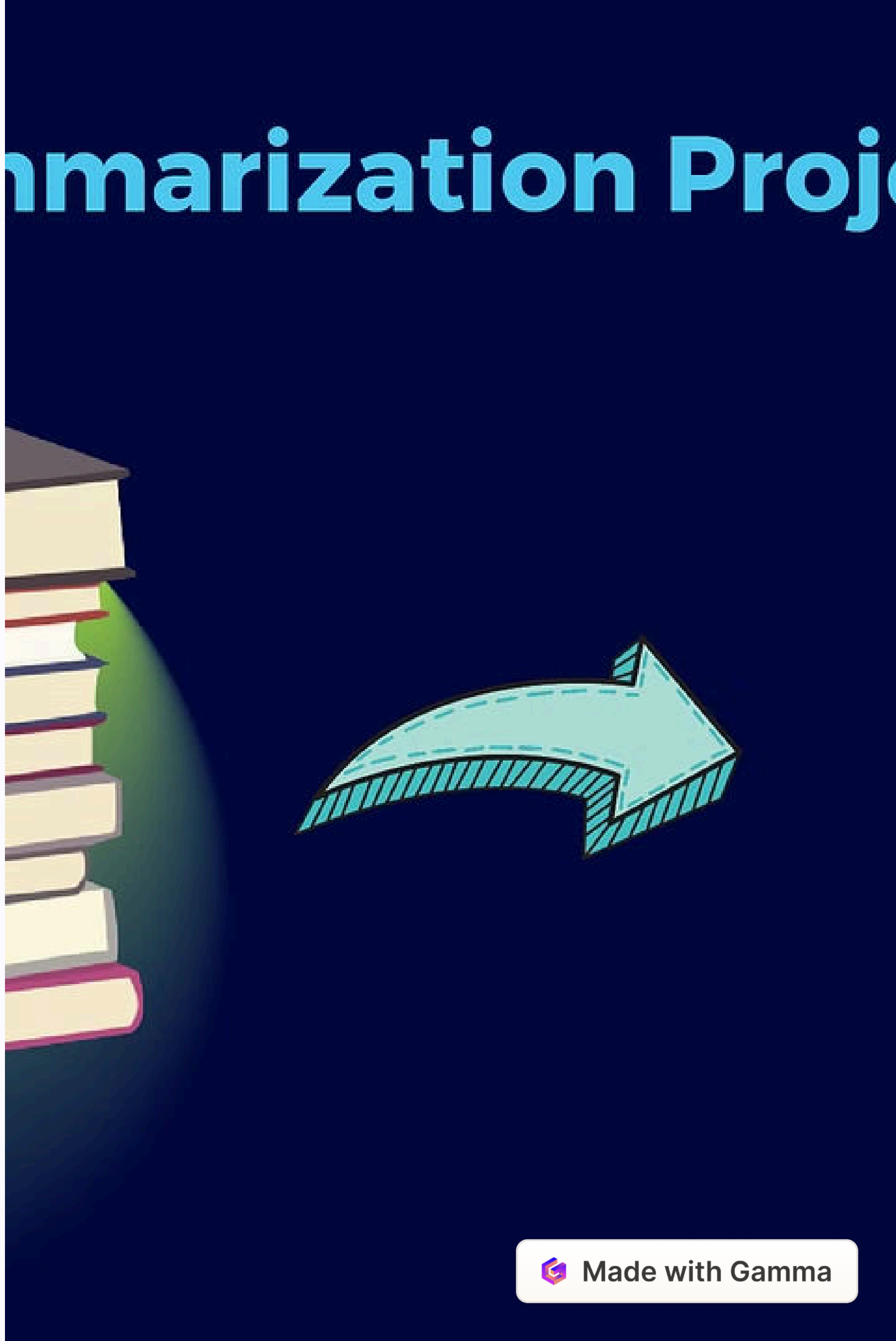
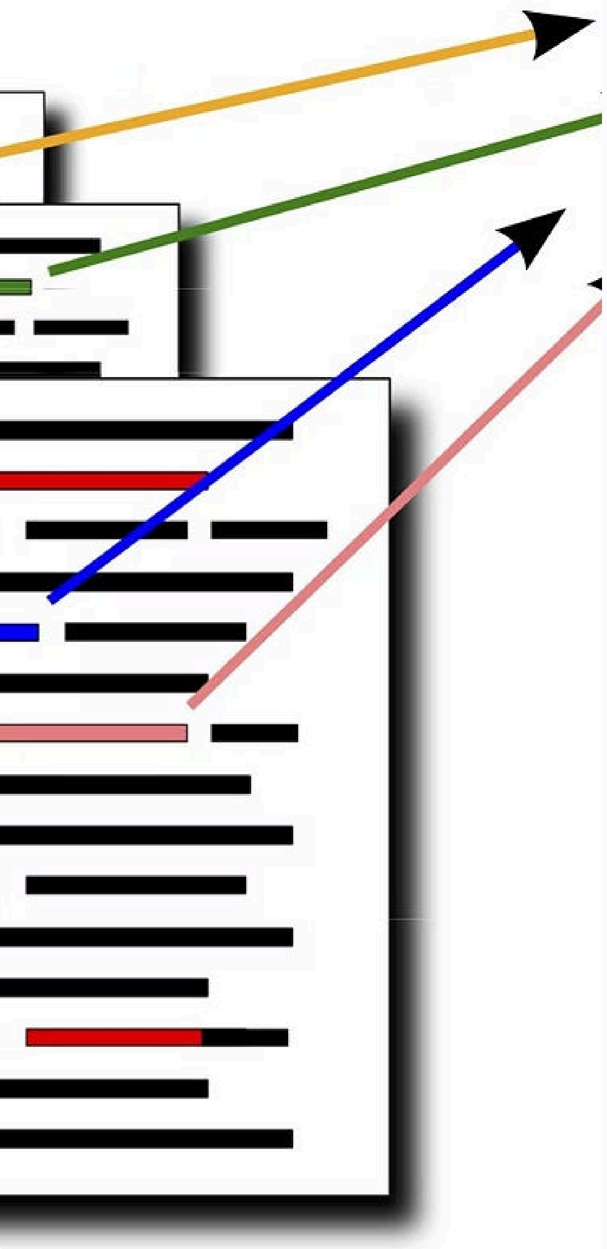
Text Summarization: Distilling the Essence





What is Text Summarization?

1 Condensing Information

Text summarization is the process of extracting the key points and essential information from a longer text, creating a concise summary.

2 Preserving Meaning

The goal is to capture the core ideas and maintain the overall meaning of the original text in a shortened form.

3 Improving Efficiency

Summarization helps readers quickly grasp the main points, making information more accessible and saving time.

