Developing a Legal Document Review Application with LangChain

Problem Statement

Legal professionals frequently work with lengthy and complex documents that require careful review to identify critical clauses, obligations, and risks. Manual review is time-consuming, error-prone, and expensive. An intelligent application is required to automatically extract text from legal PDFs, summarize essential information, and answer document-specific questions using Retrieval-Augmented Generation (RAG) powered by large language models.

Objectives

- Design and develop a web-based interface for uploading and reviewing legal PDF documents.
- Extract, process, and embed document text to enable efficient semantic search.
- Implement a Retrieval-Augmented Generation (RAG) pipeline to answer user queries using document context.
- Generate concise and meaningful summaries of uploaded legal documents.
- Ensure a secure, reliable, and user-friendly experience suitable for cloud deployment.

Tasks

PDF Text Extraction

- Use PyPDF2 to read and extract text content from uploaded PDF files.
- Implement robust error handling for scanned, image-based, or encrypted PDF documents.

Text Chunking and Embedding

- Split extracted text into manageable chunks using RecursiveCharacterTextSplitter.
- Generate vector embeddings using the Gemini embedding model.
- Store and manage embeddings efficiently using FAISS for fast similarity-based retrieval.

Question Answering

- Retrieve the most relevant document chunks from FAISS based on the user query.
- Provide the retrieved context and user question to Gemini 1.5 Flash through LangChain.
- Display clear, accurate, and context-aware answers to the user.

Summarization

- Generate a concise summary using the most relevant document chunks.
- Ensure the summary highlights key legal clauses, obligations, and risks in a clear format.

Configuration and Deployment

- Securely manage API keys and configuration values using environment variables.

- Support local development and enable easy deployment to a Community Cloud platform.
- Document setup, usage, and deployment steps clearly in a comprehensive README file.

Required Feature Suggestions

- PDF text preview after successful extraction.
- Semantic search-powered question answering.
- Automated and on-demand document summarization.
- Session state management for storing and reusing vector stores.
- Clear error handling and user guidance for unsupported or invalid inputs.