HARNESSING LARGE LANGUAGE MODELS FOR GRAPH GENERATION

Date: 6/5/2024
Authors:
Aniket Poojari, Sahil Pamnani

Overview The graph generation pipeline leverages OpenAI's GPT-3.5 to extract entities and relationships from different types of documents, including project briefs, people's profiles, and Slack messages. This pipeline involves several key steps, from reading the files to generating prompts and processing responses. Below is a detailed report on the code structure, execution, and outcomes.

1. Settings

- **Environment Variables:** Load environment variables using dotenv to configure the OpenAI API.
- **OpenAI Configuration:** API type, key, base, and version are set up for making requests.

2. Helper Functions

process_gpt

- **Purpose:** Sends a prompt to the OpenAI API and retrieves the response.
- **Parameters:**
  - file_prompt: The text prompt sent to the API.
  - system_msg: The system message providing context to the API.
- **Returns:** The content of the API response after a brief sleep to avoid rate limits.

extract_entities_relationships

- **Purpose:** Processes all files in a specified folder to extract entities and relationships.
- **Parameters:**
  - folder: Folder containing the files to be processed.
  - prompt_template: Template used for generating the prompt for each file.
- **Returns:** A list of JSON objects containing extracted entities and relationships.
- **Process:**
  - Reads files from the specified folder.
  - Uses a predefined prompt template to create prompts for the API.
  - Calls process_gpt for each file and appends the results to a list.

ingestion_pipeline

- **Purpose:** Combines extracted entities and relationships from multiple folders.
- **Parameters:**
  - folders: A dictionary where keys are folder names and values are corresponding prompt templates.
- **Returns:** A combined list of JSON objects from all folders.

- **Process:**
  - Iterates through each folder and its prompt template.
  - Calls extract_entities_relationships for each folder and combines the results.

3. Defining Prompts

Project Prompt Template

- Extracts Project, Technology, and Client entities from project briefs.
- Generates relationships: project|USES_TECH|technology and project|HAS_CLIENT|client.

People Prompt Template

- Extracts Person, Project, and Technology entities from people's profiles.
- Generates relationships: person|HAS_SKILLS|technology and project|HAS_PEOPLE|person.

Slack Prompt Template

- Extracts Person and SlackMessage entities from Slack messages.
- Generates relationships: personid|SENT|slackmessageid.

4. Running the Pipeline

The ingestion_pipeline function is executed with the defined folders and their respective prompt templates.

Execution Summary

- **Files Processed:**
  - **People Profiles:** Extracts entities and relationships from people's profiles.
  - **Project Briefs:** Extracts entities and relationships from project briefs.
  - **Slack Messages:** Extracts entities and relationships from Slack messages.
- **Performance:**
  - The total time taken for the pipeline execution is logged.
- **Output:**
  - A list of JSON objects containing extracted entities and relationships.

5. Conclusion

The graph generation pipeline effectively extracts structured information from various unstructured text documents using OpenAI's GPT-3.5. It
automates the process of identifying entities and relationships, making it a powerful tool for knowledge extraction and management.