The Importance of Text Summarization

Information Overload

In the digital age, we are bombarded with vast amounts of text. Summarization helps us navigate and process this information efficiently.

Decision Making

Summaries provide a concise overview, allowing decision-makers to quickly identify key insights and make informed choices.

Knowledge Retention

Summaries aid in retaining and understanding the most salient points, making the information more memorable.

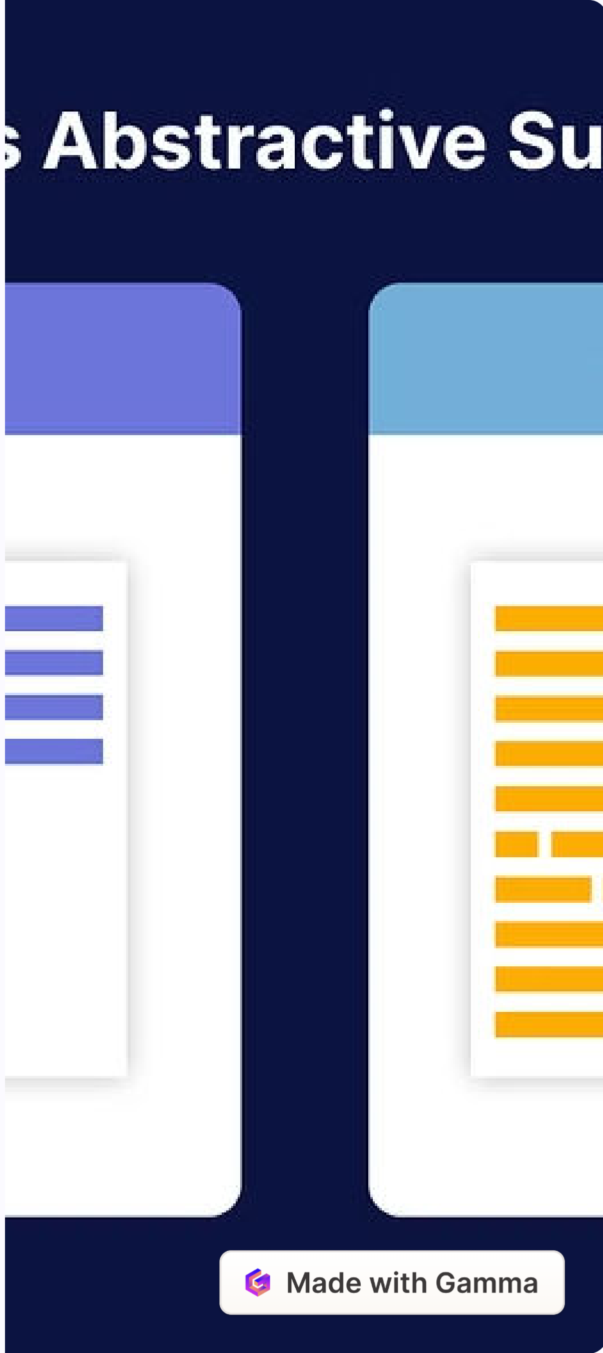
Extractive vs. Abstractive Summarization

1 Extractive

Selects and extracts the most important sentences or phrases from the original text to form the summary.

2 Abstractive

Generates new text that captures the key ideas, using advanced language processing to rephrase and combine information.



Software Requirements

- **Jupyter Notebook**: For developing and testing the summarization models
- **NLTK (Natural Language Toolkit)**: For tokenization, stop words, etc.
- **Transformers**: For pretrained models like BERT, T5, and Pegasus.
- **scikit-learn**: For LSA, TF-IDF, and other traditional machine learning method
- **NumPy**: Numerical operations
- **Matplotlib**: Data visualization
- **rouge_score**: Evaluation metrics for summarization
- **Torch**: Deep learning framework

Identifying Gaps and Challenges

- **Coherence and Context**: Many existing systems struggle with producing coherent and contextually accurate summaries.
- **Scalability**: Handling large datasets efficiently remains a significant challenge.

Evaluation Metrics: There is a need for more comprehensive metrics that capture human judgment effectively.

Techniques for Text Summarization

Frequency Analysis

Identifying the most frequently occurring words and phrases to determine the most important content.

TF-IDF (Term Frequency-Inverse Document Frequency)

Semantic Analysis

Understanding the meaning and relationships between concepts to extract the most relevant information.

LSA (Latent Semantic Analysis)

Machine Learning

Leveraging advanced algorithms to learn patterns and generate summaries based on training data.

BERT (Bidirectional Encoder Representations from Transformers)

T5 (Text-To-Text Transfer Transformer)

Pegasus

Graph-based Methods

Modeling the text as a graph and using network analysis to identify the most central and influential components.

TextRank

System Design

Data Collection/Preprocessing

- **Preprocessing**: Tokenization, stop word removal, text cleaning.

Feature Extraction

- **Sentence Embeddings**: Using BERT or similar models for embeddings.

Summarization Methods

- **Traditional Methods**: TF-IDF, TextRank, LSA.
- **Machine Learning Models**: BERT, T5, Pegasus.

Postprocessing

- **Sentence Selection**: For extractive summarization.
- **Text Generation**: For abstractive summarization.

Evaluation

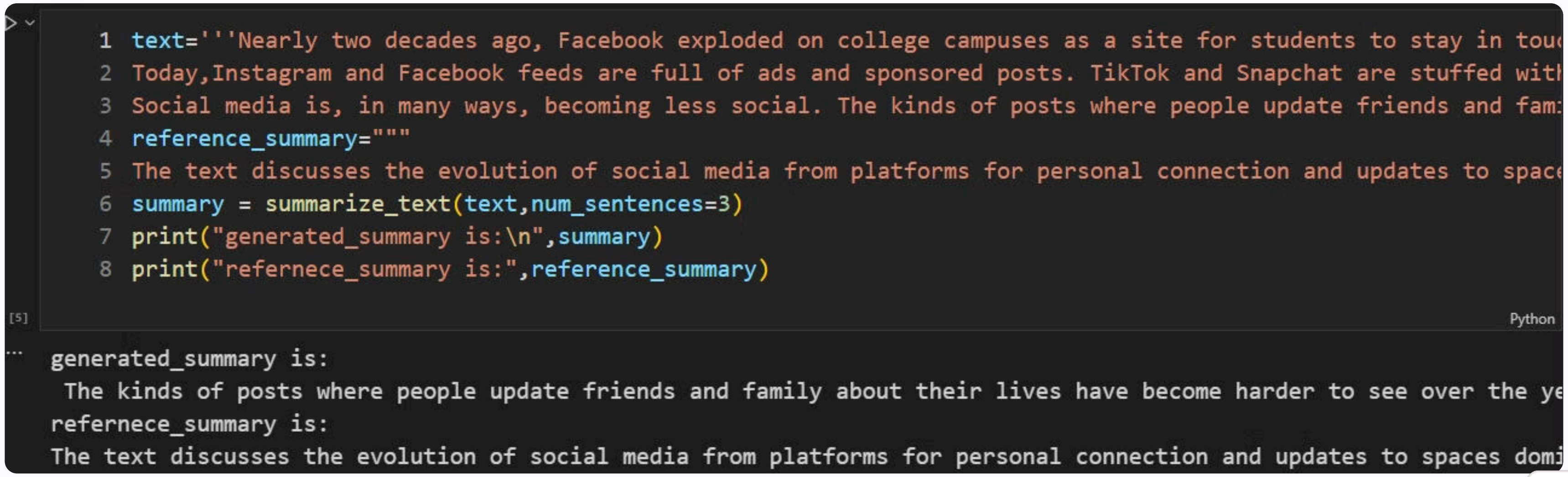
- **Metrics**: ROUGE score calculation.

