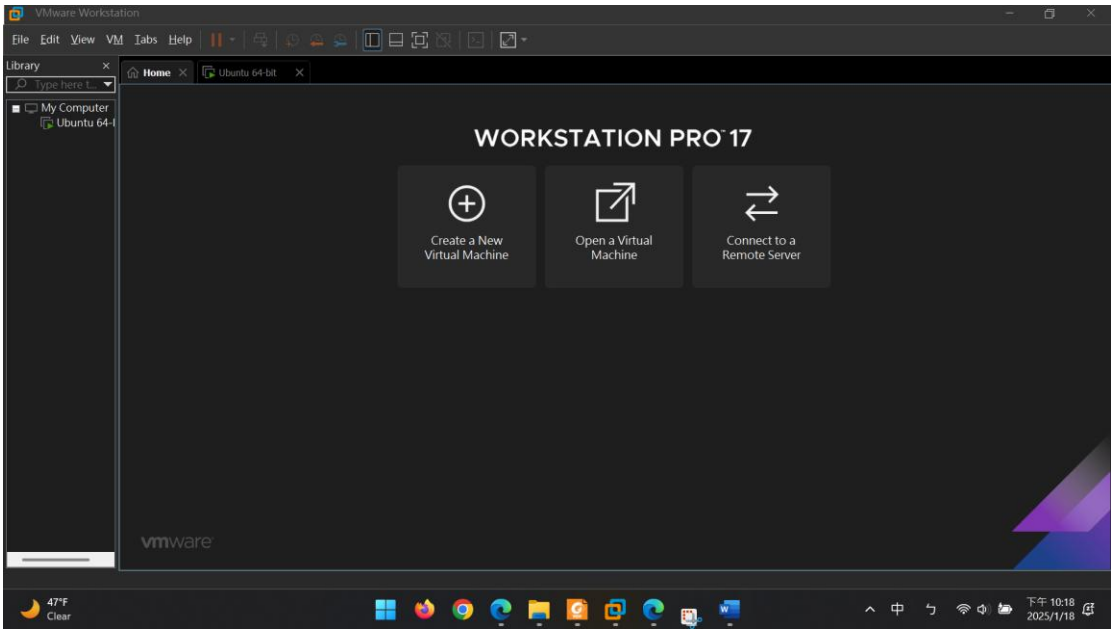
DSCI560 - Lab 1

USC-ID: 1934208430

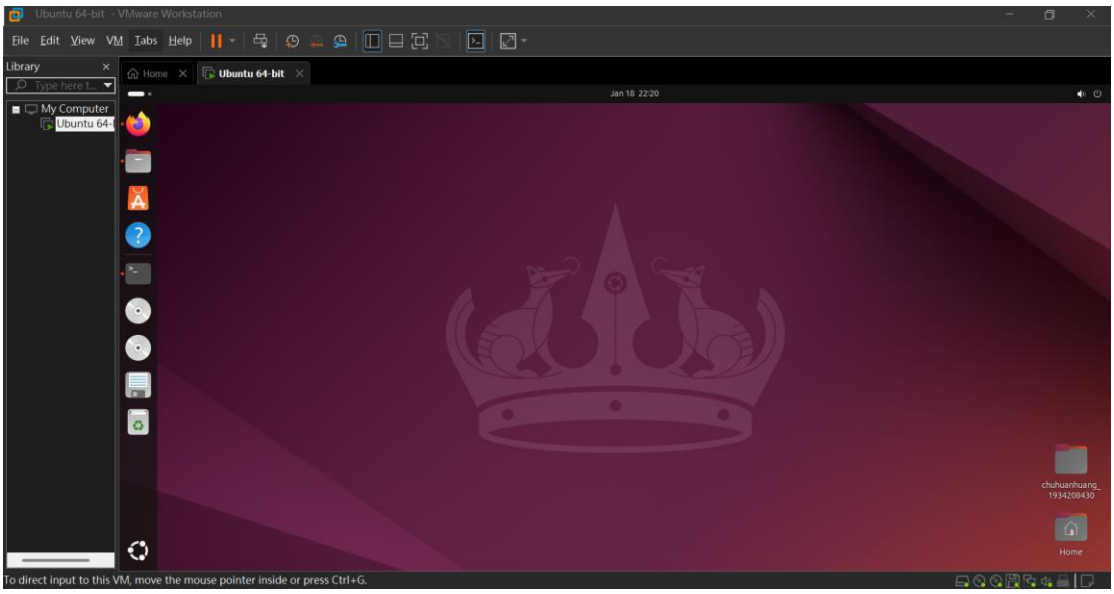
Name: Chu-Huan Huang

1. Installation and Setup

I. Install VMware



II. Download Ubuntu ISO and create VM



III. Install python

```
chris@chris-VMware-Virtual-Platform:~$ python3 --version
Python 3.12.3
chris@chris-VMware-Virtual-Platform:~$ pip --version
pip 24.0 from /usr/lib/python3/dist-packages/pip (python 3.12)
```

