Report on Text Summarization Project:

1. Dataset Description and Preprocessing

The dataset used in this project consists of textual data, likely sourced from news articles or other structured text sources. The primary goal is to extract meaningful summaries while preserving key information.

Preprocessing Steps:

- **Text Cleaning:** Removal of special characters, HTML tags, and extra spaces.
- **Tokenization:** Splitting text into individual words or subwords.
- **Stopword Removal:** Eliminating common words that do not contribute to summarization.
- **Lemmatization/Stemming:** Converting words to their base or root forms.
- **Handling Missing Data:** Ensuring completeness and consistency in the dataset.

2. Models Implemented and Rationale for Selection

Several models were considered for text summarization:

- **Extractive Summarization (e.g., TextRank, LexRank):** Suitable for selecting key sentences from the original text.
- **Abstractive Summarization (e.g., T5, BART, Pegasus):** Used for generating human-like summaries by understanding the text context.

Final Model Choice:

- Transformer-based models such as **BART** and **T5** were implemented due to their strong performance in abstractive summarization tasks.
- The selection was based on their ability to generate coherent, concise, and contextually accurate summaries.

3. Key Insights and Visualizations

- **ROUGE Score Analysis:** Evaluated model performance by comparing generated summaries with reference summaries.
- **Word Frequency Distribution:** Visualized common words and key themes in the dataset.
- **Comparison of Extractive vs. Abstractive Summaries:** Showed differences in summarization approaches and effectiveness.

4. Challenges Faced and Solutions

- **Handling Long Texts:**
    - Used truncation and sliding window approaches to manage token limits.
- **Model Performance Optimization:**
    - Fine-tuned transformer models on domain-specific data for better accuracy.
- **Data Quality Issues:**
    - Applied advanced preprocessing techniques to improve text coherence.

Conclusion

The project successfully implemented and evaluated text summarisation techniques using NLP models. Transformer-based approaches provided high-quality summaries, and performance was validated using standard metrics like ROUGE. Future work could involve optimizing inference speed and exploring multilingual summarization.