Implementation

Text Rank:

Text Rank is a graph-based ranking algorithm for text summarization. Similar to PageRank, it identifies the most important sentences in a document by ranking them based on their connectivity in a graph.

- **Step 1: Sentence Tokenization-** Splitting the text into individual sentences.
- **Step 2: Sentence Similarity Calculation-** Calculating the similarity between sentences to create edges in the graph.
- **Step 3: Graph Construction-** Building a graph where sentences are nodes and edges represent similarity scores.
- **Step 4: Sentence Ranking-** Applying the PageRank algorithm to rank sentences based on their importance.
- **Step 5: Summary Generation-** Selecting top-ranked sentences to form the summary.



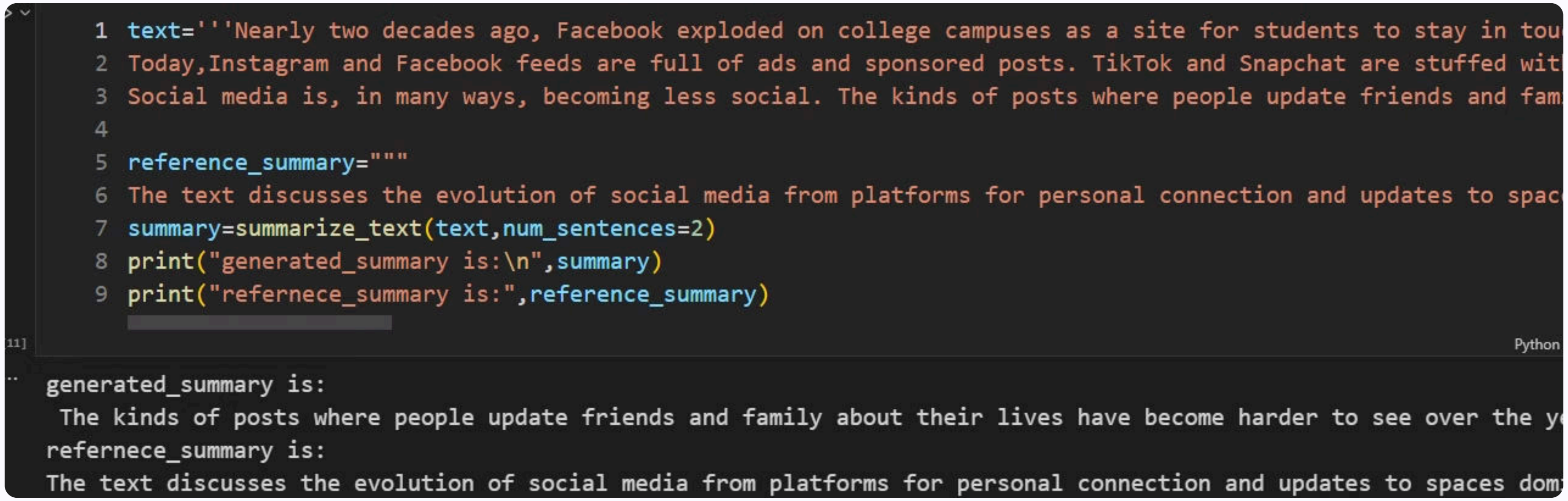
```python
text='''Nearly two decades ago, Facebook exploded on college campuses as a site for students to stay in touc
Today,Instagram and Facebook feeds are full of ads and sponsored posts. TikTok and Snapchat are stuffed wit
Social media is, in many ways, becoming less social. The kinds of posts where people update friends and fam
reference_summary="""
The text discusses the evolution of social media from platforms for personal connection and updates to space
summary = summarize_text(text,num_sentences=3)
print("generated_summary is:\n",summary)
print("refernece_summary is:",reference_summary)
```

```
generated_summary is:
 The kinds of posts where people update friends and family about their lives have become harder to see over the ye
refernece_summary is:
The text discusses the evolution of social media from platforms for personal connection and updates to spaces domi
```

TF-IDF (Term Frequency-Inverse Document Frequency):

TF-IDF is a numerical statistic that reflects the importance of a word in a document relative to a corpus. It helps in identifying the most significant words in a document for text summarization.

- **Step 1: Tokenization-** Splitting the text into individual words or terms.
- **Step 2: Term Frequency Calculation:** Calculating the frequency of each term in the document.
- **Step 3: Inverse Document Frequency Calculation:** Calculating the inverse document frequency for each term.
- **Step 4: TF-IDF Score Calculation:** Multiplying term frequency with inverse document frequency to get TF-IDF scores.
- **Step 5: Summary Generation:** Selecting sentences with the highest TF-IDF scores.