2. Get Familiar with Linux and Python

I. Playing around with Linux Terminal

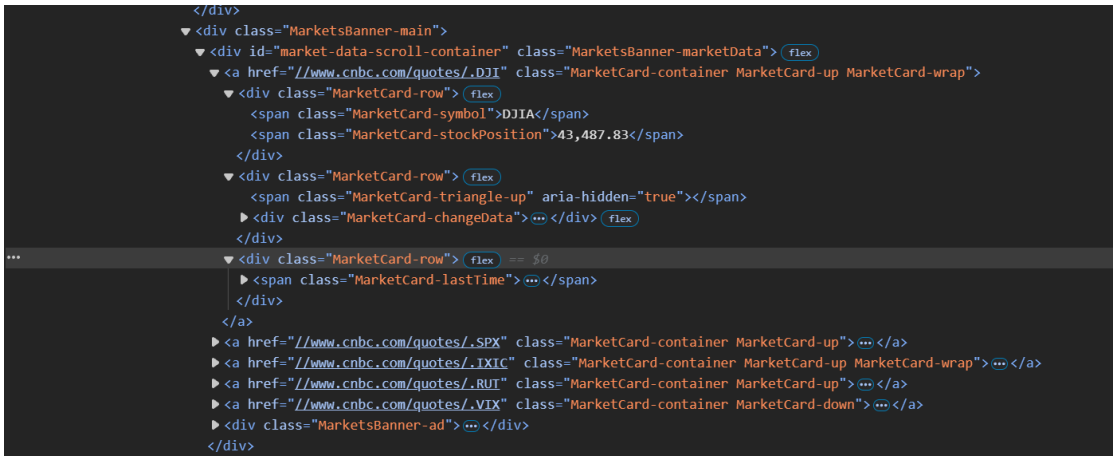
II. A basic Python Script

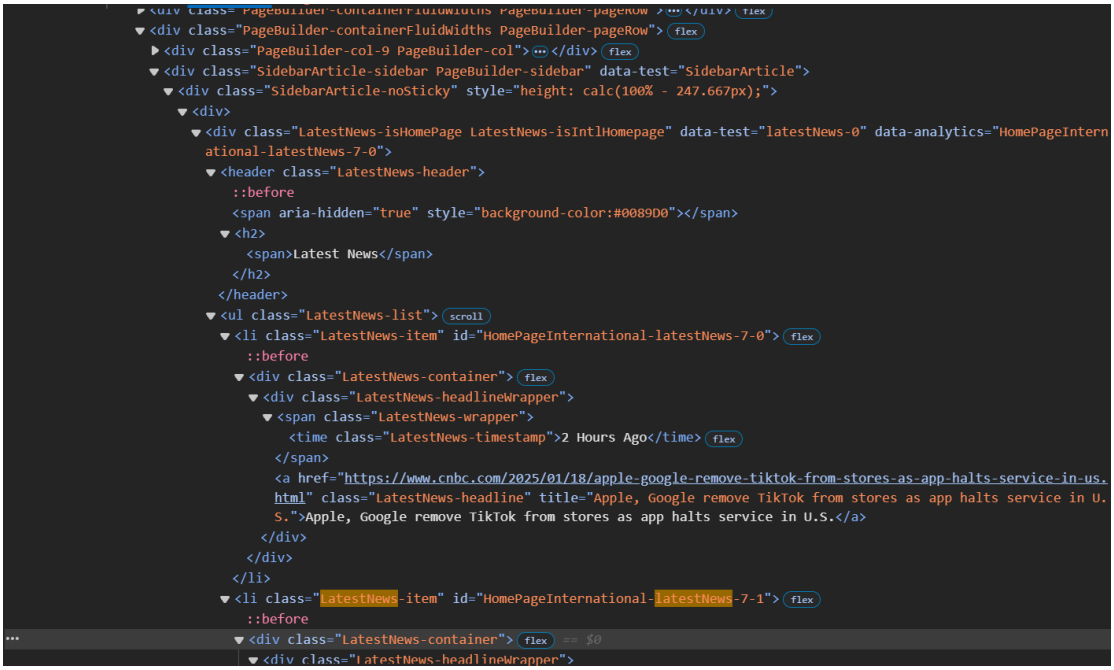
```
chris@chris-VMware-Virtual-Platform:~$ ls
Desktop  Documents  Downloads  Music  Pictures  Public  snap  Templates  Videos
chris@chris-VMware-Virtual-Platform:~$ cd Desktop
chris@chris-VMware-Virtual-Platform:~/Desktop$ ls
chuhuanhuang_1934208430
chris@chris-VMware-Virtual-Platform:~/Desktop$ cd chuhuanhuang_1934208430
chris@chris-VMware-Virtual-Platform:~/Desktop/chuhuanhuang_1934208430$ ls
data  scripts
chris@chris-VMware-Virtual-Platform:~/Desktop/chuhuanhuang_1934208430$ cd scripts
chris@chris-VMware-Virtual-Platform:~/Desktop/chuhuanhuang_1934208430/scripts$ ls
chromedriver  data_filter.py  task_1.py  task_1.py.save  web_scraper.py
chris@chris-VMware-Virtual-Platform:~/Desktop/chuhuanhuang_1934208430/scripts$ python3 task_1.py
Enter username: CHris
Hello, CHris!
```

III. Python Web-scraping Task

Here, due to the *requests* package cannot not get the html I needed (that can snap the section of Market Banner & Latest News I showed below). Thus, I choose to use selenium to crawl the URL.

