HELPMATE_AI Project Report

1. Introduction

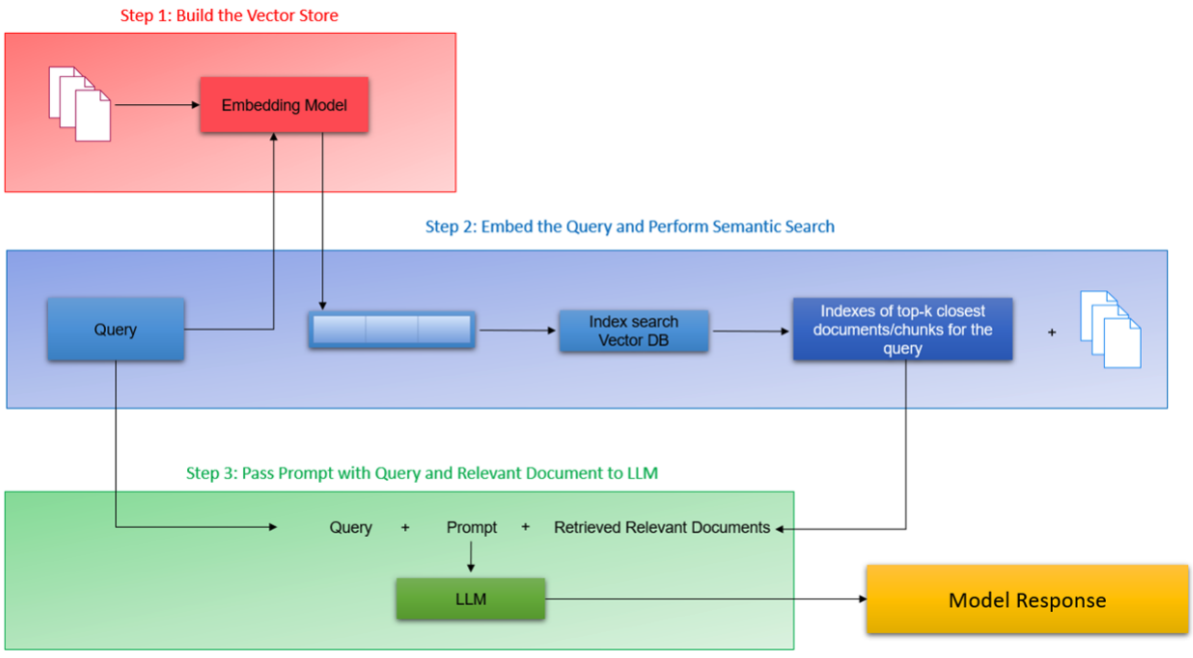
The HELPMATE_AI project aimed to develop a semantic search system utilizing a RAG (Retrieval-Augmented Generation) pipeline. The system was designed to extract information from PDF documents, generate vector representations, and implement a caching layer to improve performance.

2. Objectives

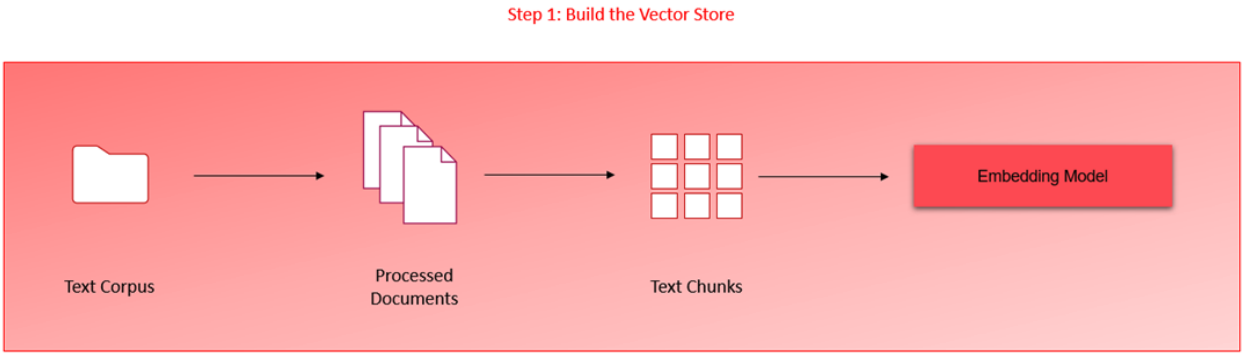
- To develop an efficient document retrieval system with a focus on semantic search.
- To generate structured data from PDF documents for indexing.
- To enhance system performance with caching of queries and results.

3. System Design

The system was designed with a three-layer RAG pipeline.