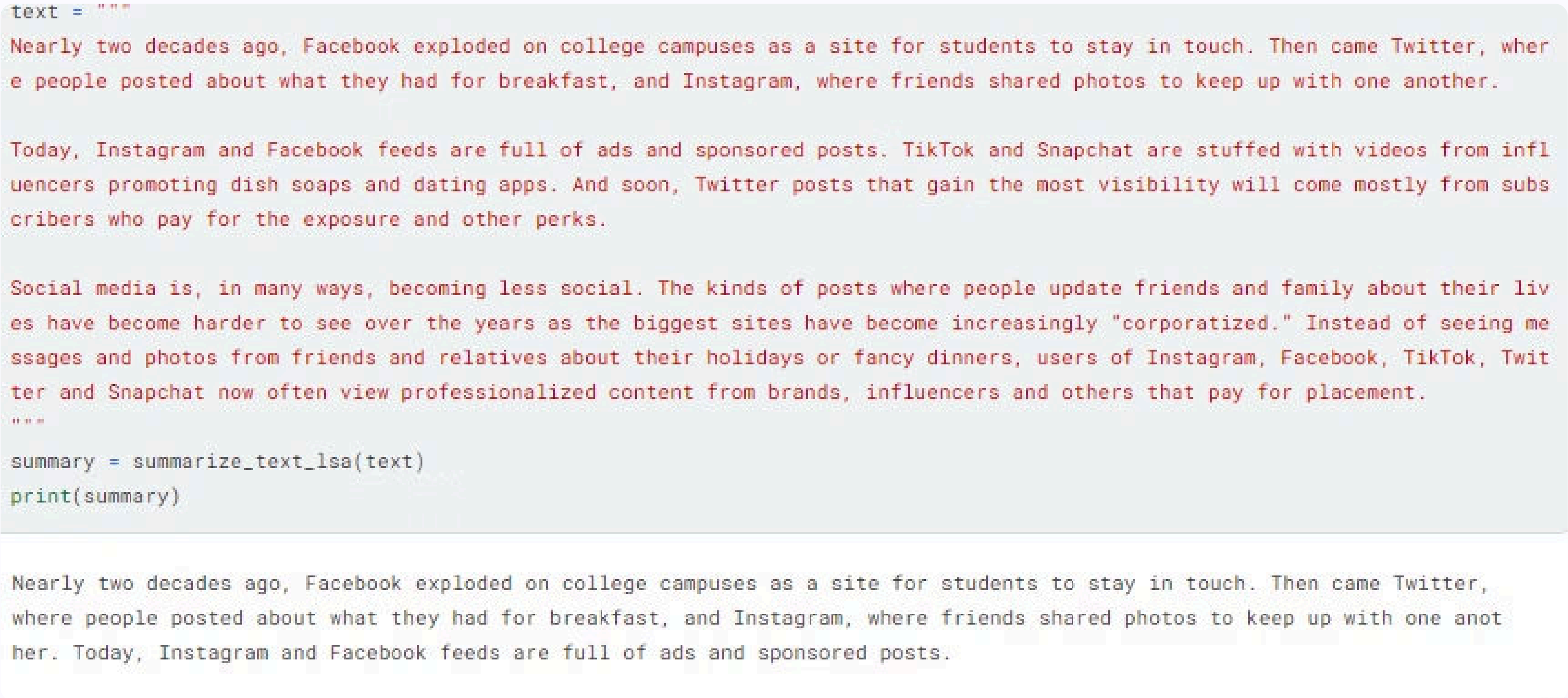
```python
text='''Nearly two decades ago, Facebook exploded on college campuses as a site for students to stay in tou
Today,Instagram and Facebook feeds are full of ads and sponsored posts. TikTok and Snapchat are stuffed wit
Social media is, in many ways, becoming less social. The kinds of posts where people update friends and fam

reference_summary="""
The text discusses the evolution of social media from platforms for personal connection and updates to spac
summary=summarize_text(text,num_sentences=2)
print("generated_summary is:\n",summary)
print("refernece_summary is:",reference_summary)
```

```
generated_summary is:
 The kinds of posts where people update friends and family about their lives have become harder to see over the ye
refernece_summary is:
The text discusses the evolution of social media from platforms for personal connection and updates to spaces dom
```

Latent Semantic Analysis (LSA):

It is a technique in natural language processing that analyzes relationships between a set of documents and the terms they contain by producing a set of concepts related to the documents and terms.

- **Step 1: Term Frequency Matrix-** Creating a term frequency matrix.
- **Step 2: Singular Value Decomposition (SVD)-** Applying SVD to reduce the dimensionality of the term-document matrix.
- **Step 3: Concept Identification-** Identifying key concepts in the text.
- **Step 4: Sentence Scoring-** Scoring sentences based on their relevance to the identified concepts.
- **Step 5: Summary Generation-** Selecting top-scoring sentences to form the summary.



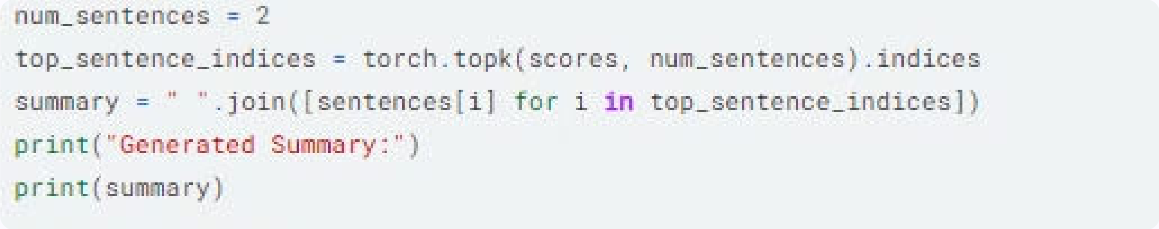
```
text = """
Nearly two decades ago, Facebook exploded on college campuses as a site for students to stay in touch. Then came Twitter, where people posted about what they had for breakfast, and Instagram, where friends shared photos to keep up with one another.

Today, Instagram and Facebook feeds are full of ads and sponsored posts. TikTok and Snapchat are stuffed with videos from influencers promoting dish soaps and dating apps. And soon, Twitter posts that gain the most visibility will come mostly from subscribers who pay for the exposure and other perks.

Social media is, in many ways, becoming less social. The kinds of posts where people update friends and family about their lives have become harder to see over the years as the biggest sites have become increasingly "corporatized." Instead of seeing messages and photos from friends and relatives about their holidays or fancy dinners, users of Instagram, Facebook, TikTok, Twitter and Snapchat now often view professionalized content from brands, influencers and others that pay for placement.
"""
summary = summarize_text_lsa(text)
print(summary)
```

```
Nearly two decades ago, Facebook exploded on college campuses as a site for students to stay in touch. Then came Twitter, where people posted about what they had for breakfast, and Instagram, where friends shared photos to keep up with one another. Today, Instagram and Facebook feeds are full of ads and sponsored posts.
```

