Assignment: Build an Enhanced RAG-Based Chatbot Using Your Own PDF Files

Objective

Build and deploy a **Retrieval-Augmented Generation (RAG)** chatbot that accepts **multiple PDF documents** as input and answers questions based on the content of these documents using a **Groq-powered LLM** and a custom Gradio-based UI.

Task Overview

Each student will:

1. **Build a working RAG chatbot**
2. **Enhance it** with at least 2 improvements (see suggestions below)
3. **Deploy it** on Hugging Face Spaces
4. **Submit** a report and link to their deployed application

Base Requirements

Students must implement the following features:

1. **Upload multiple PDF files** via Gradio
2. **Extract text** from all pages
3. **Split content into semantic chunks**
4. **Retrieve the top relevant chunks** using vector similarity (TF-IDF or other method)
5. **Send the question + context** to a Groq LLM (e.g., llama3-8b-8192)
6. **Display the answer** on the Gradio interface

Deployment Requirements

- Host the app on Hugging Face Spaces
- Include the following project files:
    - app.py
    - requirements.txt
    - apt.txt (for ffmpeg or other system dependencies if used)

Enhancement Ideas (Choose Any 2 or More)

Each student must improve their app with at least **two enhancements** from this list or propose their own:

1. **Use sentence-transformers for embeddings** instead of TF-IDF
2. **Allow PDF preview or summary** before asking questions
3. **Add conversational memory/history**
4. **Include source references or page numbers in the answers**
5. **Add support for file types other than PDF (e.g., DOCX)**
6. **Improve chunking logic (e.g., using NLTK or langchain text splitters)**
7. **Add download option for chat history**
8. **Add logging or analytics of user queries**
9. **Enable voice input/output (TTS + STT)**

Submission Deliverables

Students must submit the following:

1. Link to Hugging Face Space
2. A 1-page report (PDF or Markdown) including:
    - Overview of what was built
    - Enhancements added
    - Screenshots of the running app
    - Challenges faced

Read about what is RAG, how it works (Its main components and their working).