Lab 1

Isabella Yoo

Github: https://github.com/bellayoo/DSCI_560_Lab
Youtube:
https://www.youtube.com/playlist?list=PLEli1UXVdJMiXdgTlQhIKb7eII_oNJrwY

Objective

The goal of this lab is to set up the environment for future assignments and get familiar with the Linux operating system and the Python programming language, with a focus on web scraping using the *requests* and *BeautifulSoup4* libraries. The lab was completed using WSL.

Results

2.1 The folders required for the assignment were created using *mkdir* command.

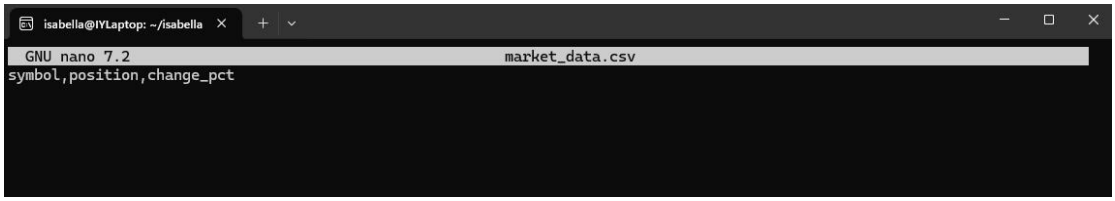
```
(venv) isabella@IYLaptop:~/isabella_isyoo$ ls
data  scripts  venv
```

2.2 A basic Python script was tested to print specified sentence with a user input.

```
(venv) isabella@IYLaptop:~/isabella_isyoo/scripts$ python3 task_1.py
name: Isabella Yoo
Hello, Isabella Yoo!
```

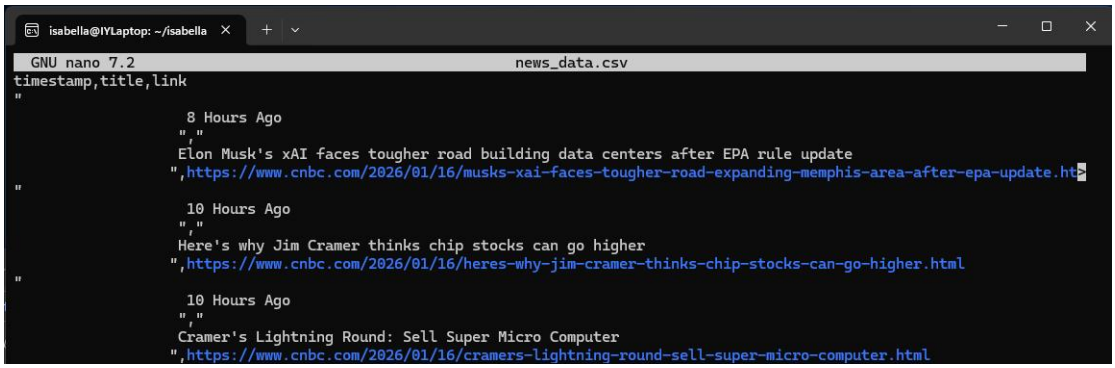
2.3 *web_scraper.py* was created to parse the given URL. The parsed HTML content was saved as *web_data.html.*

2.4 *data_filter.py* was created to read HTML file generated in the previous task. Specific sections of the web page were extracted, and the required fields were stored in Python lists. These lists were then converted into CSV files and saved in the *processed_data* folder.



The data in the Market Banner section was not stored because the content appears to be dynamically rendered using JavaScript. Retrieving it requires additional libraries

such as *Selenium*. Since this lab assignment specified the use of Requests and BeautifuSoup4 libraries, the *web_scraper.py* was not able to fetch the dynamically generated JavaScript content, resulting an empty output for the *market_data.csv.*



In contrast,the latest news section was present in the static HTML content. Thus, the resulting CSV file successfully stored required fields.