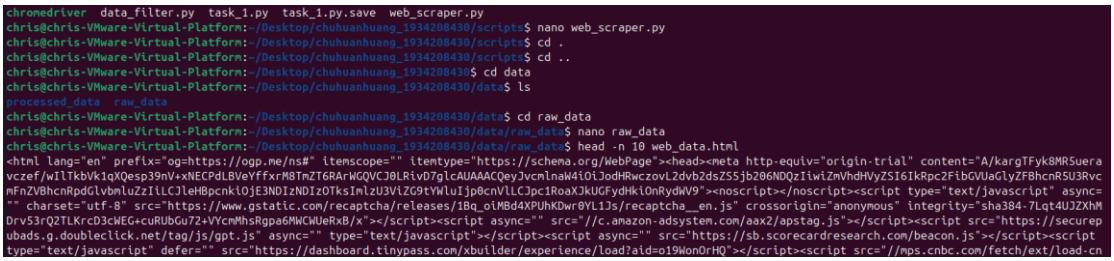
The section of the info we need after inspecting:





The code:

First 10 rows of the raw data(before extracting info we wanted):



IV. Data Filtering Task

```python
import csv
from bs4 import BeautifulSoup
import os

#open the raw data
with open("../data/raw_data/web_data.html", "r", encoding = "utf-8") as f:
  content = f.read()

#parse with BS4 indent
soup = BeautifulSoup(content, "html.parser")

#marketBanner part
market_data = []
market_data_container = soup.find('div', id = 'market-data-scroll-container', class_ = 'MarketsBanner-marketData')

print("filtering Market Banner")
if market_data_container:
  market_cards = market_data_container.find_all('a', class_ = 'MarketCard-container')
  if market_cards:
    for card in market_cards:
      symbol = card.find('span', class_ = 'MarketCard-symbol').text.strip()
      stock_pos = card.find('span', class_ = 'MarketCard-stockPosition').text.strip()
      change_pct = card.find('span', class_ = 'MarketCard-changesPct').text.strip()
      market_data.append({
        'marketCard_symbol': symbol,
        'marketCard_stockPosition': stock_pos,
        'marketCardchangePct': change_pct
        })
```

Here, use BeautifulSoup4 to parse the html and extract the information of Market Banner and Latest News by the attributes of the section where each elements located in.

```
chris@chris-VMware-Virtual-Platform:~/Desktop/chuhuanhuang_1934208430/scripts$ python3 data_filter.py
filtering Market Banner
filtering Latest News
storing results as CSV
marketBanner CSV created
LatestNews CSV created
```

The results:

Open news_data.csv ~/Desktop/chuhuanhuang_1934208430/data/processed_data

news_data.csv market_data.csv

```
timestamp,title,link
2 Hours Ago,"Apple, Google remove TikTok from stores as app halts service in U.S.",https://www.cnbc.com/2025/01/18/apple-google-remove-tiktok-from-stores-as-app-halts-service-in-us.html
10 Hours Ago,Perplexity AI makes a bid to merge with TikTok U.S.,https://www.cnbc.com/2025/01/18/perplexity-ai-makes-a-bid-to-merge-with-tiktok-us.html
13 Hours Ago,"Solana surges 12% on launch of Trump-themed meme coin, ether falls",https://www.cnbc.com/2025/01/18/crypto-market-today.html
14 Hours Ago,"What to expect from travel prices in 2025, and which spots have the best deals",https://www.cnbc.com/2025/01/18/what-to-expect-from-travel-prices-in-2025.html
15 Hours Ago,Consumer protection agencies at risk in Trump's second term: What it means for you,https://www.cnbc.com/2025/01/18/how-trumps-second-term-could-mean-the-downfall-of-the-fdic-cfpb.html
16 Hours Ago,Why the gold boom is causing a surge in illegal mining,https://www.cnbc.com/2025/01/18/why-the-gold-boom-is-causing-a-surge-in-illegal-mining.html
```

Open market_data.csv ~/Desktop/chuhuanhuang_1934208430/data/processed_data

news_data.csv market_data.csv

```
marketCard_symbol,marketCard_stockPosition,marketCardchangePct
DJIA,"43,487.83",+0.78%
S&P 500,"5,996.66",+1.00%
NASDAQ,"19,630.20",+1.51%
RUSS 2K*,"2,275.88",+0.40%
VIX,15.97,-3.80%
```