranks sentences based on centrality, selecting the most important ones for the summary

Algorithms/Techniques/Model Used

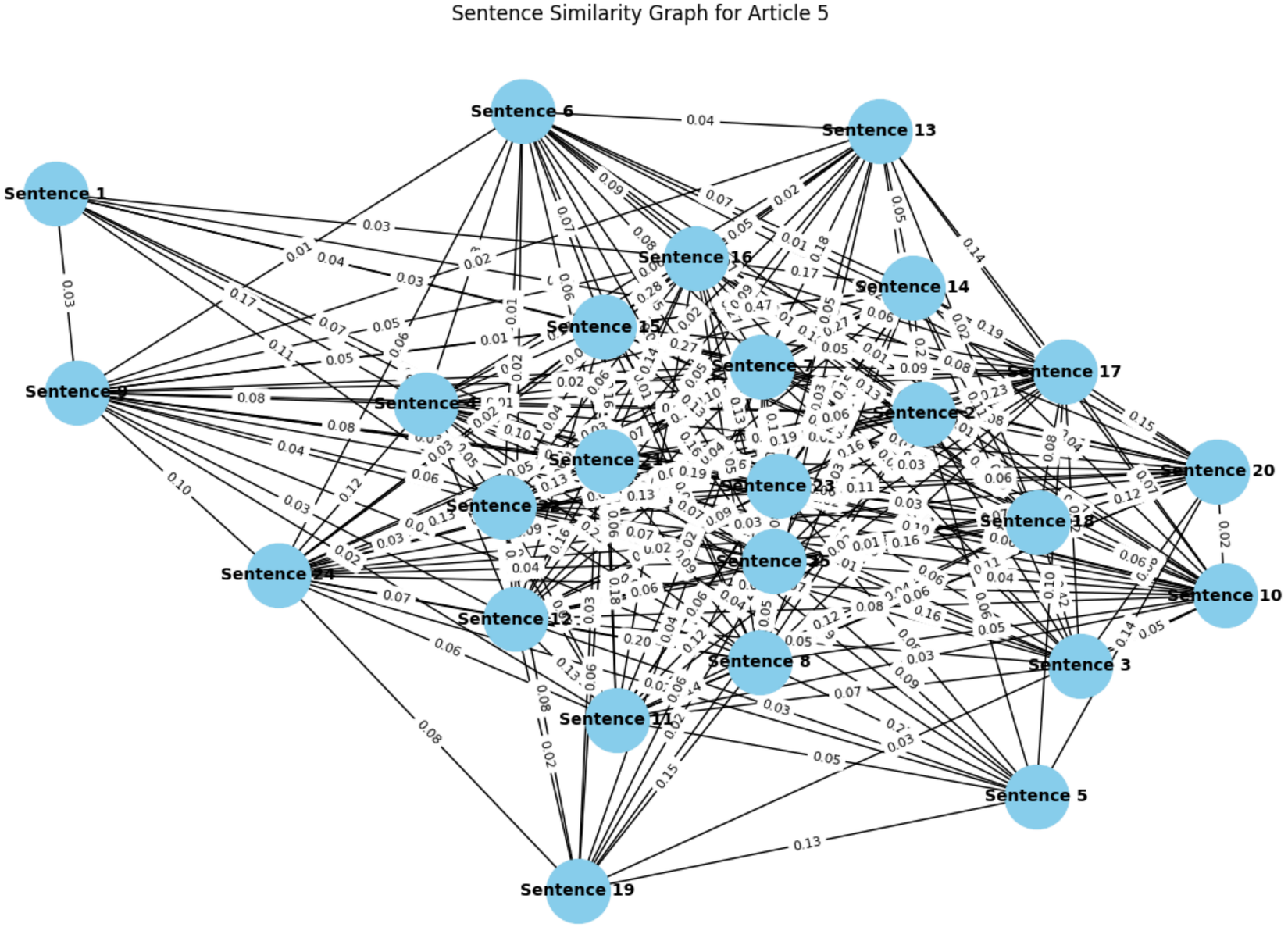
Abstractive Summarization

- **Model**: facebook/bart-base from Hugging Face Transformers
- **Type**: Pre-trained sequence-to-sequence transformer
- Architecture: Encoder-Decoder with attention mechanism
- **Fine-Tuning:** Performed on BBC News Summary dataset
- **Custom hyperparameters** (learning rate, batch size, epochs)
- **Tokenizer:** BART tokenizer with padding & truncation

