To run the code, simply run:

“python3 praw_reddit_scrape.py”

And then give it the following requested inputs and MySQL credentials.

The data from Reddit website (https://www.reddit.com/r/tech/) has been fetched using the PRAW API. Below are the steps that were followed for scraping the data from Reddit forum :

First, user registers as a developer at the website ‘https://www.reddit.com/prefs/apps/’
Then he creates an app, names it and submits
Reddit registers his created app by assigning a client ID, a secret client access code and the user agent. The user agent field is supposed to be the description of the app that the user provided while creating the app as a developer.
After having this information, user moves on and establishes a read-only instance to Reddit. Below is the command for that :

```
# Read-only instance
reddit_read_only = praw.Reddit(client_id="********* ",              # the client id provided
                               client_secret="********", # the client secret access code
                               user_agent="Scraping data from Reddit using PRAW")        # the app
description
```

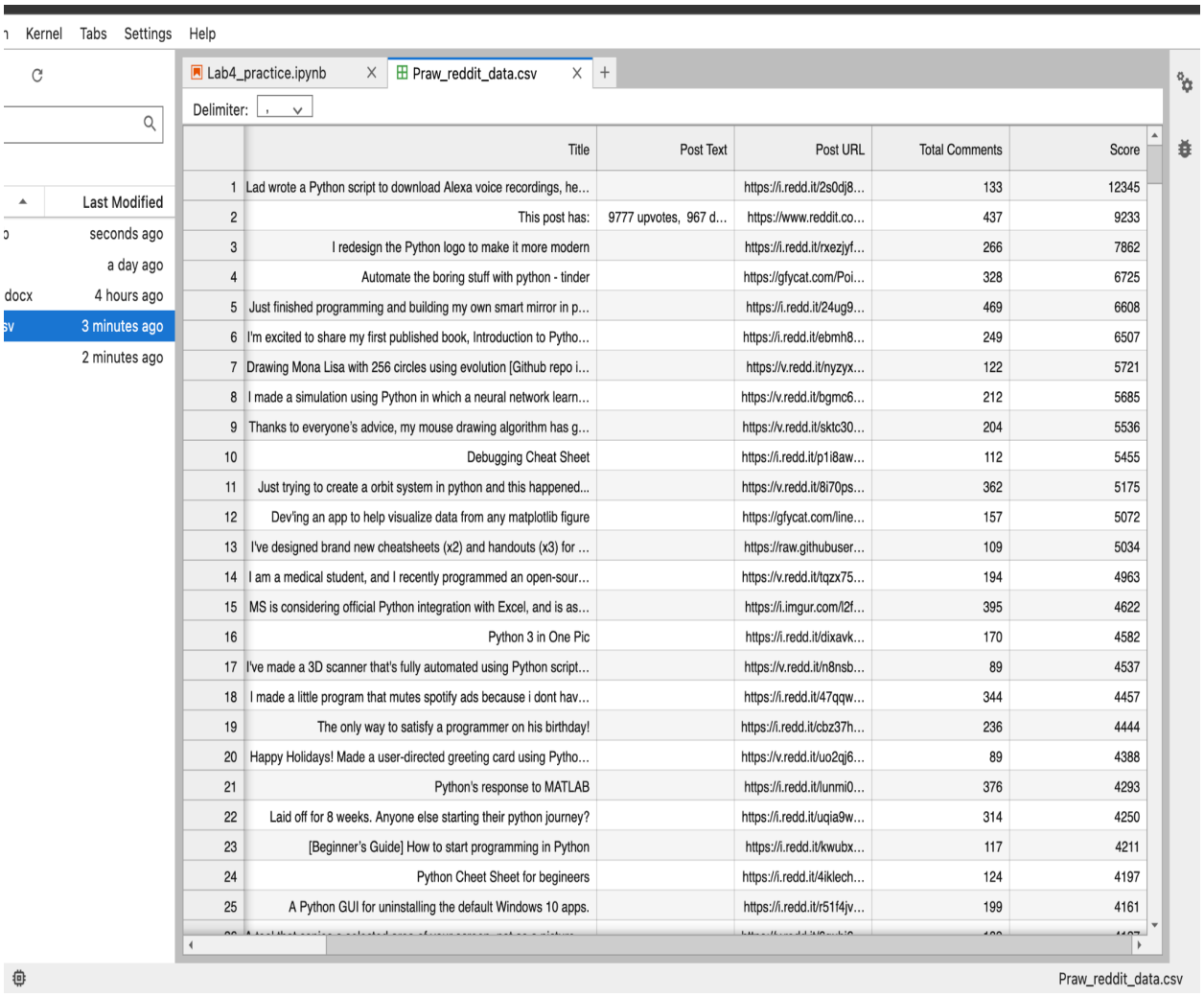
this reddit instance is now used with Praw API to fetch data from Reddit forum.
The code asks the user to enter an input which states the number of posts he wishes to fetch from the forum
The Praw reddit instance then fetches those number of posts from the Reddit forum. Below is the code syntax :

```
for post in subreddit.top(limit=input_posts): #fetching the top posts on Reddit
        print(post.title)
```

Praw API can fetch only 1000 requests at a time so we need to send multiple requests if number of posts is greater than 1000.

