BY: ARYA H

REG NO : 71762234005

MSc. Artificial Intelligence and Machine Learning

Coimbatore Institute of Technology

Text Summarization and Keyword Extraction Report

Objective

This script performs two core NLP tasks: generating a concise 3-point summary and extracting the top 5 keywords from an input paragraph.

Models and Libraries Used

For summarization, we utilize Hugging Face's transformers pipeline with the pre-trained model sshleifer/distilbart-cnn-12-6, which is a distilled version of BART fine-tuned for summarization tasks. It offers a good trade-off between performance and speed. For keyword extraction, we use the KeyBERT model, which leverages sentence-transformers to generate semantic embeddings and rank keyword phrases based on cosine similarity.

Logic

The text is first cleaned to remove unnecessary whitespace. The summarization function dynamically adjusts summary length based on the input word count, ensuring meaningful output while maintaining brevity. It returns exactly three bullet points, repeating sentences if necessary. The keyword extraction function identifies the most relevant terms or phrases using embeddings from KeyBERT.

Scope for Improvement

- Allow support for multilingual summarization using models like mBART.
- Improve redundancy handling in bullet points.
- Enhance keyword relevance by integrating domain-specific fine-tuning.
- Add visualization (e.g., word cloud) for better insight.