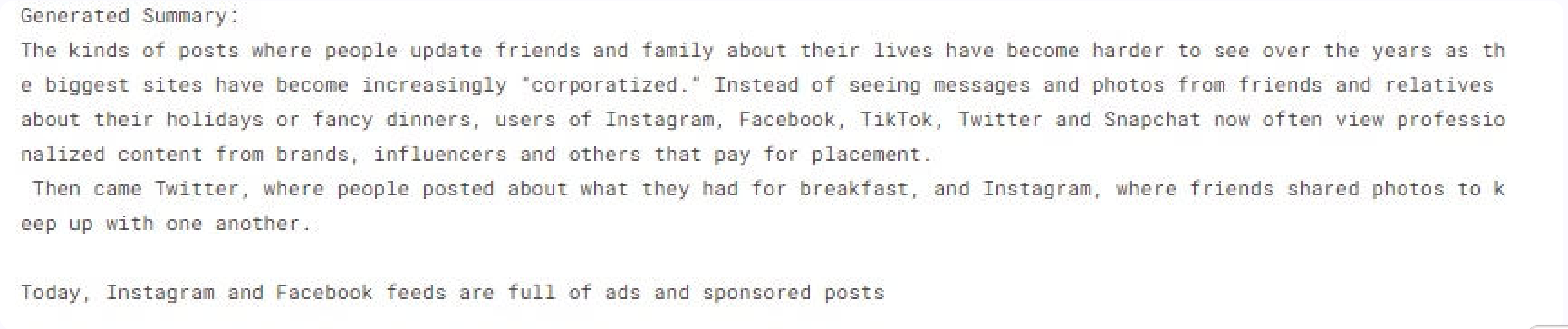
BERT (Bidirectional Encoder Representations from Transformers):

It is a transformer-based model pre-trained on a large corpus of text and fine-tuned for various NLP tasks, including summarization.

- **Step 1: Pre-trained Model Loading-** Loading a pre-trained BERT model.
- **Step 2: Text Preprocessing-** Tokenizing and encoding the text.
- **Step 3: Model Inference-** Using the model to generate embeddings and identify key sentences.
- **Step 4: Sentence Ranking-** Ranking sentences based on their embeddings.
- **Step 5: Summary Generation-** Selecting top-ranked sentences to form the summary.



```
num_sentences = 2
top_sentence_indices = torch.topk(scores, num_sentences).indices
summary = " ".join([sentences[i] for i in top_sentence_indices])
print("Generated Summary:")
print(summary)
```



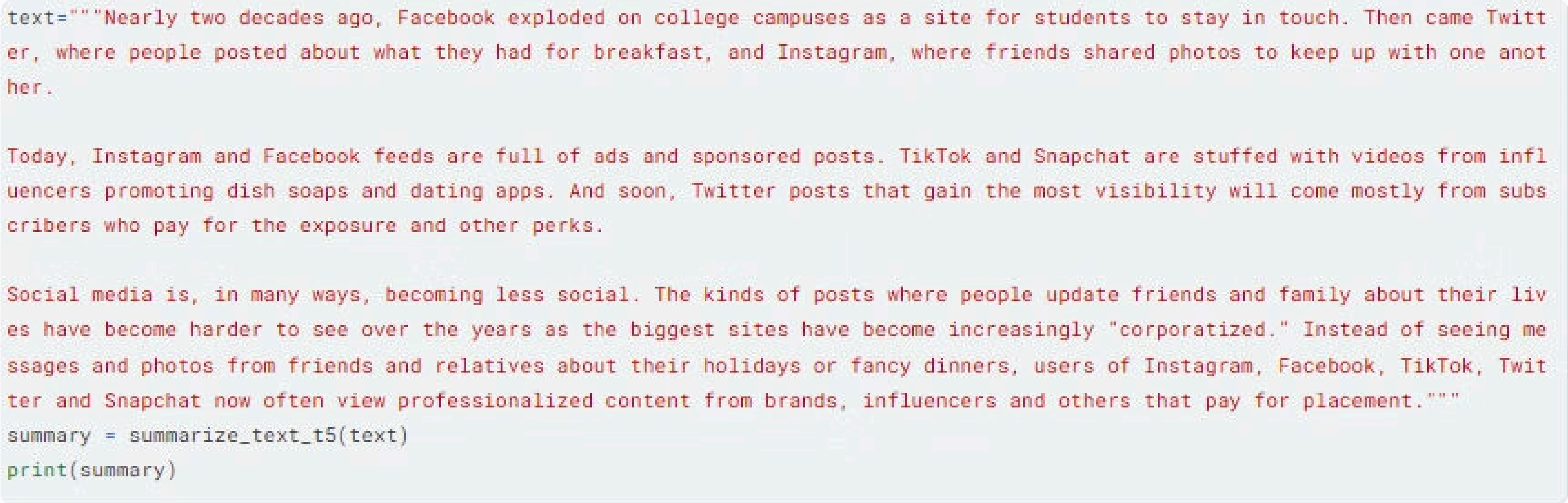
```
Generated Summary:
The kinds of posts where people update friends and family about their lives have become harder to see over the years as th
e biggest sites have become increasingly "corporatized." Instead of seeing messages and photos from friends and relatives
about their holidays or fancy dinners, users of Instagram, Facebook, TikTok, Twitter and Snapchat now often view professio
nalized content from brands, influencers and others that pay for placement.
 Then came Twitter, where people posted about what they had for breakfast, and Instagram, where friends shared photos to k
eep up with one another.

Today, Instagram and Facebook feeds are full of ads and sponsored posts
```

T5 (Text-To-Text Transfer Transformer) & Pegasus Model:

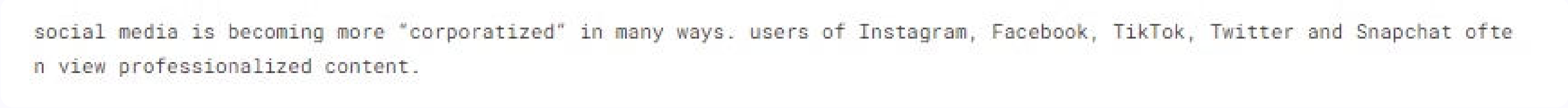
T5- It is a powerful machine learning model developed by Google AI for various text-based tasks. It's a type of encoder-decoder transformer architecture that has been pre-trained on a massive dataset of text and code.T5 can be used for both extractive and abstractive text summarization, but it depends on the fine-tuning approach.

Pegasus- It is a state-of-the-art deep learning model specifically designed for generating abstractive text summaries.

- **Step 1: Pre-trained Model Loading-** Loading a pre-trained BERT model.
- **Step 2: Text Preprocessing-** Tokenizing and encoding the text.
- **Step 3: Model Inference-** Using the model(base, small, large) to generate embeddings and identify key sentences.
- **Post-processing (Optional)**- You may perform additional processing on the generated summary to refine it (e.g., grammatical checks, sentence length adjustments, redundancy removal)
- **Step 5: Summary Generation-** Selecting top-ranked sentences to form the summary.

