

Khai phá dữ liệu phi cấu trúc với Retrieval-Augmented Generation (RAG)

Hãy tưởng tượng bạn đang phải đối mặt với hàng tá tài liệu, báo cáo dài dòng, hợp đồng, email, hay các tệp PDF, Word, PowerPoint... và bạn chỉ cần một mảnh thông tin nhỏ ẩn sâu bên trong đó. Việc tìm kiếm, tóm tắt hoặc phân tích những dữ liệu này theo cách thủ công thật sự mất thời gian và gây mệt mỏi.

Đây chính là thách thức của dữ liệu phi cấu trúc – những loại dữ liệu không tuân theo định dạng hàng cột quen thuộc như trong cơ sở dữ liệu hay bảng tính. Và đó cũng là lý do chúng ta cần đến RAG.

RAG (Retrieval-Augmented Generation) là giải pháp giúp bạn:

- **Tự động truy xuất thông tin** từ các tệp không có cấu trúc.
- **Tóm tắt nội dung** một cách thông minh.
- **Trả lời câu hỏi** dựa trên ngữ cảnh thực tế từ dữ liệu bạn cung cấp.

Trong chương này, bạn sẽ:

Làm quen với thư viện **LangChain** – công cụ chính giúp xử lý tài liệu phi cấu trúc.

Học cách xử lý nhiều loại dữ liệu: Excel, Word, PowerPoint, PDF, EPUB...

Xây dựng hệ thống **retrieval** từ các tài liệu này.

Tùy chỉnh các hàm để **truy xuất và sinh câu trả lời có ý nghĩa từ dữ liệu thực tế**.

Kết quả sau chương học:

Bạn sẽ sở hữu **bộ công cụ hoàn chỉnh** để làm việc với dữ liệu phi cấu trúc: từ việc tải, chia nhỏ văn bản, đến truy xuất thông tin và tạo nội dung có giá trị. Tất cả được áp dụng trong các **bài tập thực tế, nhiều coding và ví dụ minh họa rõ ràng**.

“ Dữ liệu phi cấu trúc có mặt ở khắp nơi - email, báo cáo, mạng xã hội, sách điện tử... và khả năng khai thác nó không chỉ là kỹ năng kỹ thuật, mà là một siêu năng lực.

Hãy cùng bắt đầu hành trình này!

Phiên bản #2

Được tạo 6 tháng 5 2025 09:57:36 bởi Đỗ Ngọc Tú

Được cập nhật 6 tháng 5 2025 09:59:02 bởi Đỗ Ngọc Tú