Giới thiệu

Mục tiêu của bài học

Bạn sẽ học cách xây dựng một hệ thống AI thông minh có khả năng:

- Tìm kiếm thông tin từ dữ liệu phi cấu trúc như file PDF hoặc ảnh.
- Trả lời câu hỏi một cách tự nhiên, giống như con người.
- Gợi ý thông minh và dẫn dắt hội thoại theo ngữ cảnh.

Ví dụ mở đầu: AI đầu bếp từ sách nấu ăn

Hãy tưởng tượng bạn đang cầm trên tay một **cuốn sách nấu ăn**, đầy những công thức ngon miệng, mẹo nấu ăn hay ho và cả những nguyên liệu bí mật.

Vấn đề đặt ra:
Làm sao để biến cuốn sách giấy này thành một trợ lý ảo mà bạn có thể hỏi:

- “Món ăn nào phù hợp với người ăn chay?”
- “Tôi hết bơ, có thể thay bằng gì?”
- “Hướng dẫn nấu súp miso là gì?”

Đó chính là điều bạn sẽ làm được sau bài học này.

Giới thiệu về RAG - Retrieval-Augmented Generation

RAG là mô hình kết hợp 2 thành phần:

1. **Retrieval (truy xuất)**: tìm kiếm thông tin từ nguồn dữ liệu lớn như tài liệu, ảnh quét, email, báo cáo...
2. **Generation (sinh)**: sử dụng mô hình ngôn ngữ (như GPT) để tổng hợp, trả lời và dẫn dắt hội thoại.

Kết quả?
Một hệ thống AI **biết tìm kiếm và hiểu rõ ngữ cảnh**, như một chuyên gia thực thụ.

Các phần học chi tiết

1. Data Conversion Mastery – Làm sạch và chuẩn hóa dữ liệu

- Chuyển đổi dữ liệu từ PDF, hình ảnh, hoặc văn bản thô sang định dạng AI có thể hiểu được.
- Công cụ: `PyMuPDF`, `pdfplumber`, `Tesseract`, v.v.

Ví dụ thực tế:
Chuyển một thực đơn nhà hàng từ ảnh chụp sang bảng Excel chứa tên món, mô tả, giá tiền, danh mục.

2. **OCR nâng cao với GPT**

- Sử dụng GPT để thực hiện OCR (Optical Character Recognition).
- Không chỉ nhận diện chữ, mà còn hiểu và **trích xuất dữ liệu có cấu trúc** (ví dụ: bảng, danh sách, bảng giá).

Ví dụ:
Trích xuất dữ liệu từ bảng PDF báo cáo doanh thu và biến thành JSON hoặc bảng Excel có thể tra cứu.

3. Tạo Embeddings và Lưu trữ thông minh với FAISS

- Tạo vector biểu diễn nội dung (embeddings) bằng OpenAI API.
- Lưu trữ trong FAISS – một hệ thống tìm kiếm tương đồng theo ngữ nghĩa.
- Cho phép truy xuất **ngay cả khi truy vấn không trùng khớp từ khóa**.

Ví dụ:
Bạn hỏi “món ăn không có gluten” → AI tìm món phù hợp dù không có cụm từ “gluten-free” trong dữ liệu.

4. Xây dựng hệ thống RAG hoàn chỉnh

- Kết hợp truy xuất và sinh văn bản.
- Dữ liệu đầu vào: file ảnh, PDF, tài liệu khách hàng, tin nhắn...
- Kết quả đầu ra: câu trả lời có dẫn chứng, thông minh và theo ngữ cảnh.

Công nghệ sử dụng:

- OpenAI API (`text-embedding-3-small`, `gpt-4`)
- FAISS
- LangChain (nếu mở rộng)
- FastAPI (để triển khai)

5. Prompt Engineering & Fine-tuning

- Thiết kế prompt giúp AI phản hồi chính xác, có kiểm soát.
- Tuỳ chỉnh hệ thống để phù hợp với ngữ cảnh riêng (ví dụ: hỗ trợ kỹ thuật, chăm sóc khách hàng, tra cứu văn bản pháp lý...).

Ví dụ:
Thêm hướng dẫn vào prompt như: “Trả lời bằng giọng điệu thân thiện, sử dụng tiếng Việt đơn giản.”

Kết quả cuối cùng

Bạn có thể:

- Đưa hàng trăm tài liệu PDF, hình ảnh, Excel vào hệ thống.
- Xây dựng trợ lý ảo cho doanh nghiệp, nhà hàng, thư viện hoặc cá nhân.
- Tạo chatbot hỗ trợ khách hàng dựa trên thông tin nội bộ doanh nghiệp.
- Biến dữ liệu tĩnh thành công cụ tương tác, nhanh chóng và thông minh.

Tóm tắt

Kỹ năng	Mô tả
Xử lý dữ liệu	Chuyển đổi dữ liệu thô sang định dạng AI
OCR nâng cao	Nhận diện & trích xuất dữ liệu từ PDF, ảnh
Embedding	Biểu diễn dữ liệu bằng vector để tìm kiếm theo ngữ nghĩa
RAG	Truy xuất + sinh văn bản từ dữ liệu thật
Prompt Engineering	Tinh chỉnh phản hồi AI

Tác giả: Đỗ Ngọc Tú

Công Ty Phần Mềm VHTSoft

Phiên bản #2
Được tạo 5 tháng 5 2025 14:01:52 bởi Đỗ Ngọc Tú
Được cập nhật 6 tháng 5 2025 09:52:34 bởi Đỗ Ngọc Tú