**RAG pipeline using PDF document**

## **Overview**

This Retrieval-Augmented Generation (RAG) pipeline extracts information from a PDF, retrieves relevant context based on a query, and generates answers using a large language model (LLM). It combines vector search with FAISS and text generation with a HuggingFace model.

## **Pipeline Architecture**

Chunks

Embeddings

Vectore Store

PDF Document

User Query

Query Embedding

Context Retrieval

Prompt (Query + Context)



* Input Query: User provides a question in natural language.
* Text Extraction:
  + Extracts text from specific chapters of the PDF.
  + Splits the extracted text into smaller chunks of size 500.
* Embedding Generation:
  + Encodes text chunks into numerical vectors using SentenceTransformer Embedding.
* **Vector Indexing**
* Store embeddings in a FAISS index for fast similarity search.
* Vector Search:
  + Do similarity search between the query and stored embeddings.
  + Retrieves the top-3 most relevant chunks for the query.
* Context Augmentation:
  + Combines the retrieved chunks into context for giving to LLM model.
* LLM Generation:
  + LLM model generates answer based on the given context and query.
  + Uses google/flan-t5-base LLM model for generating the response.

## **Future Enhancements**

1. Use Larger Context:
   * Enhance the retrieval mechanism to capture more relevant chunks.
   * Use reranking for getting the most relevant chunks for the query.
2. Chunk Overlap:
   * Add overlapping in chunking to ensure continuity between chunks.
3. Fine-Tune the Model:
   * Fine-tune the LLM for your dataset for improved responses.
4. **Use better LLM**

* Use larger LLMs like llama3 for better results.

1. **Use better embedding and retrieval methods**

* Try with different embedding and retrieval methods which gives better results.

1. **Improve the prompt**

* Custom prompts having main role in the LLM response.
* Currently used the basic prompting only.
* Try with different modifications on the prompt and analyze the results.

1. **Save the vector store into file**

* If the PDF data is large, then we can save the faiss vector store into a serialized format so that vector store creation need to be done at once only.
* When a new query comes, load the vector store from the saved file and do the retrieval.