Document Search and Summarization using Retrieval-Augmented Generation (RAG)

Overview

This project presents a Retrieval-Augmented Generation (RAG) system designed for efficient document search and summarization using Large Language Models (LLMs). The system combines semantic retrieval with neural text generation to provide concise, query-focused summaries from a large document corpus. The complete pipeline is implemented using Hugging Face models, FAISS for similarity search, and a Gradio-based interface, and is fully executable in Google Colab.

Objective

The primary objective of this project is to design and implement a system capable of accurately retrieving relevant documents and generating meaningful summaries. Specifically, the system focuses on scalable semantic search, high-quality abstractive summarization, and seamless user interaction, while strictly adhering to the assignment requirements.

Dataset

The AG News dataset is used as the document corpus for this project. It is a publicly available dataset provided by Hugging Face and consists of news articles from four categories: World, Sports, Business, and Technology. The dataset is automatically downloaded during execution, and no manual data preparation is required.

Models and Tools

Sentence embeddings are generated using the sentence-transformers/all-MiniLM-L6-v2 model. Summarization is performed using the facebook/bart-large-cnn transformer model. FAISS is employed as the vector database for efficient similarity search, while Gradio is used to build an interactive web-based user interface. Evaluation is conducted using ROUGE-1, ROUGE-2, and ROUGE-L metrics.

Methodology

The workflow begins with loading and preprocessing text data from the AG News dataset. Dense vector embeddings are then generated for each document and indexed using FAISS. When a user submits a query, the system retrieves the top-K most relevant documents based on semantic similarity. These documents are passed to a transformer-based summarization model to generate concise and coherent summaries. Performance is evaluated for both retrieval accuracy and summarization quality.

Evaluation

Retrieval performance is assessed by verifying whether relevant documents appear within the top-K retrieved results. Summarization quality is evaluated using ROUGE metrics, which measure n-gram overlap between generated summaries and reference text segments.

User Interface

A Gradio-based web interface allows users to enter natural language queries, adjust the number of retrieved documents, control summary length, and view retrieved documents along with generated summaries in real time.

Execution Instructions

To run the project, open the provided notebook in Google Colab and execute all cells sequentially. The dataset will be downloaded automatically, embeddings will be generated and indexed, and the Gradio interface will launch for interactive use.

Challenges and Solutions

Short input documents can reduce summarization quality; this is addressed by dynamically adjusting summary length based on input size. Efficient retrieval from large document collections is achieved through FAISS-based vector indexing, ensuring fast and scalable similarity search.

Conclusion

This project demonstrates a practical and scalable approach to document retrieval and summarization using Retrieval-Augmented Generation. By integrating semantic search with transformer-based summarization, the system delivers accurate, efficient, and user-friendly document understanding capabilities.