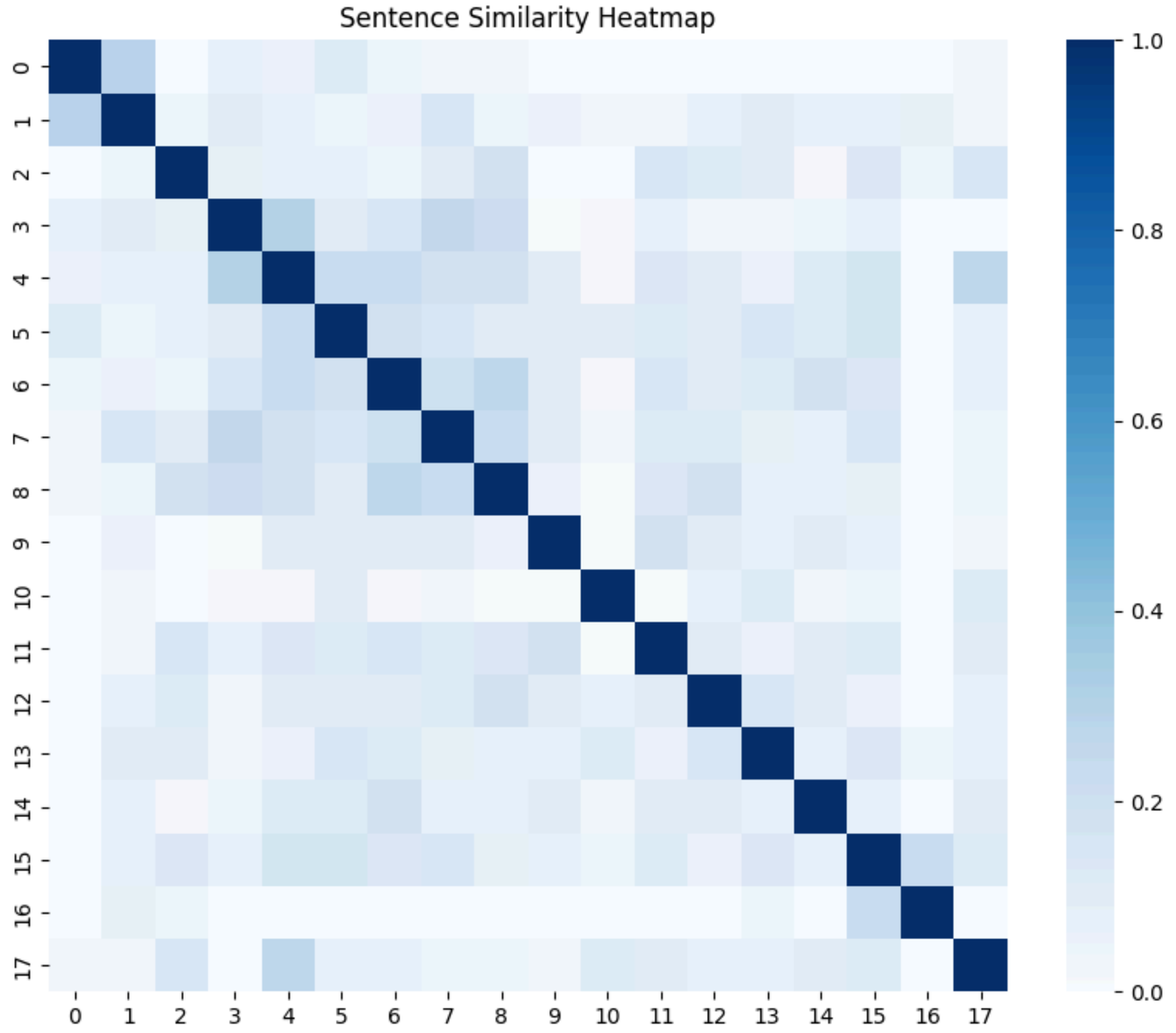
Extractive Summariser Flow



Sentence Graph



Heatmap for sentence Similarity



Evaluation Metrics

Used ROUGE (Recall-Oriented Understudy for Gisting Evaluation) metrics:

