Name: Yiyang Jia, USC ID: 3882757155

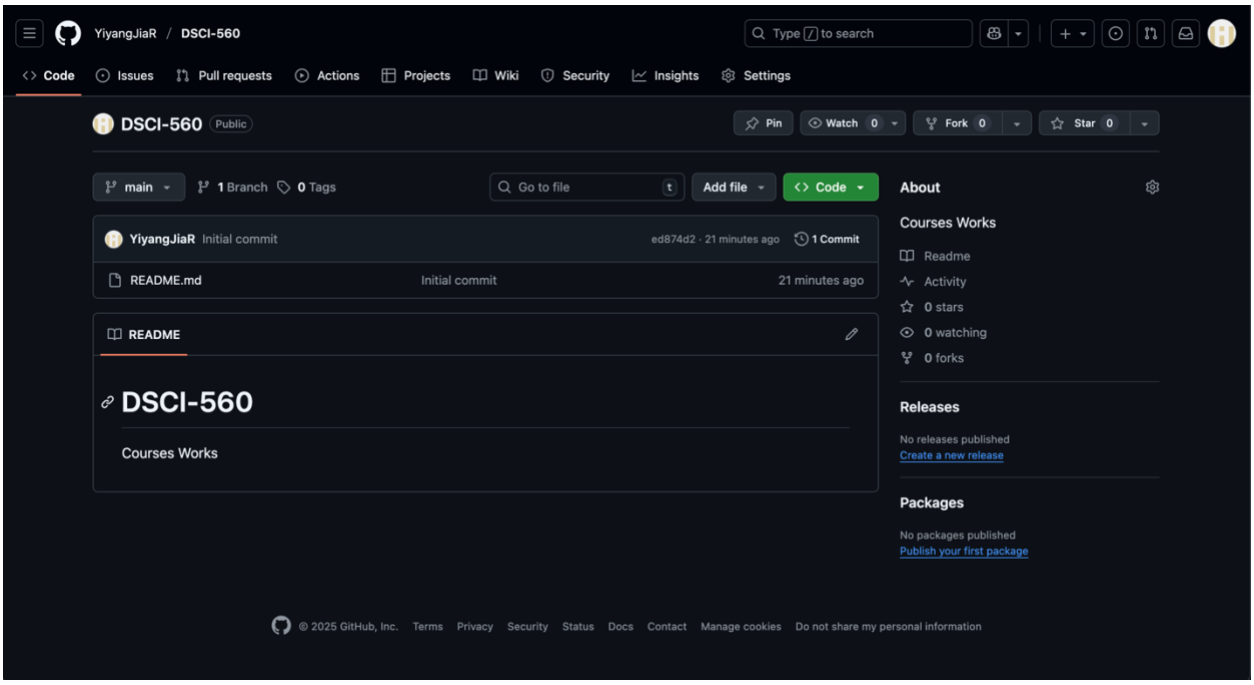
Github link: https://github.com/YiyangJiaR/DSCI-560/tree/main

Lab 1 Solution:

1. Installation & Setup

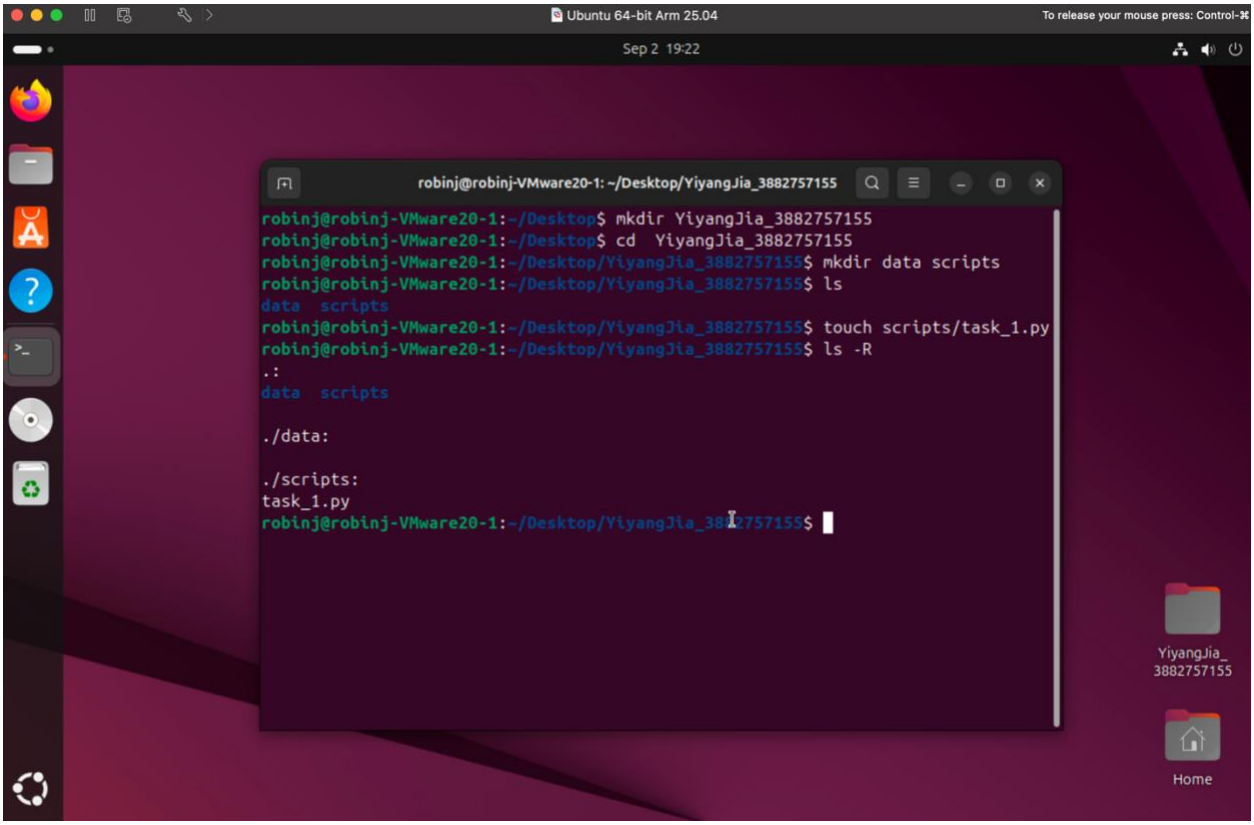
- Using Mac and M chips so download the VMware fusion and Ubuntu ARM version to create the new virtual machine.

