PDF Semantic Retrieval Using LangChain and FAISS

Khwaish Khandelwal
Dept of Artificial Intelligence and Data Science
Dr. Akhilesh Das Gupta Institute of Professional Studies, New Delhi

Abstract- *In today's information-driven environment, PDF documents serve as a primary medium for storing academic, corporate, legal, and technical knowledge. However, retrieving specific and meaningful information from large PDF files remains a major challenge, especially when relying on traditional keyword-based search methods that fail to capture deeper semantic meaning. This project, "PDF Semantic Retrieval System using LangChain and FAISS", addresses this challenge by developing an intelligent, context-aware retrieval system capable of understanding user queries and locating the most relevant sections within PDF documents.*

The proposed system extracts text from PDF files, segments it into context-preserving chunks, and generates high-dimensional semantic embeddings using transformer-based models. These embeddings are stored in FAISS, a high-performance vector search library optimized for large-scale similarity search.

I. INTRODUCTION

Digital documents, especially PDF files, have become the dominant medium for distributing information in academic, industrial, legal, and corporate environments. Although PDFs provide portability and structural consistency, they pose significant limitations when users attempt to retrieve specific information from large documents. Traditional keyword-based search methods fall short when dealing with complex language, varying terminology, or context-dependent information. With the advancement of Natural Language Processing and semantic search technologies, there is an opportunity to build more intelligent systems capable of understanding user intent rather than relying solely on literal keyword matches.

This project aims to design and develop a **PDF Semantic Retrieval System using LangChain and FAISS**, which enables context-aware search across PDF documents.

A. Challenges

- PDF documents often contain varied layouts, tables, images, and text structures that make automated extraction difficult. Text may not be sequential, and important information can be embedded in different formats.
- Traditional search tools depend on exact word matches, failing to capture semantic meaning. If the user query uses different terminology than the document, relevant information may be missed entirely.
- Academic papers, legal reports, corporate documents, and technical manuals can span hundreds of pages, making manual scanning extremely time-consuming and inefficient for users.
- Keyword search is unable to infer relationships between concepts, leading to irrelevant or superficial results.
- As the number and size of documents increase, conventional search systems struggle to maintain performance and accuracy.

These challenges highlight the necessity of a more intelligent retrieval mechanism that goesbeyond literal text matching and instead understands the meaning behind user queries.

B. Need for This System

The increasing use of PDF documents in academic, corporate, and technical environments has highlighted a major gap in efficient information retrieval. Traditional keyword-based search methods are limited to exact text matching and fail to capture the contextual meaning behind user queries.

This means that even if relevant information exists within a document, users may not find it unless the exact keyword appears in the text. As PDFs grow longer and more complex, containing dense research content, legal clauses, or technical instructions, manually scanning them becomes time-consuming, inefficient, and prone to error. These limitations create an urgent requirement for a more intelligent system that goes beyond keyword matching.

II. LITERATURE REVIEW

The evolution of document retrieval systems has progressed significantly from simple keyword-based methods to advanced semantic search frameworks powered by deep learning. Early retrieval systems relied on Boolean search and statistical scoring mechanisms such as Term Frequency–Inverse Document Frequency (TF-IDF) and BM25. While these methods were efficient for straightforward matching, they lacked an understanding of the meaning behind user queries, often resulting in irrelevant search results when synonyms or paraphrased language was used.

The proposed **PDF Semantic Retrieval System using LangChain and FAISS** addresses these gaps by combining advanced text chunking methods, high-performance vector search, and context-aware query processing into a unified system. By leveraging the strengths of LangChain for orchestration and FAISS for efficient retrieval, this system is designed to provide accurate, meaningful, and fast access to relevant information from single or multiple PDF documents, making it highly applicable across academic, corporate, legal, and research domains.

III. OBJECTIVES AND SCOPE OF WORK

A. *Objectives*

The central objective of this project is to design, develop, and evaluate a semantic retrieval system specifically tailored for PDF documents, enabling users to extract meaningful and contextually relevant information using natural language queries. This project aims to overcome the limitations of traditional keyword-based search methods, which rely solely on literal word matching and often fail to capture the true intent behind a query. By integrating the LangChain framework with the FAISS vector database, the system will leverage semantic embeddings to ensure that document retrieval is based on conceptual meaning rather than superficial keyword similarity.

B. *Scope of Work*

The scope of this project covers the complete development lifecycle of the PDF Semantic Retrieval System using LangChain and FAISS, from initial requirements gathering to system deployment and evaluation. The project will focus specifically on semantic retrieval in PDF documents, ensuring that it can serve diverse application domains such as academic research, legal document analysis, corporate knowledge management, and technical documentation search.

