LAB No. 14

Document Loading using LangChain for Retrieval-Augmented Generation (RAG)

This lab introduces students to Document Loaders in LangChain, a key component of **Retrieval-Augmented Generation (RAG)** systems. Students will learn how to load, preprocess, and structure data from different document formats such as text and PDF files. By converting documents into LangChain's Document objects, students will understand how external knowledge can be prepared and supplied to large language models for improved, context-aware responses.

LAB Objectives

- Understand the role of document loaders in RAG
- Load data from multiple file formats using LangChain
- Inspect document metadata and content
- Prepare documents for downstream tasks like chunking and retrieval

Theory:

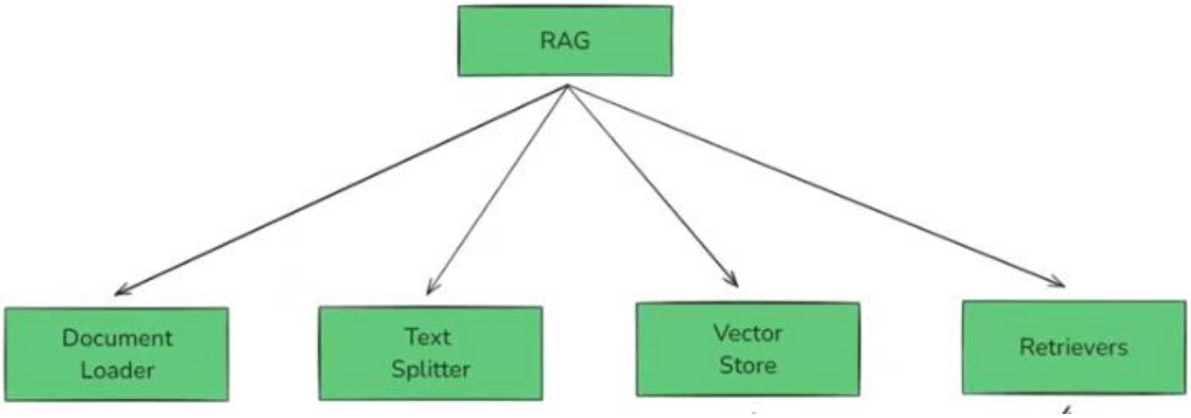
It is a technique that combines information retrieval with language generation, where a model retrieves relevant documents from a knowledge base and then uses them as context to generate accurate and grounded responses.

