Lab 6 - XÂY DỰNG HỆ THỐNG RAG VỚI LlamaIndex và Gemini

Format: w6_MSSV_HoTen

I. Kết quả đạt được

Kỹ năng kỹ thuật

- Xử lý đa dạng định dạng tài liệu (PDF, TXT, DOCX)
- Thiết kế và triển khai hệ thống RAG hoàn chỉnh
- Phân đoạn văn bản (chunking) hiệu quả
- Làm việc với vector embeddings và semantic search
- Sử dụng API của Gemini (LLM và embeddings)
- Tích hợp và cấu hình nhiều loại indexes khác nhau
- Xây dựng và tối ưu router query engines

Kiến thức và kỹ năng tổng quát

- Hiểu sâu về nguyên lý hoạt động của hệ thống RAG
- Quản lý metadata và cải thiện khả năng truy xuất thông tin
- Đánh giá và tối ưu hóa hiệu suất hệ thống
- Kỹ năng debugging và troubleshooting
- Tư duy thiết kế hệ thống modular
- Khả năng áp dụng RAG vào các bài toán thực tế
- Tích hợp frontend và backend trong ứng dụng AI

II. Bài tập

Yêu cầu:

- Tạo một hệ thống RAG sử dụng ít nhất 3 loại tài liệu khác nhau (PDF, TXT, và DOCX)
- Cấu hình các loại query engine khác nhau cho từng loại tài liệu
- **Có thể** xây dựng một giao diện đơn giản để đặt câu hỏi và hiển thị kết quả

Hướng dẫn:

1. Sử dụng `SimpleDirectoryReader` để tải nhiều loại tài liệu khác nhau
2. Tạo các metadata cho từng loại tài liệu (ví dụ: node.metadata["file_type"] = "pdf")
3. Sử dụng loại index khác nhau cho từng loại tài liệu
4. **Có thể** sử dụng Streamlit hoặc Gradio để tạo giao diện người dùng đơn giản hoặc dùng terminal

Đánh giá:

- Khả năng tích hợp và xử lý nhiều loại tài liệu
- Hiệu suất trả lời câu hỏi từ các nguồn tài liệu khác nhau
- Tính năng và tính dễ sử dụng của giao diện người dùng

III. Tài liệu tham khảo

- **Theory/ w7-Build RAG with LLamaIndex, Gemini, OpenAI**
- https://docs.llamaindex.ai/en/stable/module_guides/querying/router/