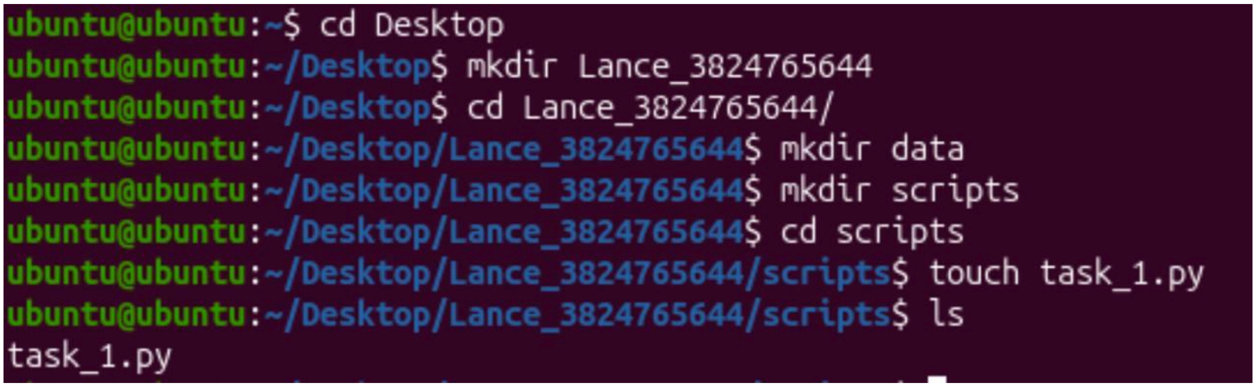
Name: Lance Dsilva
USCID: 3824765644
DSCI 560 - Lab 1 Report

Directory Creation and Setup

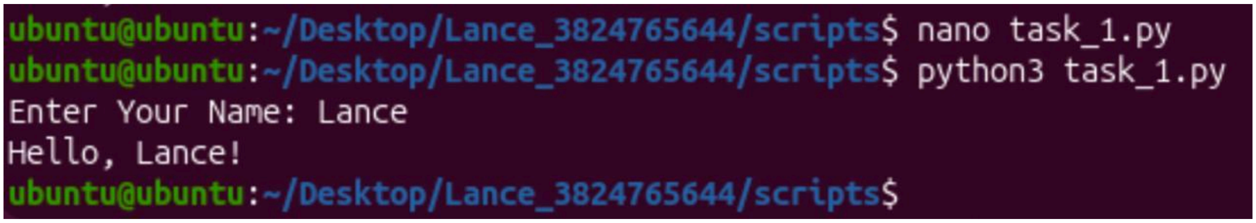
In this step, a new directory named `Lance_3824765644` was created on the Ubuntu Desktop to organize all files related to Lab 1. Inside this directory, two subfolders named `data` and `scripts` were created to separate datasets from Python scripts. An empty Python file `task_1.py` was created inside the `scripts` using Linux terminal commands (`mkdir`, `cd`, `touch`, and `ls`).



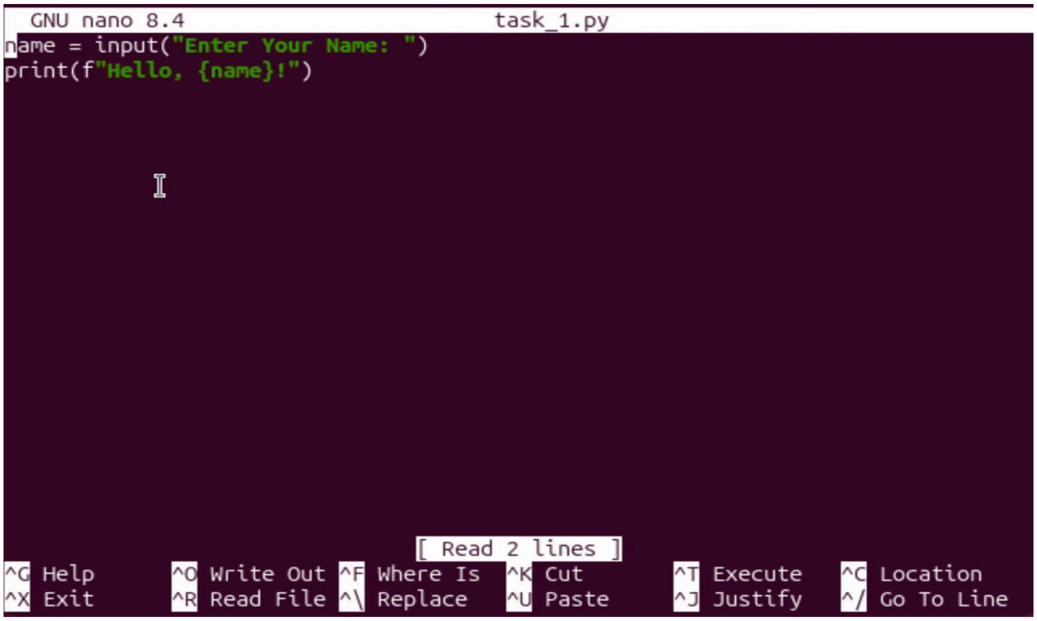
```
ubuntu@ubuntu:~$ cd Desktop
ubuntu@ubuntu:~/Desktop$ mkdir Lance_3824765644
ubuntu@ubuntu:~/Desktop$ cd Lance_3824765644/
ubuntu@ubuntu:~/Desktop/Lance_3824765644$ mkdir data
ubuntu@ubuntu:~/Desktop/Lance_3824765644$ mkdir scripts
ubuntu@ubuntu:~/Desktop/Lance_3824765644$ cd scripts
ubuntu@ubuntu:~/Desktop/Lance_3824765644/scripts$ touch task_1.py
ubuntu@ubuntu:~/Desktop/Lance_3824765644/scripts$ ls
task_1.py
```

