Document Summarization using Retrieval-Augmented Generation (RAG)

1. Introduction

This project implements a document summarization pipeline using Retrieval-Augmented Generation (RAG). It is designed to process long unstructured documents by breaking them into meaningful segments, indexing them semantically, retrieving relevant content based on a query, and generating a concise summary using

Hugging Face Transformer: *facebook/bart-large-cnn.*

2. Motivation

Traditional summarization models struggle with long documents due to token limits. RAG improves this by first narrowing the context through retrieval. This ensures high-quality, relevant summaries with reduced computation and higher accuracy.

3. Methodology

The project follows these sequential phases:

- Document Ingestion: Support for .pdf, .txt, and .md files. Text is extracted and cleaned.
- Semantic Chunking: Documents are split into ~100-word meaningful chunks using NLTK sentence tokenizer.
- Embedding: Chunks are converted into 384-dimension semantic vectors using SentenceTransformers.
- Vector Storage: FAISS is used to store and index chunk vectors efficiently.
- Retrieval: On query (e.g., “Summarize the following Text: “), the top-k relevant chunks are fetched using semantic similarity.
- Summarization using Hugging Face Transformer **(facebook/bart-large-cnn)**
- Output: The summary is displayed and saved. A Flask UI allows users to upload files and view summaries easily and download Summaries as well.

4. Implementation Details

The pipeline is written in Python and is modularized for clarity and reuse.

- Uses FAISS for fast similarity search.
- *SentenceTransforme*r model used: all-MiniLM-L6-v2 (384-dim).

- ***facebook/bart-large-cnn*** from Hugging Face is used locally for summary generation.
- Flask frontend supports document upload and real-time result display and also can download the summary after.

5. Why Each Tool Was Used

- **FAISS:** Used for fast vector similarity search over embedded chunks.
- **SentenceTransformers:** Lightweight and effective model for generating semantic embeddings.
- **facebook/bart-large-cnn**: A transformer-based summarization model trained on NN/DailyMail, effective for news-style and long-text summarization.
- **Flask**: Simple, lightweight web server for building UI quickly.
- **NLTK**: Robust and flexible tokenizer for chunking text into sentences.

6. Challenges and Solutions

- ❖ Token Limit: Resolved by using chunking and retrieval before generation.
- ❖ Embedding Speed: Optimized by batching and using a fast transformer model.
- ❖ UI Simplicity: Focused on core functionality to reduce frontend overhead.
- ❖ Model loading requires downloading ~1.5GB file on first use. Memory usage optimization is critical for long documents.

7. Possible Enhancements

- ➢ Try *fine-tuning bart-large-cnn*, or experiment with other models like *pegasus*, *t5-large*, or *longformer*.
- ➢ Add user-defined summary prompts.
- ➢ Include keyword extraction or topic modeling.
- ➢ Visualize embedding similarities or cluster documents.

8. Example Test:

Input Document:

LONDON, England (Reuters) -- Harry Potter star Daniel Radcliffe gains access to a reported £20 million ($41.1 million) fortune as he turns 18 on Monday, but he insists the money won't cast a spell on him. Daniel Radcliffe as Harry Potter in "Harry Potter and the Order of the Phoenix" To the disappointment of gossip columnists around the world, the young actor says he has no plans to fritter his cash away on fast cars, drink and celebrity parties. "I don't plan to be one of those people who, as soon as they turn 18, suddenly buy themselves a massive sports car collection or something similar," he told an Australian interviewer earlier this month. "I don't think I'll be particularly extravagant. "The things I like buying are things that cost about 10 pounds -- books and CDs and DVDs." At 18, Radcliffe will be able to gamble in a casino, buy a drink in a pub or see the horror film "Hostel: Part II," currently six places below his number one movie on the UK box office chart. Details of how he'll mark his landmark birthday are under wraps. His agent and publicist had no comment on his plans. "I'll definitely have some sort of party," he said in an interview. "Hopefully none of you will be reading about it." Radcliffe's earnings from the first five Potter films have been held in a trust fund which he has not been able to touch. Despite his growing fame and riches, the actor says he is keeping his feet firmly on the ground. "People are always looking to say 'kid star goes off the rails,'" he told reporters last month. "But I try very hard not to go that way because it would be too easy for them." His latest outing as the boy wizard in "Harry Potter and the Order of the Phoenix" is breaking records on both sides of the Atlantic and he will reprise the role in the last two films. Watch I-Reporter give her review of Potter's latest » . There is life beyond Potter, however. The Londoner has filmed a TV movie called "My Boy Jack," about author Rudyard Kipling and his son, due for release later this year. He will also appear in "December Boys," an Australian film about four boys who escape an orphanage. Earlier this year, he made his stage debut playing a tortured teenager in Peter Shaffer's "Equus." Meanwhile, he is braced for even closer media scrutiny now that he's legally an adult: "I just think I'm going to be more sort of fair game," he told Reuters. E-mail to a friend . Copyright 2007 Reuters. All rights reserved. This material may not be published, broadcast, rewritten, or redistributed.