- Create new public repo of DSCI 560



2. Playing Around with Linux Terminal

- Open the Linux Terminal and write the commands below to achieve the requirements. Just use the mkdir to create new directory and nano to get into the .py file to edit and cat to show what the files got after run it.



3. A basic python script

```
robinj@robinj-VMware20-1:~/Desktop/YiyangJia_3882757155$ cd scripts
robinj@robinj-VMware20-1:~/Desktop/YiyangJia_3882757155/scripts$ nano task_1.py
robinj@robinj-VMware20-1:~/Desktop/YiyangJia_3882757155/scripts$ cat task_1.py
name = input("Enter your name: ")
print(f"Hello, {name}!")
robinj@robinj-VMware20-1:~/Desktop/YiyangJia_3882757155/scripts$ python3 task_1.py
Enter your name: Robin
Hello, Robin!
robinj@robinj-VMware20-1:~/Desktop/YiyangJia_3882757155/scripts$
```

4. Python Web Scraping Task

I first installed the requests and beautifulsoup4 libraries and created the web_scraper.py file inside the scripts and the raw_data and processed_data inside the web_scraper.py file when nano inside to edit it. When first write the web_scraper.py I just use the request.get without headers inside the function, however, it does not work so I need to add the real browser headers then it works.

As requested, fetching CNBC world page HTML and saves it as web_data.html inside the raw_data which also inside the data directory. The results show the 10 headlines of the web_data.html file. You will see the web_scraper.py in the GitHub repo link.

```
robinj@robinj-VMware20-1:~/Desktop/YiyangJia_3882757155/scripts$ nano web_scraper.py
robinj@robinj-VMware20-1:~/Desktop/YiyangJia_3882757155/scripts$ python3 web_scraper.py
[OK] Saved: /home/robinj/Desktop/YiyangJia_3882757155/data/raw_data/web_data.html
robinj@robinj-VMware20-1:~/Desktop/YiyangJia_3882757155/scripts$ ls
task_1.py  web_scraper.py
robinj@robinj-VMware20-1:~/Desktop/YiyangJia_3882757155/scripts$ head -n 10 data/raw_data/web_data.html
head: cannot open 'data/raw_data/web_data.html' for reading: No such file or directory
robinj@robinj-VMware20-1:~/Desktop/YiyangJia_3882757155/scripts$ head -n 10 ../data/raw_data/web_data.html
<!DOCTYPE html>
<html itemscope="" itemtype="https://schema.org/WebPage" lang="en" prefix="og=https://ogp.me/ns#">
 <head>
  <meta content="website" property="og:type"/>
  <meta content="International: Top News And Analysis" property="og:title"/>
  <meta content="CNBC International is the world leader for news on business, technology, China, trade, oil prices, the
 Middle East and markets." property="og:description"/>
  <meta content="https://www.cnbc.com/world/" property="og:url"/>
  <meta content="CNBC" property="og:site_name"/>
  <meta content="max-image-preview:large" name="robots"/>
  <meta content="telephone=no" name="format-detection"/>
```

5. Data Filtering Task

```
robinj@robinj-VMware20-1:~/Desktop/YiyangJia_3882757155$ cat data/processed_data/market_data.csv
marketCard_symbol,marketCard_stockPosition,marketCard-changePct
robinj@robinj-VMware20-1:~/Desktop/YiyangJia_3882757155$
```

```
robinj@robinj-VMware20-1:~/Desktop/YiyangJia_3882757155/scripts$ nano data_filter.py
robinj@robinj-VMware20-1:~/Desktop/YiyangJia_3882757155/scripts$ python3 data_filter.py
[OK] Wrote: /home/robinj/Desktop/YiyangJia_3882757155/data/processed_data/market_data.csv and /home/robinj/Desktop/Yiya
ngJia_3882757155/data/processed_data/news_data.csv
```

I did the data filtering and when get to the market_data.csv and the news_data.csv it shows just the headers, and I think it's the html problem and as shown below my web_data.html only shows the CSS class names but not the data-symbol attributes and I try to redefine the headers, but I does not work out. I do not know how to fix it, but I think the logic of the code is correct and match the lab requirements. I will

try to fix out.