Basic Python Script Execution

The `task_1.py` file was opened using the `nano` text editor. A simple Python script was written that prompts the user to enter their name and then prints a greeting message in the format “Hello, [name]!”. The script was saved and executed using the `python3` command, confirming that Python was correctly installed and functioning within the Ubuntu virtual machine.



```
ubuntu@ubuntu:~/Desktop/Lance_3824765644/scripts$ nano task_1.py
ubuntu@ubuntu:~/Desktop/Lance_3824765644/scripts$ python3 task_1.py
Enter Your Name: Lance
Hello, Lance!
ubuntu@ubuntu:~/Desktop/Lance_3824765644/scripts$
```

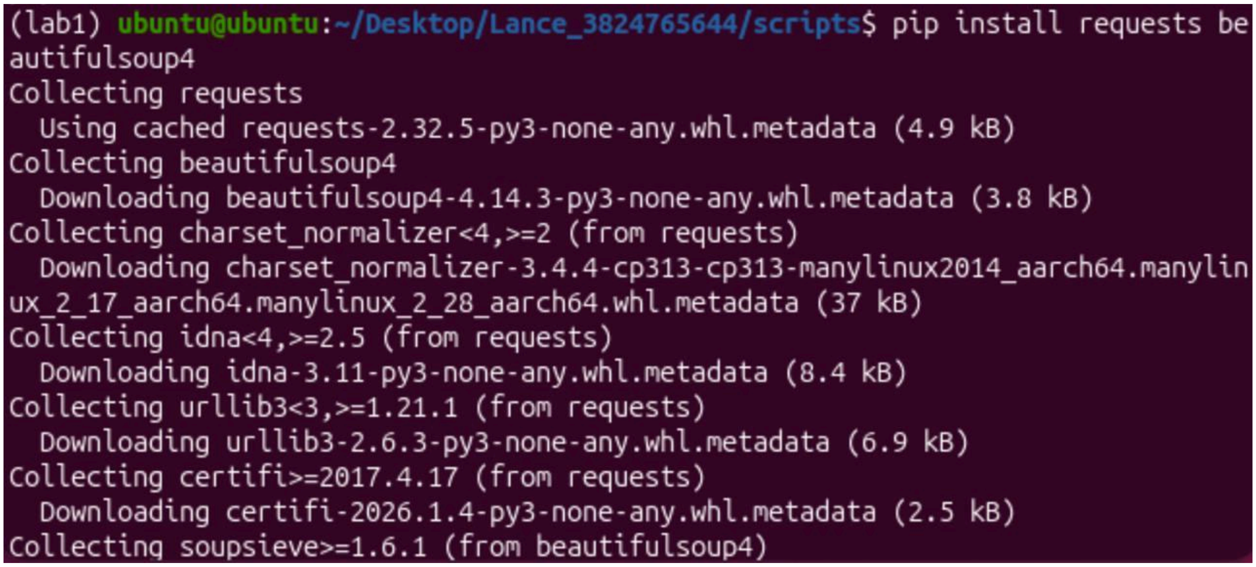


```
name = input("Enter Your Name: ")
print(f"Hello, {name}!")
```

```
[ Read 2 lines ]
```

Installing Required Python Libraries

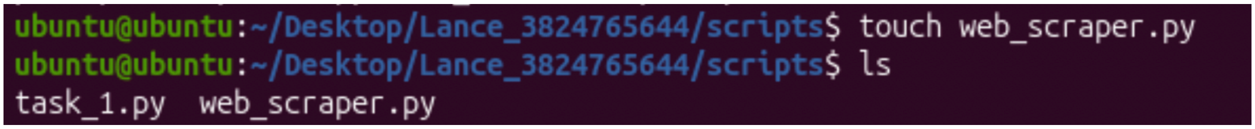
The required Python libraries for web scraping, namely `requests` and `beautifulsoup4`, were installed using `pip`. The successful installation confirmed that `pip` was correctly configured in the Python virtual environment. These libraries are required to fetch web content and parse HTML structures for subsequent scraping tasks.