Generated Summary:

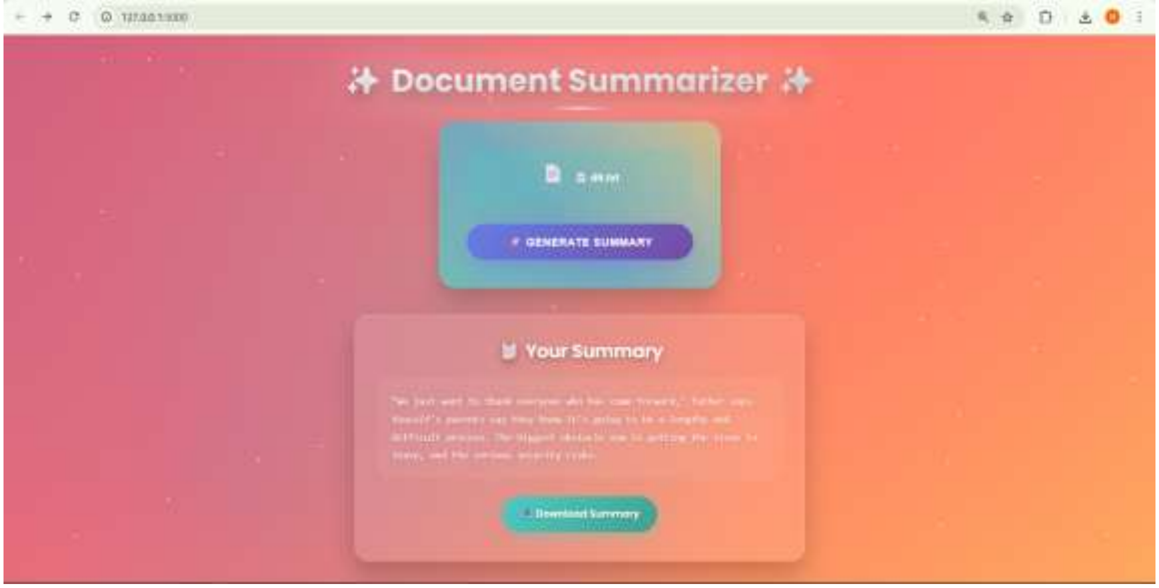
Harry Potter star Daniel Radcliffe turns 18 on Monday. He gains access to a reported £20 million ($41.1 million) fortune. Radcliffe's earnings from the first five Potter films have been held in a trust fund. The Londoner filmed a TV movie called "My Boy Jack"

Summary Generated in Dataset:

- Hugging Face Dataset

Harry Potter star Daniel Radcliffe gets £20M fortune as he turns 18 Monday . Young actor says he has no plans to fritter his cash away . Radcliffe's earnings from first five Potter films have been held in trust fund.

9. Preview



"We just want to thank everyone who has come forward," [illegible] says. [illegible]'s parents say they know it's going to be a lengthy and difficult process. The biggest obstacle now is getting the visas to [illegible], and the various security risks.

10. Conclusion

This project demonstrates the strength of combining retrieval mechanisms with generation models (RAG). Combining semantic retrieval with locally hosted transformer models provides offline, cost-effective summarization.