Title:
Retrieval-Augmented Summarization Pipeline using MiniLM, FAISS, and BART

Author: Riyyan Kharal

Objective:
The goal of this project is to build a summarization system that efficiently handles long documents by retrieving only the most relevant portions (top-k chunks) and summarizing them using a pre-trained language model. This balances efficiency and accuracy by limiting the input to the summarizer while maintaining content relevance.

Methodology:
Our pipeline is designed as a modular system with the following stages:

1. **Chunking:**
   The input article is split into fixed-length chunks of 500 words to ensure manageable and semantically meaningful inputs.
2. **Embedding:**
   Each chunk is embedded into a dense vector using the SentenceTransformer model all-MiniLM-L6-v2, chosen for its lightweight size and high semantic retrieval quality.
3. **FAISS Vector Search:**
   The dense vectors are indexed using Facebook's FAISS (FlatL2) for efficient top-k similarity search. We retrieve the top 5 chunks most relevant to the prompt: "Summarize this document".
4. **Summarization:**
   The retrieved chunks are concatenated (up to 4000 characters) and passed to the facebook/bart-large-cnn summarization model using HuggingFace's pipeline.

Models Used:

- **Embedding:** all-MiniLM-L6-v2 (Sentence Transformers)
- **Retrieval:** FAISS IndexFlatL2
- **Summarization:** facebook/bart-large-cnn via HuggingFace Transformers

Pipeline Diagram:

Input Document

↓

Chunking (500 words)

↓

Sentence Embedding (MiniLM)

↓

FAISS Index + Retrieval (Top-5)

↓

Concatenation of Chunks

↓

Summarization (BART)

↓

Output Summary

Results & Analysis

Sample Results (CNN/DailyMail Dataset):

Article	Input Tokens	Summary Tokens	Top Chunk Similarity (max)	Summary Output
1	627	53	0.0943	Palestinian Authority joins ICC; Gaza conflict highlighted.
2	700	53	0.0342	Injured dog survives multiple life-threatening incidents.
3	682	49	0.0472	Iranian diplomat Zarif's political background and nuclear talks.

Key Findings:

- The pipeline effectively condenses multi-paragraph articles into 2–3 sentence summaries.
- Despite some repeated chunks in top-k retrieval (due to FAISS L2 vs cosine mismatch), the summaries remain coherent and informative.
- BART handles concatenated context well within the 4000-character limit.
- FAISS + MiniLM is performant, with retrieval and embedding taking <1s per document.

Limitations:

- Some duplicate chunks appear in the top-k retrieved set, reducing diversity.
- BART may truncate critical information if context exceeds its max length.
- Similarity via L2 distance in FAISS may not always align with cosine similarity ranking.