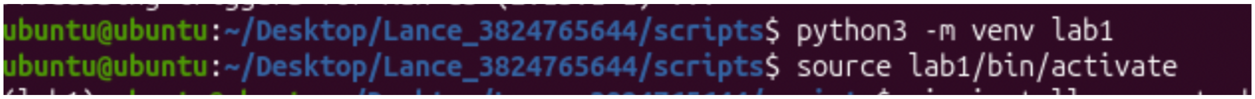
```
(lab1) ubuntu@ubuntu:~/Desktop/Lance_3824765644/scripts$ pip install requests be
autifulsoup4
Collecting requests
  Using cached requests-2.32.5-py3-none-any.whl.metadata (4.9 kB)
Collecting beautifulsoup4
  Downloading beautifulsoup4-4.14.3-py3-none-any.whl.metadata (3.8 kB)
Collecting charset_normalizer<4,>=2 (from requests)
  Downloading charset_normalizer-3.4.4-cp313-cp313-manylinux2014_aarch64.manylin
ux_2_17_aarch64.manylinux_2_28_aarch64.whl.metadata (37 kB)
Collecting idna<4,>=2.5 (from requests)
  Downloading idna-3.11-py3-none-any.whl.metadata (8.4 kB)
Collecting urllib3<3,>=1.21.1 (from requests)
  Downloading urllib3-2.6.3-py3-none-any.whl.metadata (6.9 kB)
Collecting certifi>=2017.4.17 (from requests)
  Downloading certifi-2026.1.4-py3-none-any.whl.metadata (2.5 kB)
Collecting soupsieve>=1.6.1 (from beautifulsoup4)
```

Web Scraper Script Creation and HTML Saving

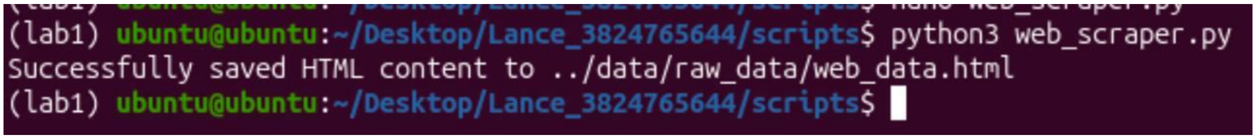
A Python script named `web_scraper.py` was created to fetch the CNBC World News webpage using the `requests` library with a custom User-Agent header. The retrieved HTML content was parsed and formatted using BeautifulSoup and saved as `web_data.html` inside the `data/raw_data` directory.



```
ubuntu@ubuntu:~/Desktop/Lance_3824765644/scripts$ touch web_scraper.py
ubuntu@ubuntu:~/Desktop/Lance_3824765644/scripts$ ls
task_1.py  web_scraper.py
```



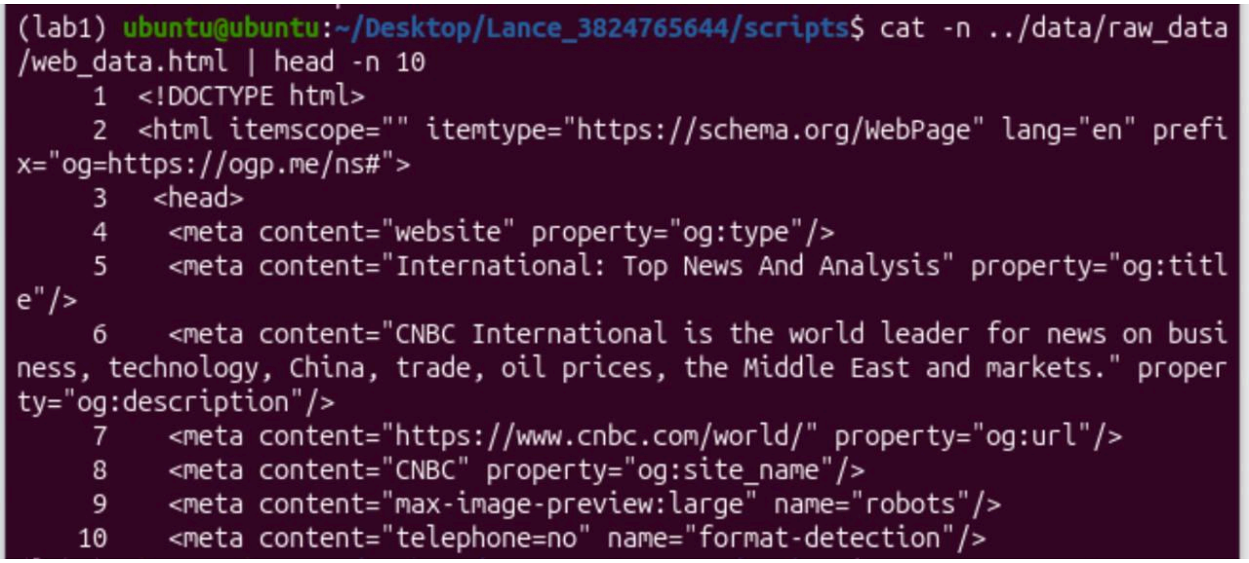
```
ubuntu@ubuntu:~/Desktop/Lance_3824765644/scripts$ python3 -m venv lab1
ubuntu@ubuntu:~/Desktop/Lance_3824765644/scripts$ source lab1/bin/activate
```



```
(lab1) ubuntu@ubuntu:~/Desktop/Lance_3824765644/scripts$ python3 web_scraper.py
Successfully saved HTML content to ../data/raw_data/web_data.html
(lab1) ubuntu@ubuntu:~/Desktop/Lance_3824765644/scripts$ 
```