- ROUGE-1: Unigram (single word) overlap
- ROUGE-2: Bigram (two-word) overlap
- ROUGE-L: Longest Common Subsequence (LCS)

Models Comparison

Extractive Summarizer

Retain original sentences that best represent the main content.

Rouge 1 : 0.67
Rouge 2: 0.56
Rouge L : 0.4588

Abstractive Summarizer

Generates new sentences by understanding and rephrasing the content.

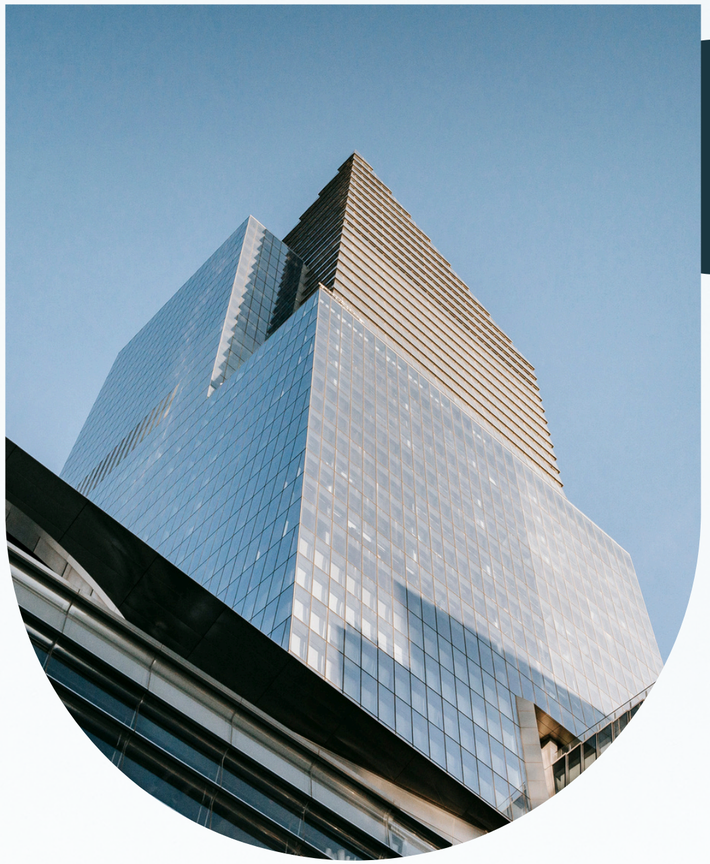
Rouge 1 : 0.64
Rouge 2: 0.55
Rouge L : 0.46

Future Aspects

- **YouTube Video Summarizer:** Integrating text from the transcript, visual elements, and audio cues to produce a more comprehensive summary.
- **Research Paper Summarizer:** A research paper summarizer that can read and process research papers, extracting key sections like the abstract, methodology, results, and conclusions, and then condensing these into a concise summary.
- **Medical Report Summarizer:** A medical report summarizer that can analyze clinical notes, lab results, and patient histories, summarizing them into concise, easy-to-understand reports.

Conclusion

- This project demonstrates the effectiveness of automated text summarization in condensing large volumes of information into brief, meaningful summaries.
- By implementing both extractive and abstractive techniques, we were able to compare traditional statistical methods with modern deep learning approaches for summarizing text.
- The summarizer can be a valuable tool in domains like news aggregation, research, education, and content curation, helping users save time and focus on key information



Presented By

- Sayali Khedkar 612203092
- Jay Kolhe 612203096
- Tejas Kolhe 612203097
- Varad Kothekar 612203099

THANK YOU