

Thiết lập môi trường xử lý dữ liệu không có cấu trúc với LangChain

Trong phần này, chúng ta sẽ **thiết lập môi trường làm việc** để xử lý **dữ liệu không có cấu trúc** bằng thư viện **LangChain** và các công cụ liên quan.

Chúng ta sẽ thực hiện điều này thông qua **Google Colaboratory**, một công cụ tuyệt vời để viết và chạy mã Python trực tuyến, đồng thời dễ dàng tích hợp với **Google Drive**.

Bước 1: Tạo môi trường làm việc trên Google Colab

- Mở thư mục `unstructured_data` trong dự án `rack`
- Nhấn chuột phải → **New More** → **Google Collaboratory**

Đây sẽ là nơi bạn viết các đoạn mã để xử lý dữ liệu.

Bước 2: Cài đặt các thư viện cần thiết

Thư viện chính:

- `langchain` → Xử lý dữ liệu và tạo pipeline thông minh
- `langchain-community` → Loader cho các định dạng không cấu trúc như Excel, PDF, EPUB...
- `openai` → Tạo embeddings, LLM
- `faiss-cpu` → Dùng để xây dựng hệ thống truy xuất vector (retrieval system)

```
!pip install langchain-community langchain openai faiss-cpu
```

Bước 3: Thiết lập API Key cho OpenAI

Sử dụng Google Colab, bạn có thể lưu API Key dưới dạng dữ liệu người dùng:

```
from google.colab import userdata

OPENAI_API_KEY = userdata.get('janai_course')
```

Bước 4: Kết nối với Google Drive

Google Drive sẽ là nơi chứa các file dữ liệu như `.xlsx`, `.pdf`, `.docx`, `.epub`...

```
from google.colab import drive
drive.mount('/content/drive')
```

Sau đó bạn đổi thư mục làm việc sang thư mục dữ liệu:

```
%cd /content/drive/MyDrive/your-folder-path
```

Bước 5: Import các module quan trọng

```
# Load dữ liệu Excel
from langchain_community.document_loaders import UnstructuredExcelLoader

# Chia nhỏ dữ liệu
from langchain.text_splitter import RecursiveCharacterTextSplitter

# Gọi LLM và tạo Embeddings
from langchain_openai import ChatOpenAI, OpenAIEmbeddings

# Vector store - FAISS
from langchain.vectorstores.faiss import FAISS

# Hiển thị Markdown trong notebook
from IPython.display import display, Markdown
```

Tổng kết

Trong bài giảng này, bạn đã:

- Tạo Google Colab notebook cho dự án
- Cài đặt các thư viện cần thiết cho xử lý dữ liệu không cấu trúc
- Kết nối API OpenAI
- Liên kết với Google Drive để truy cập dữ liệu
- Import các thành phần chính từ LangChain

Tác giả: Đỗ Ngọc Tú
Công Ty Phần Mềm VHTSoft

Phiên bản #1

Được tạo 6 tháng 5 2025 15:23:16 bởi Đỗ Ngọc Tú

Được cập nhật 6 tháng 5 2025 15:31:24 bởi Đỗ Ngọc Tú