SUMMARIZATION OF NLP TECHNIQUES

Project Overview

This repository contains implementations of summarization techniques, BLEU/ROUGE scoring, advanced LLM techniques, and interactive Gradio applications for various NLP tasks. Below is a detailed overview of each file and its purpose.

1. Abstractive Summarization

- **File Name**: ABSTRACTIVE_SUMMARIZATION.ipynb

- **Description**:
  Implements abstractive summarization techniques using models like T5, BART, and other large language models (LLMs). Abstractive summarization generates summaries that go beyond simply extracting text; instead, they rephrase or condense the original text while maintaining the semantic essence.

- **Purpose**:

  o   Provide a more human-like summary of text.

  o   Suitable for applications like news summarization, content creation, and more.

- **How It Works**:

  o   Pre-trained models (e.g., T5, BART) are loaded using Hugging Face Transformers.

  o   Input text is tokenized and passed through the model for inference.

  o   Output summaries are post-processed to improve readability.

2. Extractive Summarization

- **File Name**: EXTRACTIVE_SUMMARIZATION.ipynb

- **Description**:
  Utilizes algorithms such as Luhn, LexRank, and KMeans for extractive summarization. This approach involves identifying and selecting the most important sentences from the text based on statistical and semantic features.

- **Purpose**:

  o   Provides a concise version of text while retaining the original wording.

  o   Commonly used for document summarization in legal, academic, or corporate contexts.

- **How It Works**:

  o   Luhn: Ranks sentences based on term frequency and position.

  o   LexRank: Calculates sentence importance using a graph-based approach.

  o   KMeans: Clusters similar sentences and selects representatives for each cluster.

3. Interactive Gradio Application for Summarization

- **File Name**: GRADIO.ipynb

- **Description**:
  Creates an interactive Gradio application for text summarization. Users can upload text, enter URLs, or provide PDFs to generate summaries in real time.

- **Purpose**:
  - Makes summarization models accessible to non-technical users.
  - Facilitates real-time experimentation with summarization techniques.

- **How It Works**:
  - Integrates the abstractive and extractive summarization models into a Gradio interface.
  - Accepts multiple input formats and returns the generated summary along with model metadata.

4. Evaluation Metrics: BLEU and ROUGE Scoring

- **File Name**: EVALUATION_BLEU_ROUGE.ipynb

- **Description**:
  Implements BLEU and ROUGE scores to evaluate summarization quality using the CNN/DailyMail dataset.

- **Purpose**:
  - BLEU: Measures n-gram precision by comparing generated summaries with reference summaries.
  - ROUGE: Focuses on recall, comparing overlaps of unigrams, bigrams, and longer n-grams.

