Low-Level Design (LLD) Documentation for PDF Text Extraction and Question Answering API

1. Overview

This API provides two main endpoints:

1. Upload PDF: Processes a PDF document to extract text and create a vector store for efficient text retrieval.
2. Ask Question: Retrieves relevant content from the uploaded PDF in response to user queries.

2. Function Definitions

extract_text_from_pdf(pdf_bytes)

- Input: pdf_bytes - Binary data of the uploaded PDF file.
- Process:
  1. Opens the PDF file using fitz.open() with the pdf_bytes stream.
  2. Iterates through each page of the PDF.
  3. Extracts text from each page and appends it to the text string.
- Output: A single text string containing the extracted content of the PDF.
- Purpose: To obtain all the readable text from the PDF for further processing.

process_pdf_and_create_vector_store(pdf_bytes)

- Input: pdf_bytes - Binary data of the PDF.
- Process:
  1. Calls extract_text_from_pdf() to retrieve text from the PDF.
  2. Splits the extracted text into smaller chunks using CharacterTextSplitter, with a chunk size of 300 characters.
  3. Each chunk is encapsulated as a Document object for compatibility with LangChain's vector store.
  4. Uses the FAISS library to create a vector store, embedding each document with the pre-initialized HuggingFace model.
- Output: A FAISS vector store containing embeddings of text chunks.
- Purpose: To enable quick similarity search over the PDF content by embedding and storing chunks of text.

3. API Endpoints

1. **/upload/ (POST)**:
   - Functionality: Accepts a PDF file upload, extracts and processes the PDF text, and creates a vector store.
   - Process:
     - Reads PDF bytes and stores them in app.state.
     - Processes bytes with process_pdf_and_create_vector_store.

        - Returns a success message upon processing.
    - Error Handling: Raises errors if PDF reading or processing fails.
2. **/ask/ (POST)**:
    - Functionality: Accepts a question related to the uploaded PDF, retrieves the most relevant content from the vector store, and returns it as an answer.
    - Process:
        - Checks for an existing vector store in app.state.
        - Uses FAISS retriever to retrieve and return the best-matching text chunk.
    - Error Handling: Raises HTTP error if no PDF is uploaded or processed.