Lab 4 Part 1 - Reddit API Data Extraction

Important Files and Folders

Make sure files are present in to root directory of the project

1. `.env` - File with environment variables
2. `ca.pem` - Certifcate Authentication file for database
3. `requirements.txt` - File with dependencies to installed for the project
4. `src/models/` - Has pretrained embedding generator model and clustering model
5. `docs/` - Documentation including minutes of the meeting and readme in PDF format

Setup

- For setup move into `reddit-scraper` directory

```
cd reddit-scraper
```

- Create and active a virtual environment

```
virtualenv venv
source venv/bin/activate  # or "venv\Scripts\activate" on Windows
```

- To install the necessary libraries

```
pip install -r requirements.txt
```

Running the script

- Make sure the all the important files listed above are present at the correct location
- Virtual Environment is created and activated
- To run the script
    - Move into the `/src` directory

    ```
    cd src
    ```

    - Run the `main.py` script

    ```
    python main.py <time-in-minutes>
    # Example: python main.py 3
    ```

Exceution Flow

Upon execution, the script follows the following steps:

1. Prompt for Input String:
    - The script asks the user to input a query string or the command 'quit'.
2. Check for Updates:
    - First, it checks for any updates by scraping Reddit posts from the tech subreddit page.
    - If new posts are found, it processes and stores them in the database.
    - If no new posts are found, it logs a message indicating no updates.
3. Wait for User Input or Time Interval:
    - After processing updates or checking for updates, the script waits for a specified time interval (e.g., 5 minutes).
    - During this waiting period, it remains open for user input.
    - If the specified time elapses without any input from the user, the script resumes scraping Reddit posts.
4. Handle User Input:
    - If the user inputs 'quit', the program exits gracefully.
    - If the user inputs a query string (e.g., 'pubg battlegrounds mobile india'), the script returns the top 10 similar documents and keywords matching the query string.
    - If no input is provided during the waiting period, the script resumes scraping Reddit posts.

About Scripts

File Name	Purpose
`crud.py`	Insert information of scraped posts in bulk into the database, discarding posts that already exist.
`database.py`	Establish a connection between Python and the SQL server.
`doc2vec.py`	Train a Gensim model to create embeddings for documents and calculate document similarity.
`extract.py`	Clean and preprocess text from scraped Reddit posts and linked websites, and extract top keywords characterizing each document.
`main.py`	Driver program for scraping Reddit posts, preprocessing data, storing it in the database, and providing functionality to search for similar documents.
`model.py`	Define schema for data storage in the MySQL server. Create tables is doesn't exist.
`schema.py`	Create Pydantic model to validate data format before storing it in the database.
`settings.py`	Set up access to certificates and environment credentials required for connecting to the MySQL database.
`clustering.py`	Train and save clustering model, inferences for interested records