Various features about each of the posts in the list can be fetched using post.title, post.comment, post.score. This data is then cleaned, processed, and stored in a CSV file using pandas library.

Attached is a screenshot of the CSV file :



	Title	Post Text	Post URL	Total Comments	Score
1	Lad wrote a Python script to download Alexa voice recordings, he...		https://i.redd.it/2s0dj8...	133	12345
2	This post has:	9777 upvotes, 967 d...	https://www.reddit.co...	437	9233
3	I redesign the Python logo to make it more modern		https://i.redd.it/rxezjyf...	266	7862
4	Automate the boring stuff with python - tinder		https://gfycat.com/Poi...	328	6725
5	Just finished programming and building my own smart mirror in p...		https://i.redd.it/24ug9...	469	6608
6	I'm excited to share my first published book, Introduction to Pytho...		https://i.redd.it/ebmh8...	249	6507
7	Drawing Mona Lisa with 256 circles using evolution [Github repo i...		https://v.redd.it/nyzyx...	122	5721
8	I made a simulation using Python in which a neural network learn...		https://v.redd.it/bgmc6...	212	5685
9	Thanks to everyone's advice, my mouse drawing algorithm has g...		https://v.redd.it/sktc30...	204	5536
10	Debugging Cheat Sheet		https://i.redd.it/p1i8aw...	112	5455
11	Just trying to create a orbit system in python and this happened...		https://v.redd.it/8i70ps...	362	5175
12	Dev'ing an app to help visualize data from any matplotlib figure		https://gfycat.com/line...	157	5072
13	I've designed brand new cheatsheets (x2) and handouts (x3) for ...		https://raw.githubuser...	109	5034
14	I am a medical student, and I recently programmed an open-sour...		https://v.redd.it/tqzx75...	194	4963
15	MS is considering official Python integration with Excel, and is as...		https://i.imgur.com/l2f...	395	4622
16	Python 3 in One Pic		https://i.redd.it/dixavk...	170	4582
17	I've made a 3D scanner that's fully automated using Python script...		https://v.redd.it/n8nsb...	89	4537
18	I made a little program that mutes spotify ads because i dont hav...		https://i.redd.it/47qqw...	344	4457
19	The only way to satisfy a programmer on his birthday!		https://i.redd.it/cbz37h...	236	4444
20	Happy Holidays! Made a user-directed greeting card using Pytho...		https://v.redd.it/uo2qj6...	89	4388
21	Python's response to MATLAB		https://i.redd.it/lunmi0...	376	4293
22	Laid off for 8 weeks. Anyone else starting their python journey?		https://i.redd.it/uqia9w...	314	4250
23	[Beginner's Guide] How to start programming in Python		https://i.redd.it/kwubx...	117	4211
24	Python Cheet Sheet for begineers		https://i.redd.it/4iklech...	124	4197
25	A Python GUI for uninstalling the default Windows 10 apps.		https://i.redd.it/r51f4jv...	199	4161

We discussed the importance of data cleaning to ensure the quality and integrity of the data from Reddit. It was noted that the Reddit data may contain various issues such as missing values, duplicates, inconsistent formatting, and noise that need to be addressed.
We emphasized the need to identify and handle missing data points effectively, as they can impact the quality of analysis and modeling.Duplicate records, if any, should be identified and removed during the cleaning process.We discussed the use of Pandas for data pre-processing tasks.It was agreed that Pandas provides a wide range of functions for tasks such as data cleaning, data type conversion, and handling outliers.We highlighted the need to prepare the data in a format suitable for storage in an SQL database.This includes defining the schema of the database table, specifying data types, and ensuring that the data is compatible with the SQL database management system we plan to use (e.g., MySQL).

The following schema reflects the choices made during the data preprocessing step to be stored in a mySQL database.

RedditPosts Table:

- post_id: A unique identifier for each Reddit post.
- subreddit: The subreddit where the post was made.

- title: The title of the Reddit post.
- author: The author of the post.
- post_text: The content of the post.
- post_url: The URL to the Reddit post.
- created_at: The timestamp when the post was created.
- Retrieved_at: The timestamp when the post was fetched
- score: The upvote score of the post.
- num_comments: The number of comments on the post.
- is_stickied: A boolean indicating whether the post is stickied.