Local Al agent with Ollama

Introduction

This complete code workflow sets up a chatbot that uses LangChain, Ollama, and Chroma to answer questions about a pizza restaurant by retrieving and summarizing relevant reviews.

How the Solution Works

- Review Data & Embeddings: Loads reviews from restaurant_reviews.csv, embeds them using Ollama's "mxbai-embed-large" model, and stores them in a Chroma vector database.
- Retriever Setup: Creates a retriever to find the most relevant review when asked a question, supporting retrieval-augmented generation.
- LLM & Prompt Chain: Uses <code>Ollamallm</code> with "llama3.2" to process questions alongside retrieved reviews via a conversational chain.
- Chat Loop: Allows repeated user interaction; for each question, finds relevant reviews and provides an expert answer using the LLM model and prompt template.

Main Modules Used

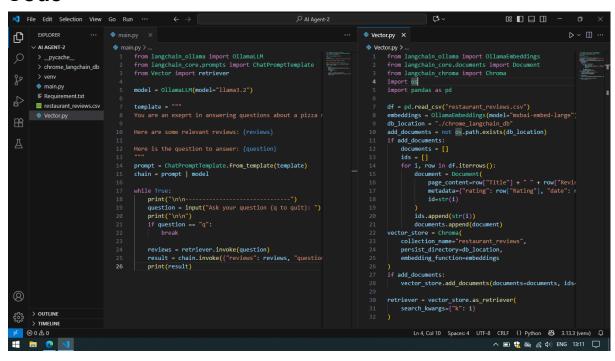
- langchain_ollama:
 - OllamaLLM: Handles large language model responses.
 - OllamaEmbeddings: Generates embeddings for review texts.
- langchain_core:
 - ChatPromptTemplate: Manages custom prompts for the chatbot.
 - Document: Wraps reviews as structured documents.
- langchain_chroma:
 - Chroma: Provides persistent vector database capabilities.
- Pandas:
 - Handles CSV data loading and processing.

Workflow Summary

- 1. Preprocessing:
 - Reviews converted to documents with metadata (rating, date).
 - Documents embedded and stored in Chroma if not already present.
- 2. Interactive Chat:
 - Receives a user question.
 - Retrieves most relevant review(s) using the retriever (semantic search).

- Combines reviews with the question using a custom prompt.
- Sends to the LLM model for expert answer generation.
- 3. Continuous Interface:
 - Users can ask multiple questions until quitting.

Code



Output

