

Đọc và xử lý dữ liệu Excel với LangChain

Trong bài học này, chúng ta sẽ:

1. Đọc dữ liệu từ file Excel
2. Phân tích và hiển thị dữ liệu
3. Chia nhỏ dữ liệu thành các **chunk** để chuẩn bị cho việc tạo **embeddings**

Bước 1: Đọc dữ liệu Excel

Chúng ta sẽ sử dụng **UnstructuredExcelLoader** từ `langchain_community` để đọc file `reviews.xlsx`. Đây là tập dữ liệu chứa các **đánh giá và bình luận** đến ngày **21 tháng 8 năm 2024**.

```
from langchain_community.document_loaders import UnstructuredExcelLoader

loader = UnstructuredExcelLoader("reviews.xlsx", mode="elements")
docs = loader.load()
```

Ghi chú: `mode="elements"` giúp tách nội dung trong Excel thành các phần tử nhỏ như từng dòng, từng ô - điều này giúp xử lý linh hoạt hơn.

Bước 2: Hiển thị dữ liệu

Hiển thị 5 phần tử đầu tiên để kiểm tra kết quả:

```
docs[:5]
```

Bạn sẽ thấy một số dữ liệu hiển thị như `TD`, `TR` - điều này thể hiện dữ liệu gốc được biểu diễn theo dạng bảng.

Bước 3: Chia nhỏ văn bản (Chunking)

Tại sao cần chunk?

- Dữ liệu lớn sẽ khó xử lý một lần
- Embedding có giới hạn độ dài token (vd: 4096 tokens)
- Chunk nhỏ giúp dễ dàng truy vấn và tìm kiếm hơn

Thực hiện chia nhỏ

```
from langchain.text_splitter import RecursiveCharacterTextSplitter

text_splitter = RecursiveCharacterTextSplitter(
    chunk_size=2000,
    chunk_overlap=200
)

chunks = text_splitter.split_documents(docs)
```

Hiển thị 5 chunk đầu tiên:

```
chunks[:5]
```

Mẹo:

- `chunk_size=2000` có thể thay đổi tùy vào độ dài dữ liệu
- Nếu file rất lớn (nhiều triệu dòng), bạn nên dùng chunk nhỏ hơn hoặc chia theo nội dung logic hơn (theo tiêu đề, đoạn...)

Ghi chú về dữ liệu

Khi hiển thị chunk, có thể bạn sẽ thấy các thẻ như `<td>` hoặc `<tr>`:

- `<td>`: thể hiện ô trong bảng
- `<tr>`: thể hiện hàng trong bảng

Điều này là bình thường khi xử lý dữ liệu dạng bảng – bạn có thể lọc hoặc xử lý thêm nếu muốn dữ liệu "sạch" hơn.

Tổng kết

Bạn vừa học cách:

- Tải dữ liệu từ file Excel bằng LangChain
- Hiển thị một phần dữ liệu
- Chia dữ liệu thành các phần nhỏ để chuẩn bị cho bước tiếp theo

Tác giả: Đỗ Ngọc Tú
Công Ty Phần Mềm VHTSoft

Phiên bản #1

Được tạo 6 tháng 5 2025 15:45:59 bởi Đỗ Ngọc Tú

Được cập nhật 6 tháng 5 2025 15:50:34 bởi Đỗ Ngọc Tú