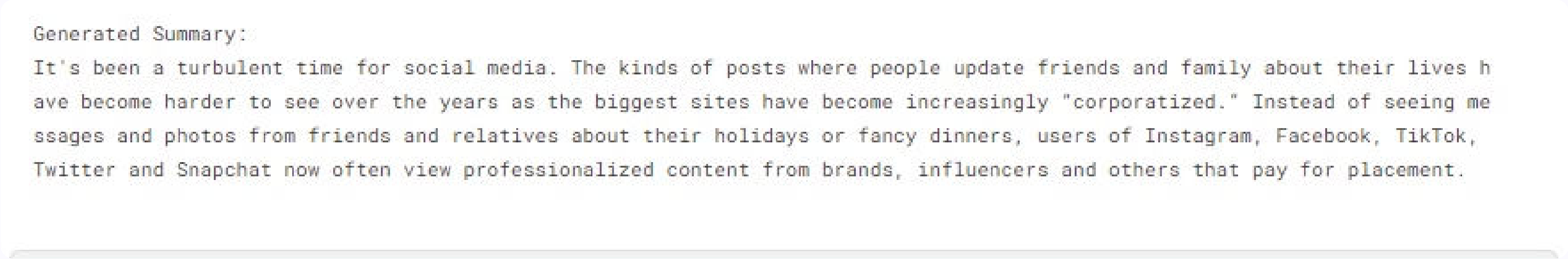


```
text="""Nearly two decades ago, Facebook exploded on college campuses as a site for students to stay in touch. Then came Twitt
er, where people posted about what they had for breakfast, and Instagram, where friends shared photos to keep up with one anot
her.

Today, Instagram and Facebook feeds are full of ads and sponsored posts. TikTok and Snapchat are stuffed with videos from infl
uencers promoting dish soaps and dating apps. And soon, Twitter posts that gain the most visibility will come mostly from subs
cribers who pay for the exposure and other perks.

Social media is, in many ways, becoming less social. The kinds of posts where people update friends and family about their liv
es have become harder to see over the years as the biggest sites have become increasingly "corporatized." Instead of seeing me
ssages and photos from friends and relatives about their holidays or fancy dinners, users of Instagram, Facebook, TikTok, Twit
ter and Snapchat now often view professionalized content from brands, influencers and others that pay for placement."""
summary = summarize_text_t5(text)
print(summary)
```



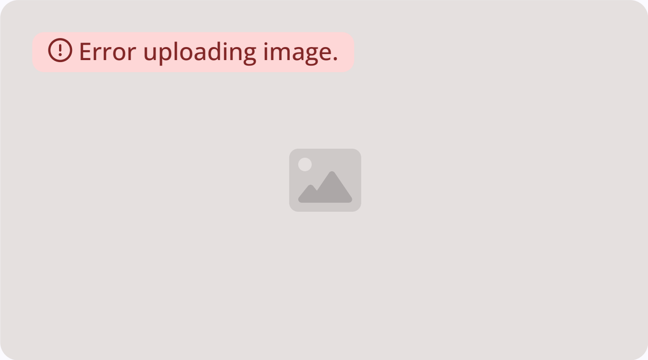
```
social media is becoming more "corporatized" in many ways. users of Instagram, Facebook, TikTok, Twitter and Snapchat ofte
n view professionalized content.
```

Pegasus summary

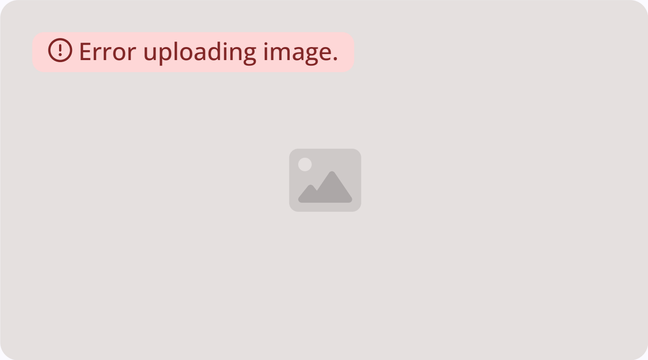


```
Generated Summary:
It's been a turbulent time for social media. The kinds of posts where people update friends and family about their lives h
ave become harder to see over the years as the biggest sites have become increasingly "corporatized." Instead of seeing me
ssages and photos from friends and relatives about their holidays or fancy dinners, users of Instagram, Facebook, TikTok,
Twitter and Snapchat now often view professionalized content from brands, influencers and others that pay for placement.
```

Summarization with T5 model



Summarization with Pegasus model



Applications of Text Summarization

News Summarization

Concise summaries of current events and breaking news.

Academic Summarization

Condensing research papers and scholarly articles.

Email Summarization

Generating quick overviews of long email threads.

Medical Summarization

Summarizing patient records and clinical notes.

Evaluating Text Summarization

ROUGE	Measures the overlap between the summary and reference texts, assessing content quality.
BLEU	Evaluates the grammatical correctness and fluency of the generated summary.
Pyramid	Assesses the informative content of the summary, focusing on important concepts.
Human Evaluation	Subjective assessments by human raters on aspects like relevance, coherence, and usefulness.