The work will begin with the parsing and preprocessing of PDF documents, whichinvolves extracting raw text from files of varying complexity. This stage will address challenges such as handling multi-page layouts, identifying headers and footers, managing multi-column formats, and preserving the logical order of text. Where applicable, the system may also integrate OCR capabilities to process scanned or image-based PDFs.

IV. METHODOLOGY

The overall workflow can be divided into the following major components: PDF Ingestion and Parsing, Text Chunking and Preprocessing, Embedding Generation, FAISS Indexing, Query Processing and Retrieval, LangChain Orchestration, and User Interface Development.

A. PDF Ingestion and Parsing

The first stage of the system involves document ingestion, where users upload one or multiple PDF files into the system. This stage useLangChain’s PDF loaders or other robust parsing libraries (e.g., PyPDF2, pdfplumber) to handle various PDF structures.

B. Text Chunking and Preprocessing

Once the PDF content is extracted, it will undergo chunking — the process of dividing large text bodies into smaller, context-rich segments. This is necessary because embedding models have token limitations and because smaller, well-defined segments
improve retrieval precision.

C. Embedding Generation

The next stage involves generating semantic embeddings for each text chunk. Embeddings are high-dimensional vector representations that capture the contextual meaning of the text, enabling the system to compare and retrieve passages based on semantic similarity rather than exact keyword matches.

D. FAISS Indexing and Vector Storage

After embedding generation, the system will index these vectors in **FAISS** (Facebook AI Similarity Search), an efficient library for nearest neighbor search in high-dimensional spaces.

E. Query Processing and Semantic Retrieval

When the user submits a query, the system will process it using the same embedding model used for document chunks. The query embedding will then be matched against the FAISS index to find the top-k most semantically similar chunks.

F. LangChain Integration

LangChain acts as the orchestration framework connecting all modules from document ingestion to retrieval, offering ready-made tools for loading PDFs, splitting text, generating embeddings, and FAISS-based vector search. It keeps the pipeline modular and extensible, enabling easy model swapping, future support for additional file formats, and integration of LLMs for tasks like query expansion or summarization.

Tentative Chapterization

Chapter1–Introduction

This chapter will provide the background, motivation, problem statement, objectives, and scope of the project. It will also define the significance of semantic retrieval and outline.

Chapter2–Literature Review

This section will review existing research and technologies in the field of document retrieval, semantic search, vector databases, and large language model integration.

Chapter3–Objectives and Scope of Work

This chapter define the objectives and the scope of the PDF Semantic Retrieval System. This includes, building a LangChain-based pipeline for document ingestion, text chunking, embedding generation, and FAISS powered semantic retrieval.

Chapter4–Methodology

This chapter will describe the systematic methodology followed for developing the PDF Semantic Retrieval System, including document ingestion, text chunking with context-aware splitting, embedding generation, vector indexing using FAISS, and semantic query retrieval.

Chapter5–Conclusion

The final chapter will summarize findings, highlight contributions, and reflect on the effectiveness of the proposed model.

V. CONCLUSION

The PDF Semantic Retrieval System using LangChain and FAISS is envisioned as a robust, intelligent, and efficient solution to the problem of retrieving meaningful information from large and complex PDF documents. Traditional keyword-based search mechanisms often fall short in identifying relevant information when the user’s query wording differs from the document’s terminology.

This project overcomes that limitation by leveraging semantic embeddings and vector similarity search, which interpret and match the conceptual meaning of a query rather than relying solely on exact keyword matches.

By combining LangChain for workflow orchestration with FAISS for fast and scalable vector storage, the proposed system ensures that retrieval operations remain accurate and responsive even with large datasets. The resulting solution is user-friendly, adaptable across multiple domains, and capable of supporting both academic and professional information retrieval needs.

REFERENCES

[1] Johnson, Jeff, Matthijs Douze, and Hervé Jégou. "Billion-scale similarity search with GPUs."

[2] "LangChain" *LangChain Documentation*, 2025, https://python.langchain.com.

[3] "Tesseract Open-Source OCR Engine.", https://github.com/tesseract-ocr/tesseract.

[4] "PyPDF2: A Pure Python PDF Library." *Python Package Index (PyPI)*, 2025, https://pypi.org/project/PyPDF2/.

[5] "Streamlit Documentation." *Streamlit*, 2025, https://docs.streamlit.io.