BÁO CÁO TIẾN ĐỘ TUẦN 3:

Trong tuần qua, em đã tìm hiểu và thực hành với các công nghệ về cơ sở dữ liệu vector, triển khai hệ thống truy vấn dữ liệu và chatbot AI. Dưới đây là tóm tắt những kiến thức đã học được:

1. So Sánh Milvus và ChromaDB

Milvus và ChromaDB là hai cơ sở dữ liệu vector hỗ trợ truy vấn dựa trên approximate nearest neighbors (ANN). Mỗt số điểm so sánh chính:

Tiêu chí	Milvus	ChromaDB
Kiến trúc	Phân tán, hỗ trợ scale	Nhẹ gọn, dễ tích hợp
Hiệu năng	Cao, hỗ trợ nhiều index	Thích hợp nhanh trong project
Khả năng truy vấn	Hỗ trợ ANN và hybrid search (BM25 kết hợp vector search), cho phép truy vấn kết hợp giữa truy vấn tìm kiếm văn bản truyền thống và truy vấn theo vector.	Tự động xây dựng metadata

Milvus thích hợp cho hệ thống quy mô lớn, trong khi ChromaDB phù hợp với các project nhanh gọn.

3. Mô Hình Embedding Trong Ollama

Ollama cung cấp các model embedding như **bge-m3** và **nomic-embed-text**

để chuyển đổi văn bản thành vector, giúp chatbot hiểu và truy vấn dữ liệu hiệu quả hơn. Em đã thử tích hợp Ollama với Milvus/ChromaDB để truy vấn theo ngữ nghĩa.

4. Triển Khai Milvus Bằng Docker trên wsl2

Sử dụng Docker giúp triển khai Milvus nhanh chóng. Các bước thực hiện:

1.Cài đặt wsl2:wsl --install

2.Cài ubuntu: wsl --install -d Ubuntu-22.04

3.Cài docker desktop

4. Chạy Milvus bằng Docker:

- Tạo file docker-compose.yml
- Chạy lệnh: docker-compose up -d để khởi tạo milvus(milvus sẽ chạy trên cổng 19530 để chấp nhận kết nối)

5. Hỏi Đáp Chatbot Với RAG

Em đã tích hợp RAG (Retrieval-Augmented Generation) với chatbot, kết hợp Ollama, Milvus/ChromaDB để truy vấn tài liệu trước khi sinh ra câu trả lời.

Bước tiến hành:

1. **Tiền xử lý**: Trích xuất nội dung tài liệu (PDF, text)
2. **Embedding**: Sử dụng nomic-embed-text tạo vector
3. **Lưu trữ**: Lưu embedding và metadata vào Milvus(ChromaDB)
4. **Truy vấn**: Khi người dùng hỏi, truy vấn embedding gần nhất
5. **Sinh câu trả lời**: Dựa trên dữ liệu tìm được

6. Kết Luận:

Việc tích hợp Milvus/ChromaDB với chatbot RAG giúp nâng cao khả năng hiểu và truy vấn dữ liệu. Việc triển khai Docker và sử dụng Attu giúp quản lý và trực quan hóa dữ liệu dễ dàng hơn.

Mục tiêu tuần tới:

1. **Tích hợp API OpenAI** vào chatbot thay thế cho model local Ollama.
2. **Tối ưu pipeline RAG** để phù hợp với OpenAI API.

3. **Dùng mongoDB** để lưu lịch sử trò chuyện với chat bot,