

Lab 4 Part 1 - Reddit API Data Extraction

Important Files

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

1. .env - File with environment variables
2. ca.pem - Certificate Authentication file for database
3. requirements.txt - File with dependencies to installed for the project

Setup

- For setup move into `reddit-scrapers` directory

```
cd reddit-scrapers
```

- 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
```

- To scrape more posts a `line 107` can be modified in the `main.py` file.

```
new_post_count = <enter your desired number>
```

- Scraping may take some time so please be patient.
- Data Collected is processed and ingested into MySQL database.

To view the data ingested data MySQL database

- Open **MySQL Workbench**. Use the following environment variables to connect to the database
 - MYSQL_USERNAME
 - MYSQL_PASSWORD
 - MYSQL_HOST
- Snapshot after a successful connection to the database

The screenshot displays the MySQL Workbench interface. The 'Query' window shows a SQL query: `select * from tech;`. The 'Result Grid' displays a table with columns: `id`, `title`, `author`, `num_comments`, `url`, `selftext`, `created_utc`, `upvote_ratio`, `score`, `num_crossposts`, `preview`, and `permalink`. The 'Output' window at the bottom shows the execution progress: `25 22:41:28 select * from tech LIMIT 0.50000`, `356 row(s) returned`, `0.047 sec / 0.938 sec`, `26 22:41:56 select * from tech LIMIT 0.50000`, `367 row(s) returned`, `0.062 sec / 1.125 sec`, `27 22:44:17 select * from tech LIMIT 0.50000`, `393 row(s) returned`, `0.062 sec / 0.672 sec`, and `28 23:10:48 select * from tech LIMIT 0.50000`, `393 row(s) returned`, `0.062 sec / 0.500 sec`.