- **How It Works**:
  - Loads the CNN/DailyMail dataset and pre-processes it for evaluation.
  - Runs BLEU and ROUGE scoring scripts to output evaluation metrics for different models.
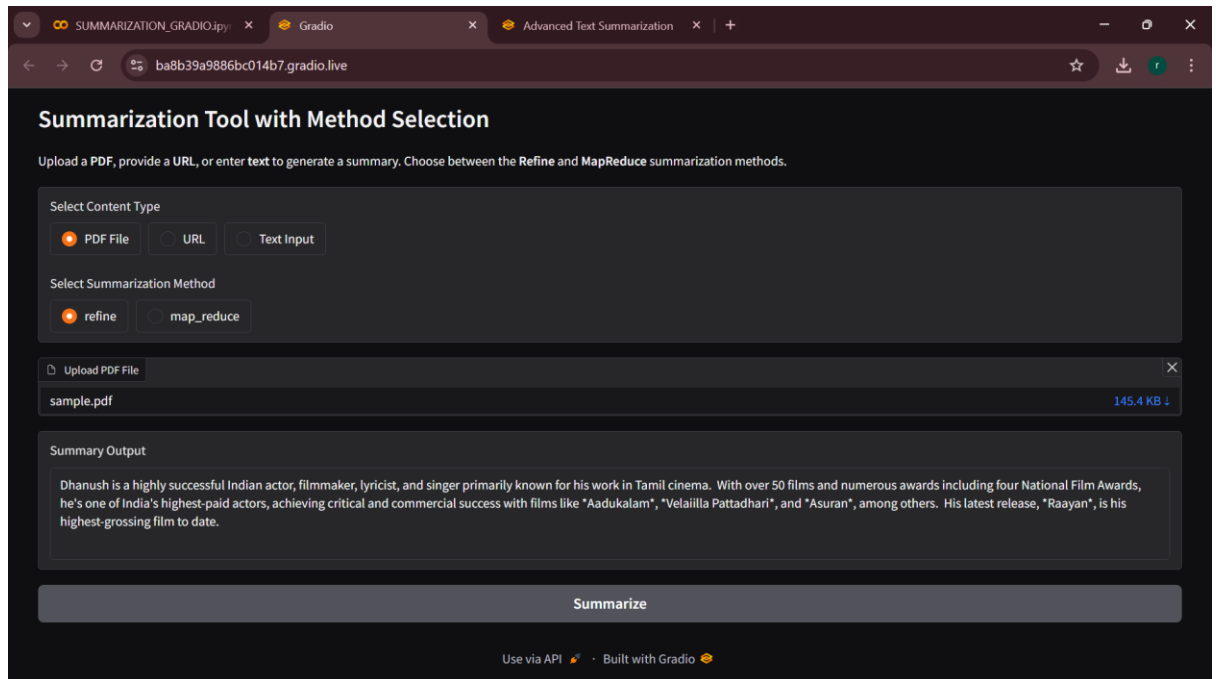
5. Advanced LLM Techniques

- **File Name**: ADVANCED_LLM_TECHNIQUES.ipynb

- **Description**:
  Demonstrates advanced techniques in LangChain, such as MapReduce and iterative summarization.

- **Purpose**:

  o Explores methods to handle large documents effectively.

  o Enhances summarization by breaking down tasks into smaller, manageable chunks.

- **How It Works**:

  o MapReduce: Splits text into smaller sections, summarizes each, and combines the results.

  o Iterative Summarization: Continuously refines summaries by summarizing summaries.

6. Gradio App for Advanced Techniques

- **File Name**: DEVELOPING_GRADIO_FOR_LLM_ADVANCED_MODELS.ipynb

- **Description**:
  Creates an interactive Gradio interface for LangChain-based advanced summarization techniques.

- **Purpose**:

  o Provides a user-friendly way to experiment with MapReduce and iterative summarization.

  o Allows input in various formats, including URLs and PDFs.

- **How It Works**:

- o Incorporates advanced techniques and integrates them into a visually appealing Gradio app.

7. Comprehensive Summarization Gradio App

- **File Name**: SUMMARIZATION_GRADIO.ipynb

- **Description**:
  A fully integrated Gradio app that combines all summarization techniques (abstractive, extractive, advanced) and supports multiple input types like URLs, PDFs, and plain text.

- **Purpose**:
  - o Acts as a one-stop solution for summarization tasks.
  - o Designed for deployment in real-world use cases such as education, journalism, and research.

- **How It Works**:
  - o Allows users to switch between different summarization approaches.
  - o Processes input data and generates summaries using user-selected methods.
  - o Outputs include the summary, metadata, and evaluation scores (if applicable).

Getting Started

Prerequisites

- Python 3.8+
- Jupyter Notebook
- Required libraries:

- pip install transformers gradio langchain rouge-score

Running the Notebooks

1. Open the notebooks in Jupyter or a similar IDE.

2. Follow the instructions in each notebook to execute the cells.

Applications

1. **Abstractive Summarization**: Ideal for generating concise yet creative summaries.

2. **Extractive Summarization**: Best for retaining the exact wording of the source text.

3. **Evaluation**: Ensures the quality of summarization models with quantitative metrics.

4. **Gradio Apps**: Provide easy-to-use interfaces for non-technical users.

This repository is designed to provide a complete suite of summarization tools, from algorithmic implementations to interactive demos.

SAMPLE SNAPSHOT OF FINAL TASK