Coding Blocks | Online

Machine Learning Online

Assignment – 3: Retrieval-Augmented Generation (RAG) using LangChain

---

Part-I: Conceptual Understanding of RAG

Objective:

To understand the concept and need for Retrieval-Augmented Generation, its architecture, components, and relevance in modern LLM applications.

Assignment Tasks

Task 1: Short Answer Questions

Answer the following in 2–4 sentences each:

1. What is the motivation behind Retrieval-Augmented Generation (RAG)?
2. Explain the difference between RAG and standard LLM-based QA.
3. What is the role of a vector store in a RAG pipeline?
4. Compare “stuff”, “map_reduce”, and “refine” document chain types in LangChain.
5. What are the main components of a basic LangChain RAG pipeline?

Task 2: RAG Pipeline Diagram

Draw or describe the flow of a RAG system showing:

- User Query
- Retriever
- Vector Store
- LLM
- Final Answer Generation

Part-II: Practical RAG Implementation with LangChain

Objective:

To implement a working LangChain RAG pipeline using OpenAI or Ollama LLMs, FAISS or Chroma vector store, and local documents.

Assignment Tasks

Task 3: Setup LangChain RAG Pipeline

Use the following stack:

- `LangChain`
- Embeddings: OpenAI, HuggingFace or Ollama
- Vector Store: `Chroma` or `FAISS`
- LLM: Ollama (Mistral, LLaMA2, gemma2)

Steps:

- Load and split a PDF/Text document into chunks
- Convert chunks to embeddings
- Store embeddings in a vector database
- Create a retriever from the vector store
- Pass user queries through a RetrievalQA chain

Document Suggestions:
Use any academic paper, product manual, or syllabus PDF.

Task 4: Test with Queries

- Ask at least 5 questions from your document and log the answers.
- Also log the **retrieved chunks** used in each answer.
- Compare results with and without using the retriever (i.e., raw LLM vs RAG).

Task 5: Customize Prompt Template

- Modify the prompt used by the LLM to:
    - Include citations
    - Add disclaimers
    - Format answers as bullet points or structured outputs

Submission Guidelines

- Submit:
    - Jupyter Notebook or Python script
    - Sample queries and output logs (answers + retrieved documents)
    - Any local document used
    - README with installation and run instructions
    - Upload the all on github.

Coding Blocks | Online