Lab4-part1

Python: 3.11.4 Pandas: 2.0.3 Numpy: 1.24.3 PRAW: 7.7.1

fetch_data_from_tech.py

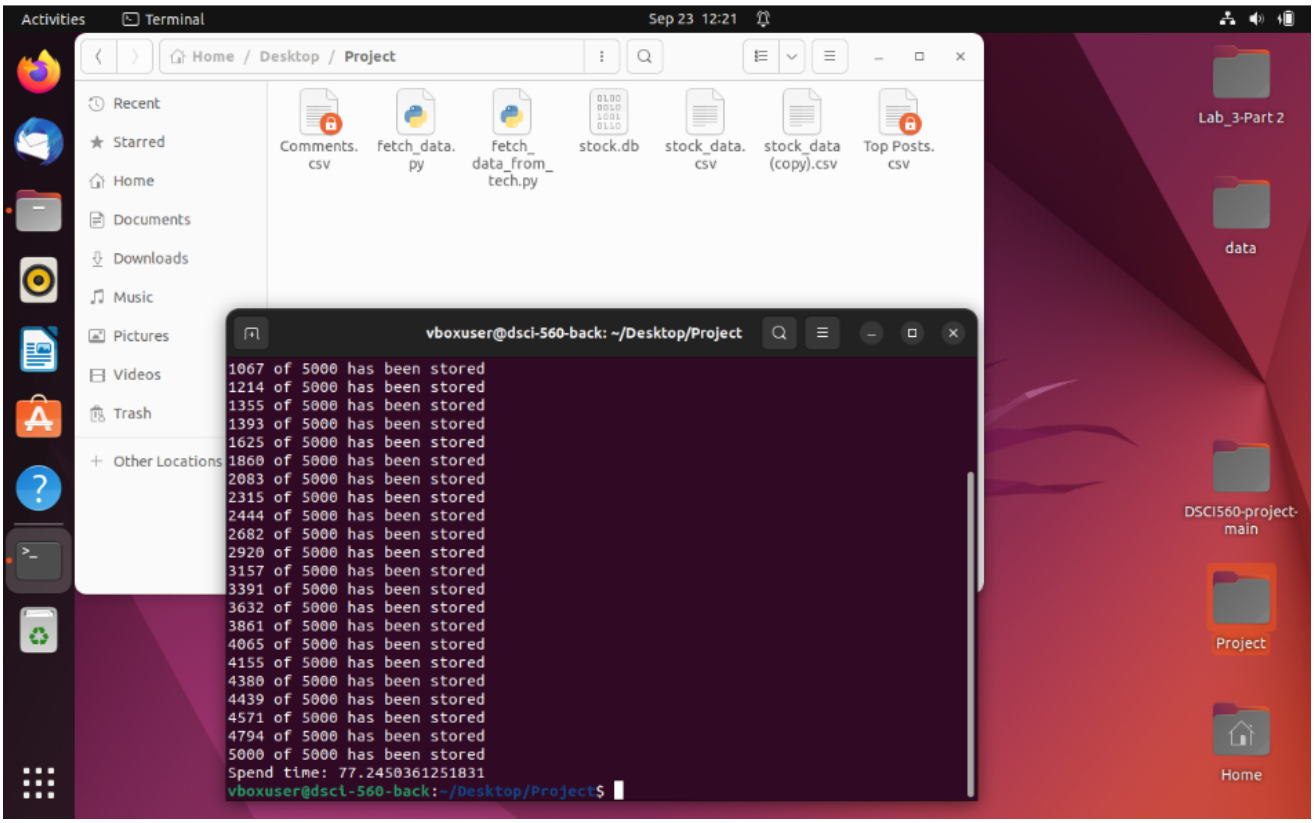
To run the script:

> python3 fetch_data_from_tech.py

Then the script will downloading some data from subreddit/tech, the default number of post is 5000 If you want to change the number of post, you can run command like:

> python3 fetch_data_from_tech.py --num 3000

You will get some information on the screen like:



Preprocessing Data

execute data_preprocessing_lab 4.py

Pandas: 2.0.3 Re: 2.2.1 Spacy: 3.6.1 nltk: 3.8.1

Dataframe after preprocessing:

1. Keywords
2. Topics
3. Remove duplicate values
4. Filter out domain from URLs

	Title	Post Text	ID	Score	Total Comments	Post URL	Domain URL	Keywords	Topics
0	CYBER WAR Anonymous leaks '776GB of Kremlin fi...	NaN	u2t0wx	17006	757	https://www.the-sun.com/tech/5106556/anonymous...	the-sun	cyber, war, anonymous, leaks, 776, gb, kremlin...	CYBER WAR Anonymous leaks, 776GB, Kremlin file...
1	Deepfake presidents used in Russia-Ukraine war	NaN	tolypc	3160	132	https://www.bbc.com/news/technology-60780142	bbc	deepfake, presidents, used, russia, ukraine, war	Deepfake presidents, Russia-Ukraine war
2	Remote Startups Will Win the War for Top Talent	NaN	wvt9c2	2343	115	https://future.com/remote-startups-hire-top-ta...	future	remote, startups, win, war, top, talent	Remote Startups, the War, Top Talent
3	Sweden returns to cold war tactics to battle f...	NaN	sls0ad	1637	228	https://www.theguardian.com/world/2022/feb/06/...	theguardian	sweden, returns, cold, war, tactics, battle, f...	Sweden, cold war tactics, fake news
4	The Ukraine War has accelerated research into ...	NaN	1692q0m	2243	73	https://www.businessinsider.com/ukraine-war-sp...	businessinsider	ukraine, war, accelerated, research, lithium, ...	The Ukraine War, research, lithium-ion battery...

Storing the data

We are using duckdb(SQL OLAP) to store the data.

1. The script that imports the preprocessed CSV into our duckdb database is "to_db.py". Quick note for replication, must use pandas 2.0.3 as there is currently a bug when working with pandas and duckdb.

```
python3 to_db.py
```