Testing

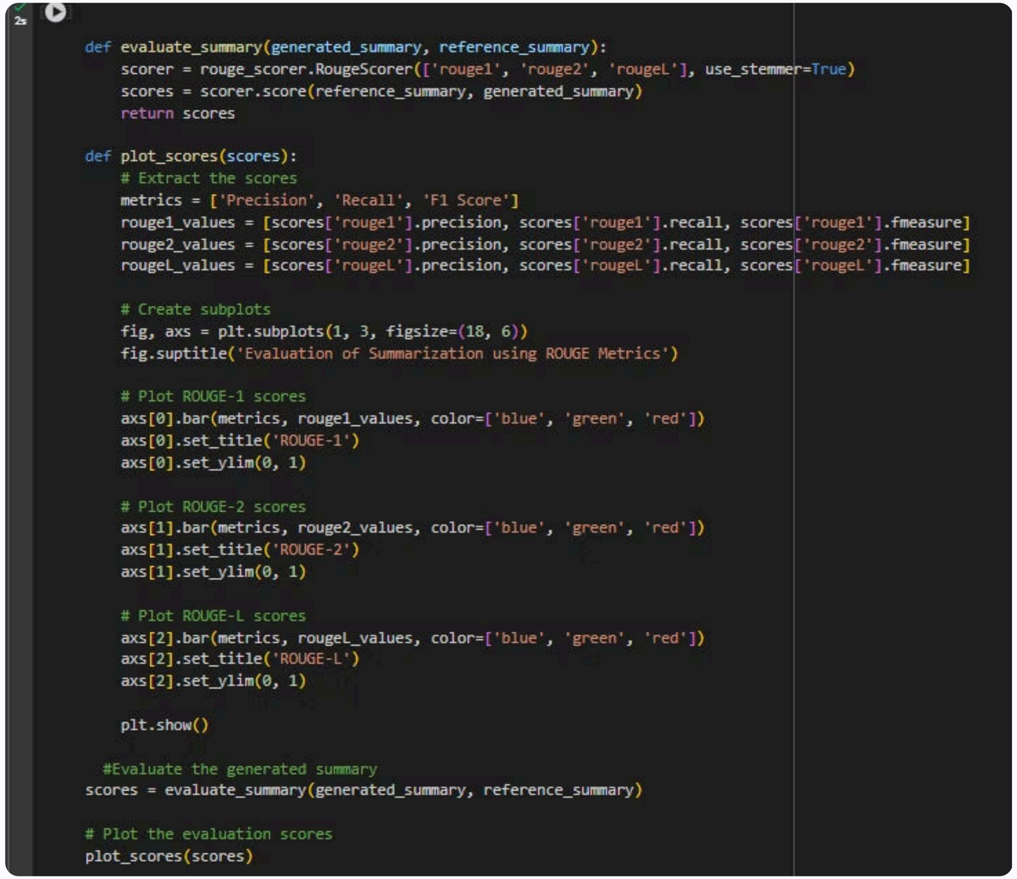
Introduction

The purpose of the testing phase is to evaluate the performance of various text summarization models on a standardized dataset. The models tested include TF-IDF, TextRank, LSA, T5, Pegasus, and a BERT-based model.

Evaluation Metrics

The performance of the models was evaluated using the following metrics:

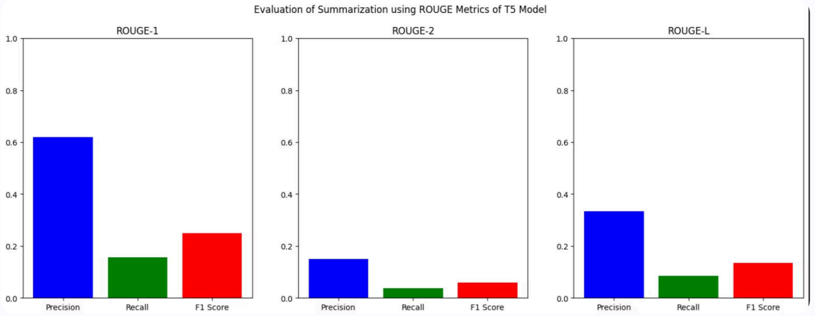
- **ROUGE-1**: Measures the overlap of unigrams between the generated and reference summaries.
- **ROUGE-2**: Measures the overlap of bigrams between the generated and reference summaries.
- **ROUGE-L**: Measures the longest common subsequence between the generated and reference summaries.

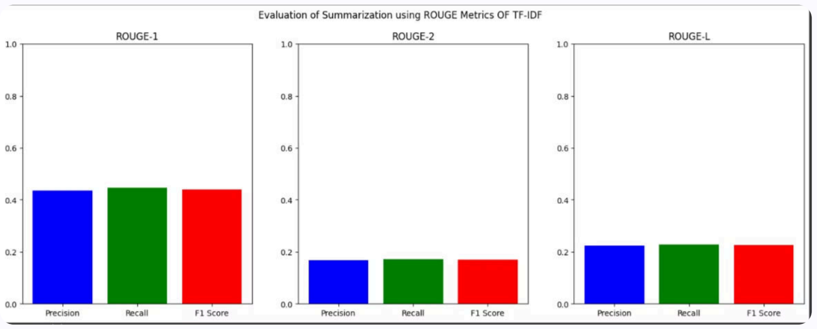


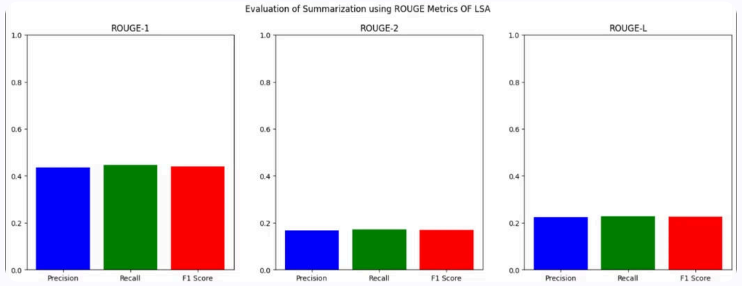
```
def evaluate_summary(generated_summary, reference_summary):
    scorer = rouge_scorer.RougeScorer(['rouge1', 'rouge2', 'rougeL'], use_stemmer=True)
    scores = scorer.score(reference_summary, generated_summary)
    return scores

def plot_scores(scores):
    # Extract the scores
    metrics = ['Precision', 'Recall', 'F1 Score']
    rouge1_values = [scores['rouge1'].precision, scores['rouge1'].recall, scores['rouge1'].fmeasure]
    rouge2_values = [scores['rouge2'].precision, scores['rouge2'].recall, scores['rouge2'].fmeasure]
    rougeL_values = [scores['rougeL'].precision, scores['rougeL'].recall, scores['rougeL'].fmeasure]

    # Create subplots
    fig, axs = plt.subplots(1, 3, figsize=(18, 6))
    fig.suptitle('Evaluation of Summarization using ROUGE Metrics')

    # Plot ROUGE-1 scores
    axs[0].bar(metrics, rouge1_values, color=['blue', 'green', 'red'])
    axs[0].set_title('ROUGE-1')
    axs[0].set_ylim(0, 1)

    # Plot ROUGE-2 scores
    axs[1].bar(metrics, rouge2_values, color=['blue', 'green', 'red'])
    axs[1].set_title('ROUGE-2')
    axs[1].set_ylim(0, 1)

    # Plot ROUGE-L scores
    axs[2].bar(metrics, rougeL_values, color=['blue', 'green', 'red'])
    axs[2].set_title('ROUGE-L')
    axs[2].set_ylim(0, 1)

    plt.show()

  #Evaluate the generated summary
scores = evaluate_summary(generated_summary, reference_summary)

# Plot the evaluation scores
plot_scores(scores)
```

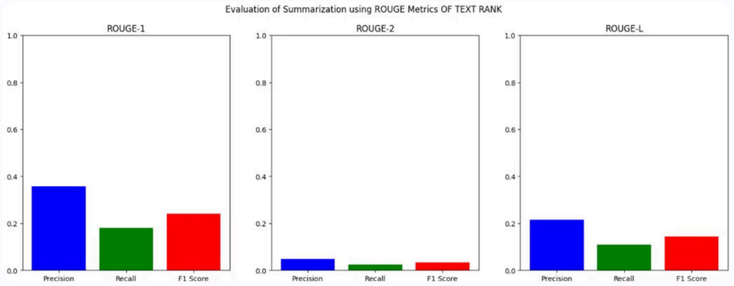
Error uploading image.