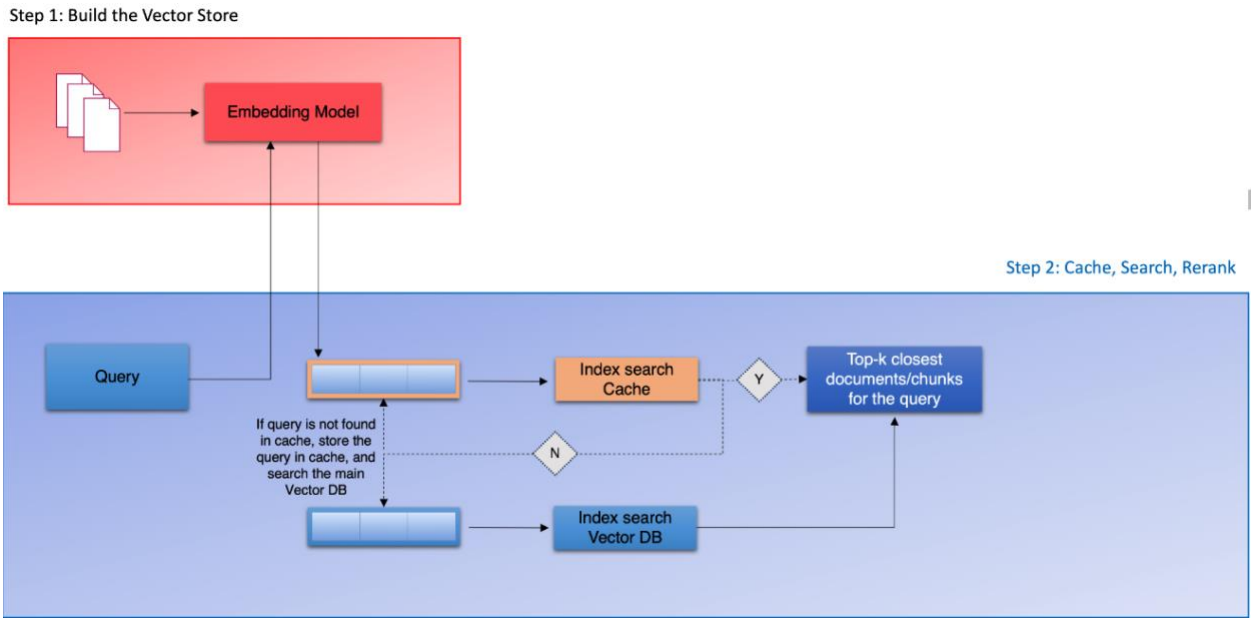


3.1. Embedding Layer



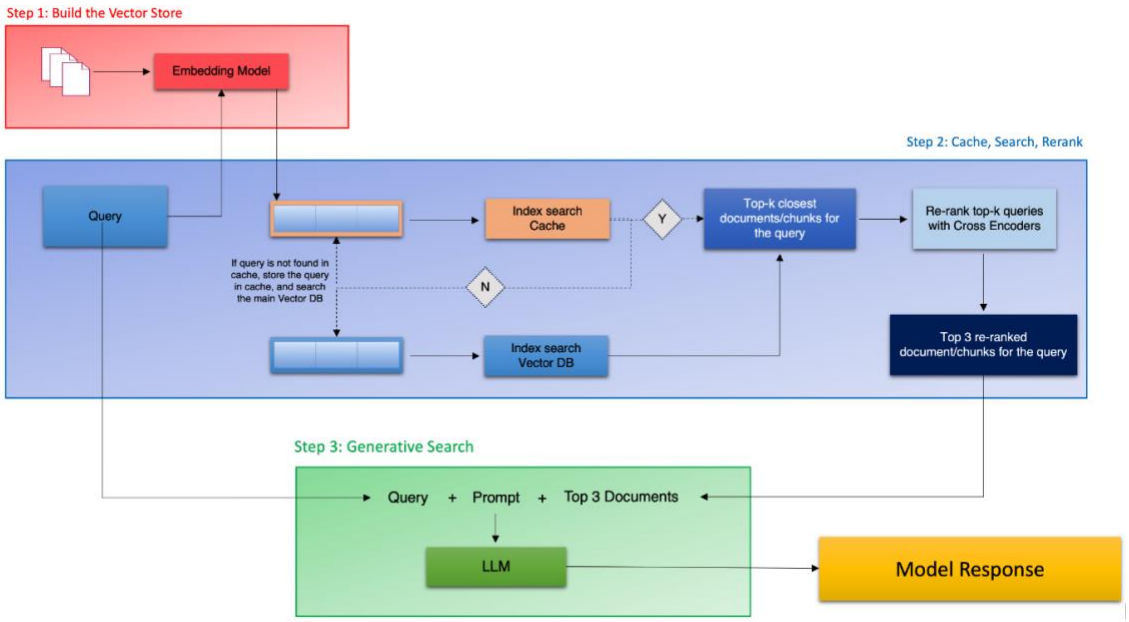
- Text and tables extracted from PDFs were converted into dataframes.
- OpenAI's text-embedding model generated vector representations stored in ChromaDB.
- Documents were processed and chunked for optimized retrieval.

3.2. Search and Rank Layer



- Semantic searches performed on queries to retrieve top K relevant documents or chunks.
- A re-ranking block utilized cross-encoders to refine search accuracy.

3.3. Generation Layer



- A well-constructed prompt, including the original query and retrieved documents, was used by a language model to generate coherent responses.

3.4. Cache Implementation

- A cache layer with a threshold for semantic similarity was integrated.
- Queries and results were stored for quick retrieval in future searches.

4. Implementation

Leveraging tools like pdfplumber, tiktoken, openai, chromaDB, and sentence-transformers, we:

- Extracted and processed PDF document content.
- Generated vector embeddings.
- Implemented the RAG pipeline for semantic search.
- Developed a cache system for efficiency.

5. Challenges and Lessons Learned

- **Document Processing:** Efficient processing is crucial, and tools like pdfplumber are essential.
- **Semantic Search:** Optimization of search parameters and thresholds is necessary for relevancy.
- **Cache Management:** Effective cache strategies are vital for system efficiency.

6. Conclusion

The project achieved its objectives, with the RAG pipeline and caching layer providing a scalable and efficient solution for document retrieval and information extraction.

The project's codebase is accessible at HelpMateAI RAG GitHub Repository.

https://github.com/MrVuTuanAnh/HELPMATE_AI/ Branches: Main