Turmerik ML Home Assignment

This project aims to showcase proficiency in ethical web scraping, data analysis with sentiment analysis, and the integration of AI for personalized communication. This readme provides guidelines for setting up and running the code, a brief discussion of the methodology, examples of collected data, and an overall data analysis along with generated personalized messages.

Author

- Vatsal Thakkar (vatsalthakkar3.vt@gmail.com)

Requirements

Hard Requirements

1. Python (3.11.X)
2. Reddit Account

Soft Requirements

1. Miniconda
2. VS Code (code editor)
3. SQLite (Extension for VSCode)
4. SQLite Viewer (Extension for VSCode)
5. macOS 14.4.1

Setup Steps

Install Python Dependencies

If you prefer to use Miniconda for managing Python environments, follow all the steps (1-5) outlined. However, if you opt not to use Miniconda, make sure you have Python version 3.11 installed, and you can skip steps 1-4, proceeding directly to step 5. I recommend using Miniconda for smoother Python environment management.

1. **Download Miniconda**: Visit the Miniconda website and download the appropriate installer for your OS.

2. **Install Miniconda**: Follow the installation instructions provided on the Miniconda website after downloading the installer.

3. **Set up Conda Environment**:

```
conda create --name turmerik_ml python=3.11
```

4. **Activate the Environment**:

```
conda activate turmerik_ml
```

5. **Install Dependencies**:

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

This command installs all necessary packages listed in `requirements.txt` into your Python environment.

Setting up the Environment Variables

To setup the environment variables for your application I prefer using the `.env` configuration file to store all the environment variable and keep the environment variables in the same place and private

1. **Create a `.env` File**: Start by creating a file named `.env` in the root directory of your application.

   Write the following command in the terminal

```
touch .env # Shell command to create a file named `.env`
```

2. **Define Environment Variables**: Inside the `.env` file, define your environment variables using the KEY=VALUE format. I have provide the `.env.example` file containing the example variables you want to define. You can just copy and paste the example variables and set your environment variables.

```
   # REDDIT API
   REDDIT_SECRET_KEY="" # Write Your Reddit Secret Key in quotes
   REDDIT_CLIENT_ID=""  # Write Your Reddit Client ID in quotes
   USER_AGENT=""        # Write Your Reddit User Agent in quotes
   REDDIT_USERNAME=""   # Write Your Reddit Username in quotes (Not
Required)
   REDDIT_PASSWORD=""   # Write Your Reddit Password in quotes (Not
Required)

   # OPENAI_API_KEY
   OPENAI_API_KEY=""    # Write Your OpenAI API Key in quotes
```

Note: For using the Reddit API you require the Reddit account and obtain your the Reddit API key (Client ID and Secret Key) from your account. You can follow the following steps if you dont have the Reddit Secret Key.

Obtaining Reddit Client ID and Secret Key

Follow these steps to obtain your Reddit client ID and secret key for authenticating with the Reddit API:

1. **Create a Reddit Account**: Sign up for a Reddit account if you don't already have one.
2. **Visit Reddit Developer App Settings**: Log in to your Reddit account and go to Reddit App Preferences.
3. **Create a New Reddit App**: Scroll down to the "Developed Applications" section and click on the "Create App" button.
4. **Fill Out the App Creation Form**: Choose the type of app you're creating, enter a name, and provide a redirect URI you can use http://localhost:8080 for testing purposes (if applicable).
5. **Obtain the Client ID and Secret Key**: After creating the app, you'll see your app's details, including the "client ID" and "secret key" (client secret).
6. **Include Client ID and Secret Key in Your Application**: Use these keys to authenticate with the Reddit API in your application. Ensure to keep them confidential and avoid hardcoding them directly in your source code.
7. **Usage**: Refer to your application's code or documentation for instructions on how to use the Reddit client ID and secret key for API authentication.

Note: Keep your client ID and secret key secure and do not share them publicly or expose them in your version control system.

🎉 Now You are ready to run the code but before that just get the overview of the procedure.

Methodology Overview

1. **Data Collection**
   - Use the PRAW library to scrape Reddit comments and related posts from specified subreddits.
   - Store comment data (comment ID, post ID, comment author, comment body, reply-to comment ID) in a SQLite database.
   - Store post data (post ID, title, author, content) in the same SQLite database.