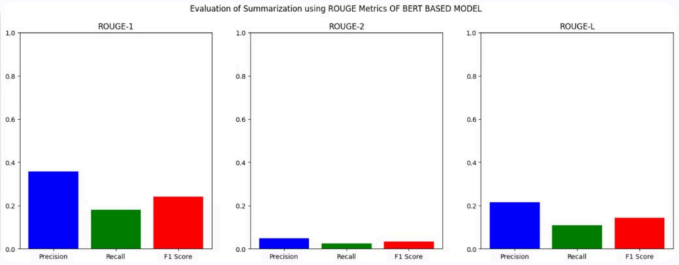
Evaluation Charts

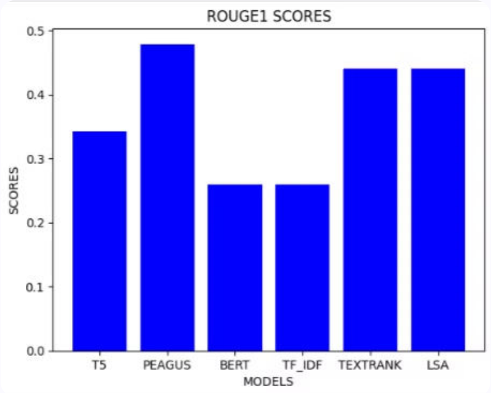


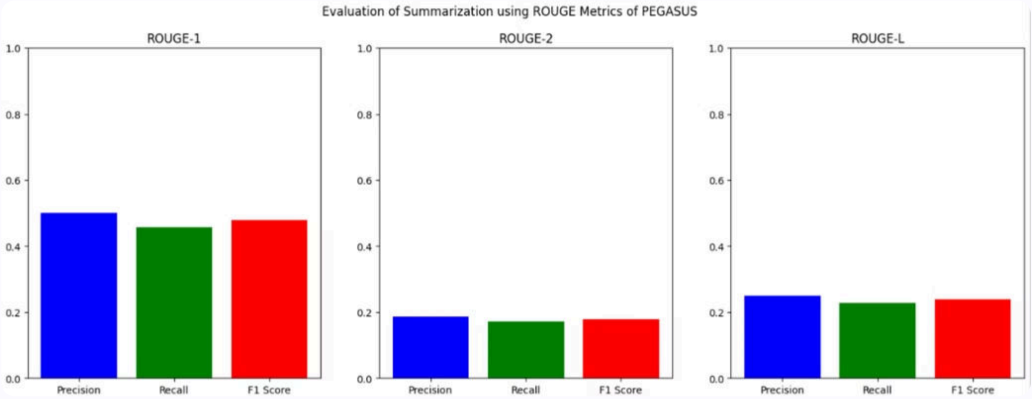


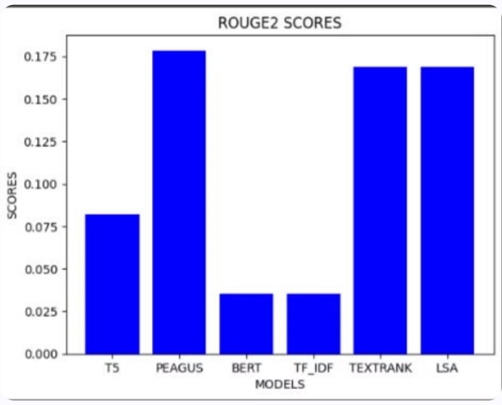


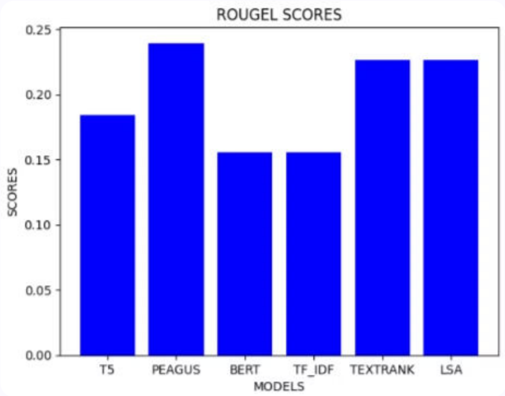












Future Enhancement

- **Hybrid Approach:** Combine strengths of extractive (TextRank/TF-IDF) and abstractive (T5/Pegasus) methods for a more comprehensive summary.
- **Hierarchical Summarization:** Generate summaries at different levels of detail to cater to various needs.
- **Domain-Specific Tuning:** Fine-tune models on domain-specific data (e.g., medical text, legal documents) for improved accuracy.
- **Multilingual Summarization:** Extend the system to handle text summarization in multiple languages.
- **Summarization with Sentiment Analysis:** Include sentiment analysis to capture the emotional tone of the summarized text.

Conclusion

Our exploration of various text summarization techniques (TextRank, TF-IDF, LSA, T5, Pegasus) revealed valuable insights:

- **Extractive vs. Abstractive:** Extractive methods (TextRank, TF-IDF) are efficient for identifying key sentences but lack the ability to rephrase information. Abstractive methods (T5, Pegasus) offer higher fidelity summaries but require more computational resources.
- **Accuracy and Resources:** Pre-trained models like T5 and Pegasus often achieve higher accuracy but come at a cost of greater computational power.
- **Task Specificity:** The ideal summarization method depends on the specific task. TextRank or TF-IDF might suffice for short summaries highlighting key points, while T5 or Pegasus could be preferred for comprehensive and informative summaries.

Ultimately, the choice of text summarization method depends on the specific needs of the application, balancing factors like desired level of abstractiveness, accuracy requirements, and available computational resources. Our project has laid a strong foundation for further exploration and development of a robust and versatile text summarization system.

References

1. Business System Bharati Vidyapeeth Deemed University College of Engineering, Pune, **Text Summarizer Using NLP (Natural Language Processing)** (JUL 2022)
2. Dr Sumathi Pawara* , Dr Manjula Gururaj Hb , Dr Niranajan N Chiplunarc bAssociate Professor,Nitte(Deemed to be University),NMAMIT, Karkala, Karnataka, India ,cProfessor, Nitte (Deemed to be University),NMAMIT, Karkala, Karnataka, India, **Text Summarization Using Document and Sentence Clustering Method** (2022)
3. Narendra Andhale Department of Computer Engineering Vishwakarma Institute of Information Technology Pune, India and L.A. Bewoor Department of Computer Engineering Vishwakarma Institute of Information Technology Pune, India .**An Overview of Text Summarization Techniques**