Bank of England NLP Project

This repository contains a set of Jupyter notebooks that process Q&A sections from PDF transcripts and perform Retrieval-Augmented Generation (RAG) on the extracted data.

Requirements

Below is a list of Python packages and dependencies used throughout these notebooks. Install any missing dependencies using `pip install <package-name>` or via `conda install <package-name>` if using Anaconda. Adjust versions and additional libraries as needed.

- **Python 3.7+**
- **Jupyter Notebook** or **JupyterLab**
- **pandas** (data manipulation and analysis)
- **numpy** (numerical computing)
- **PyPDF2** (PDF handling, reading/writing)
- **pdfplumber** (alternative PDF extraction tool)
- **openai** (OpenAI's API for embeddings, chat completions)
- **tiktoken** (token counting/encoding for OpenAI models)
- **mlflow** (experiment tracking and logging)
- **faiss** (vector similarity search for RAG)
- **python-dotenv** (to load environment variables from .env files)
- **requests** (HTTP requests, if needed for data fetching)
- **dotenv** (part of python-dotenv for environment management)
- **time**, **os**, **sys**, **json**, **csv**, **re**, **hashlib**, **typing**, **pathlib**, **pickle**, **datetime** (all standard library modules)

Notebooks

1. **pdf_QnA_section_extractor.ipynb**

   - **Purpose**: Extracts the Q&A portion from PDF transcripts.
   - **Output**: Saves the extracted Q&A sections as individual files in the `Extracted` folder.

2. **Q&A_pdf_to_json.ipynb**

   - **Purpose**: Takes the extracted Q&A files, applies a prompt-based approach (for example, summarization or question-answering), and converts the results into JSON.
   - **Output**: Stores the resulting JSON files in the `Processed` folder.

3. **JSON_page_number_update_folder_based.ipynb**

   - **Purpose**: Reads the processed JSON files, checks for the correct page numbers (using the Q&A start page), and updates each JSON file with accurate page references.

4. **rag_stable_output.ipynb**

  - **Purpose**: Performs the final Retrieval-Augmented Generation step. It picks up the updated JSON files, runs queries against them, and demonstrates table-like outputs for the results.

Workflow Summary

1. **Extract Q&A**
   Run `pdf_QnA_section_extractor.ipynb` to split transcripts and collect Q&A segments.

2. **Convert to JSON**
   Use `Q&A_pdf_to_json.ipynb` to process extracted Q&A files and output them in JSON format.

3. **Update Page Numbers**
   Execute `JSON_page_number_update_folder_based.ipynb` to ensure each Q&A section has the correct page numbers.

4. **Run RAG**
   Finally, run `rag_stable_output.ipynb` to conduct retrieval-augmented queries, demonstrating how to effectively query the Q&A data.