Benefits of using RAG

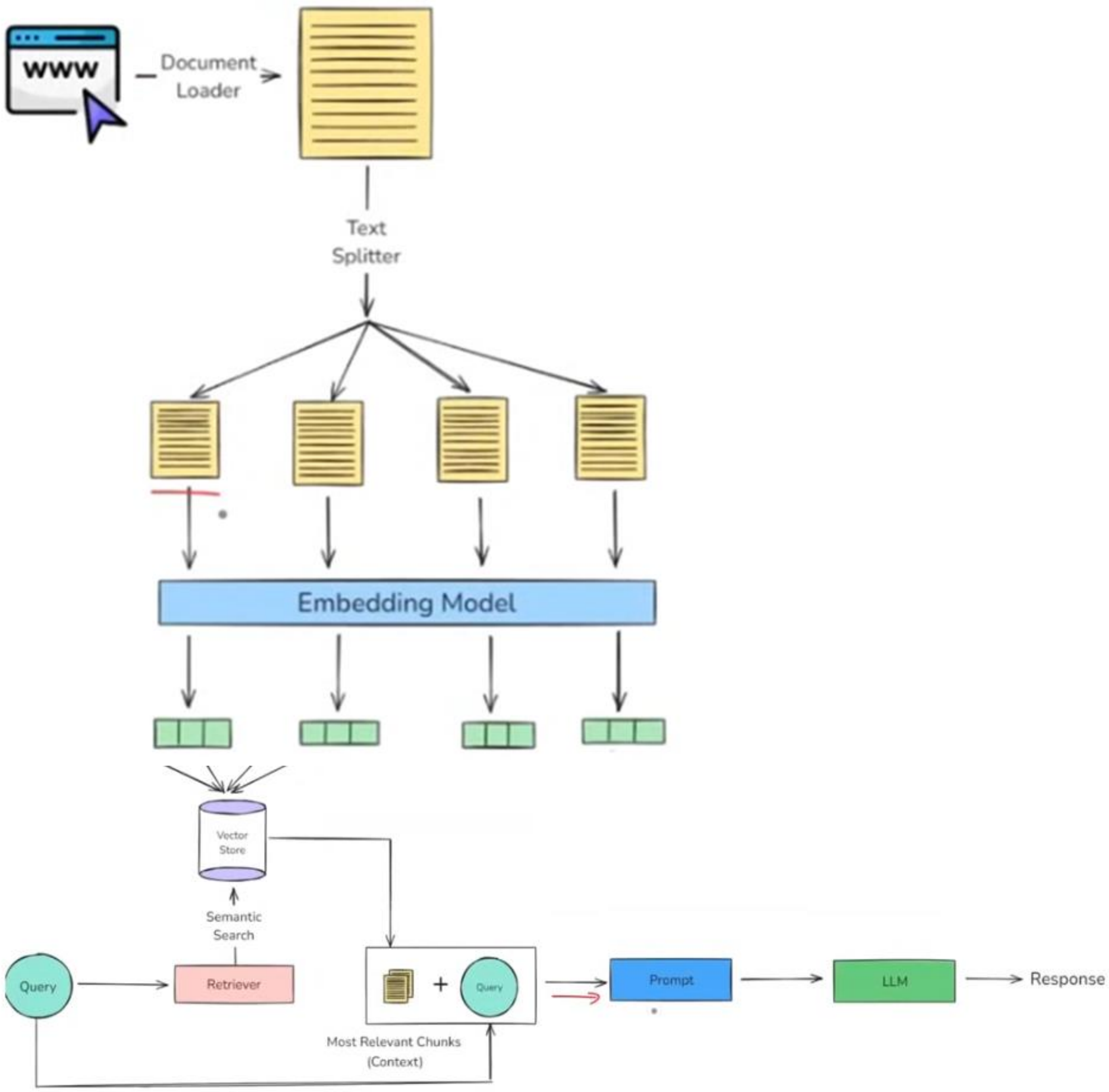
1. Use of up to date information
2. Better privacy

No limit of document size

Flow Methodology of RAG



Complete Flow diagram of RAG



Tools & Libraries

- Python 3.9+
- Required libraries:
  - langchain
  - langchain-community
  - pypdf

- unstructured

Lab Tasks (Practice Steps)

Task 1: Environment Setup

- Create a virtual environment
- Install required LangChain libraries
- Verify installation

Task 2: Understand the Main Concept – Document Loaders

- Study the role of **Document Loaders** in RAG
- Explain how loaders convert raw data into LangChain Document objects

Task 3: Load PDF Data (PyPDFLoader)

- Load lecture_notes.pdf
- Count total pages
- Display content of first page
- Attach code with output screenshot

Task 4: Load Web Data (WebBaseLoader)

- Use **WebBaseLoader** to load a webpage
- Extract main textual content
- Observe metadata (URL source)
- Attach code with output screenshot

Task 5: Load Structured Data (CSVLoader)

- Load students.csv
- Inspect how rows are converted into documents
- Print one document sample
- Attach code with output screenshot

Task 6: Compare All Loaders

Students must compare:

- Content format
- Metadata fields
- Attach code with output screenshot

Lab Questions

1. Show the working of all above lab tasks
2. What is the role of document loaders in RAG?
3. Why is metadata important in LangChain documents?
4. Difference between TextLoader and PyPDFLoader?
4. What happens if a PDF has scanned images instead of text?
5. Why is directory-based loading useful in real applications?
6. How does document quality affect RAG performance?

LAB Assessment

Student Name		**LAB Rubrics**	CLO3 , P5, PLO5
		Total Marks	10
Registration No		**Obtained Marks**	
		Teacher Name	Dr. Syed M Hamedoon
Date		**Signature**	