Database Schema

Table: posts

Field	Type	Description
post_id	TEXT	Primary key, unique identifier for each post
title	TEXT	Title of the post
author	TEXT	Name of the author of the post
content	TEXT	Content of the post

Table: comments

Field	Type	Description

Field	Type	Description
comment_id	TEXT	Primary key, unique identifier for each comment
post_id	TEXT	Foreign key referencing the related post
comment_author	TEXT	Name of the author of the comment
comment_body	TEXT	Content of the comment
reply_to_comment_id	TEXT	ID of the comment being replied to (if any) or Post id

2. **Related Post Retrieval**
    - Retrieve the related post for each comment using the parent ID from the SQLite database.
    - Fetch post content (title and self-text if available) to include in personalized messages.
3. **Sentiment Analysis**
    - Lot of Research such as Sentiment Analysis in the Era of Large Language Models: A Reality Check shows that LLMs are able to understand the sentiments of the input and provide the repose based on that so we do not need a separate sentiment analysis model but we can just use LLM to get the sentiment of the comment and based on that context and then that LLM will generate the personalize message.
4. **Message Generation**
    - Use the OpenAI API (GPT-3) for generating personalized messages sensing the sentiment of the comment directly.
    - Input comment content, related post content into GPT-3 to generate engaging messages.
    - You can find the details regarding prompt and system prompt which are used form `src/generate_message.py` file.
5. **Output Generation**
    - Generate a JSON file containing personalized messages for each commenter.
    - Include author ID, comment content, whether it's a reply to a post or a comment, and the personalized message.

How to run the code

All the code is provided in the `src/` directory. It will contain the following files.

```
src
├── __init__.py
├── generate_message.py         # Used To Generate the personalized
Message Form the
└── scrape_data.py              # Used to scrape comments and Posts from
Reddit for particular topic
```

1. **Running `scrape_data.py`**: This python script will scrape the comments and posts related for the particular topic and store them in the SQLite database (`reddir_data.py`)

```
# Details regarding the Usage. All the arguments are optional but can
be overridden
usage: scrape_data.py [-h] [--topic TOPIC] [--limit LIMIT] [--depth
{shallow,deep}]

Scrape Reddit data

options:
  -h, --help            show this help message and exit
  --topic TOPIC         Topic to search for eg. "clinical trial"
  --limit LIMIT         Number of posts to scrape per subreddit eg. 5
  --depth {shallow,deep}
                        Search depth for subreddits default is shallow
```

Write the following line in terminal to start scraping.

```
 python src/scrape_data.py # You can add arguments and override the
default arguments for different topic

  #OR

 python src/scrape_data.py --topic "clinical trial" --limit 5 --depth
shallow
```

   - After Running this script SQLite database named `reddit_data.db` will be created in the root directory of the project. You can look at the database if you are using the VS Code extension `SQLite` and `SQLite Viewer`

Example Database:

Table: posts

post_id	title	author	content
omqnox	Clinical Trials Discussion Thread - Week of 2021-07-18	ClinicalTrialsBot	Here you can talk about specific clinical trials, random studies or experiences with them, or the clinical trials process like recruitment, compensation etc.

Table: comments

comment_id	post_id	comment_author	comment_body	reply_to_comment_id

comment_id	post_id	comment_author	comment_body	reply_to_comment_id
h7k4j6a	omqnox	RachelRei	My trial has had a rash of people canceling their study day appointments last minute! This costs me $200 in cancellation fees. I can't charge them these fees since they are volunteering their time. Any suggestions on steps to take? We already send out plenty of reminders and communicate openly.	omqnox

2. **Running** `generate_message.py`: This Python script will create a JSON file containing the personalized messages to the commenter based on their comments Using the LLM.

   Write following line in the terminal to run the code.

   ```
   python src/generate_message.py
   ```

   Example Output in JSON File :


   ```
   {
       "author_id": "h7k4j6a",
       "comment": "RachelRei",
       "is_reply_to_post": true,
       "personalized_message": "Hi RachelRei,\n\nI'm sorry to hear about
   the challenges you're facing with last-minute cancellations in your
   clinical trial. It's understandable that this can be frustrating,
   especially when it results in financial losses. Have you considered
   exploring strategies to improve participant engagement and commitment,
   such as providing additional support or incentives to encourage
   attendance?\n\nClinical trials are crucial for advancing medical
   research and improving healthcare outcomes, and your dedication to
   managing these challenges is commendable. Your experience highlights
   the importance of continuous improvement in the clinical trial
   process. If you're open to it, participating in a clinical trial
   yourself could provide valuable insights and contribute to the
   advancement of medical knowledge.\n\nIf you're interested, I'd be
   happy to provide more information on how to get involved in clinical
   trials or discuss any questions you may have. Your contribution to
   ```


```
research efforts is truly appreciated!\n\nBest regards,\n[Your Name]"
}
```

Evaluation of Results:

Manual analysis of the responses generated by the LLM (OpenAI's GPT-3) showed promising results in providing personalized and engaging messages to Reddit commenters. The methodology adopted in this project demonstrates the potential of AI-driven communication in enhancing user engagement and interaction on online platforms.