Viewing Raw HTML Output

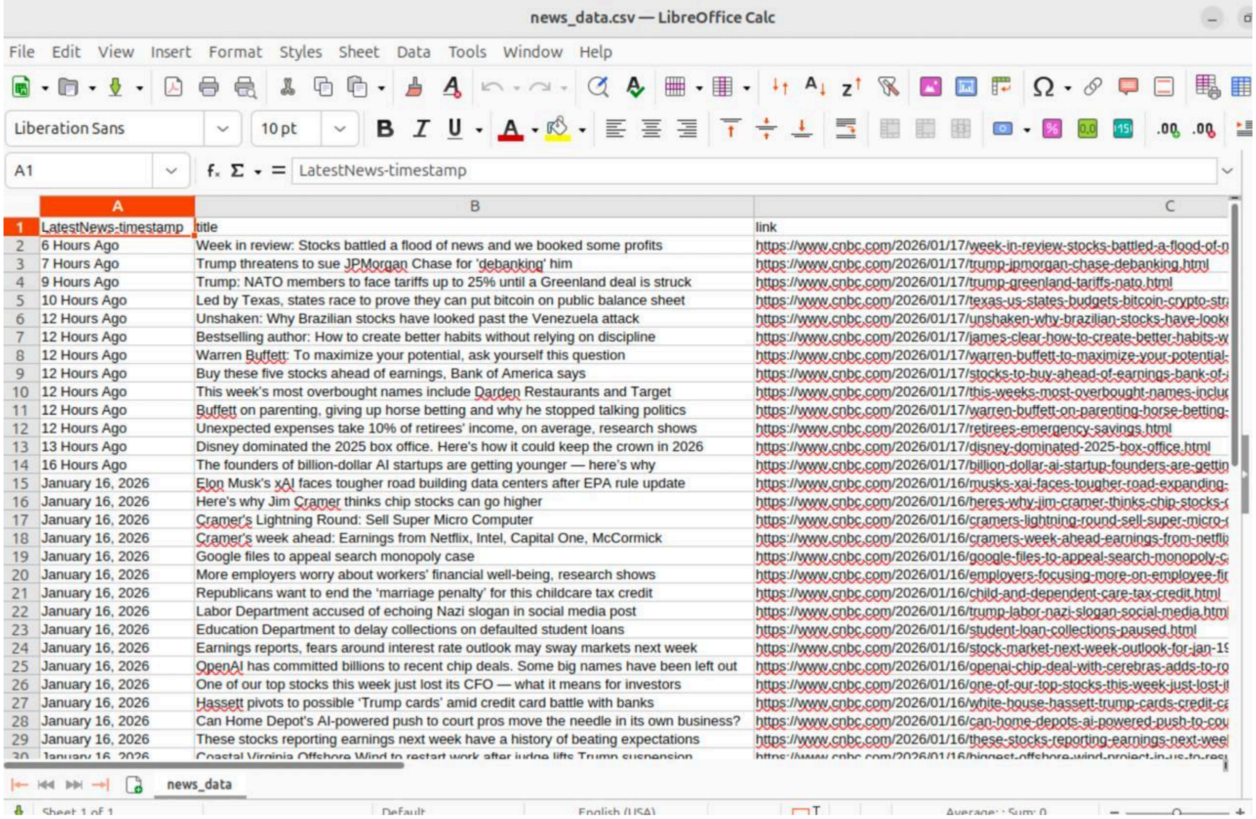
The first ten lines of the saved `web_data.html` file were displayed in the terminal using the `cat` command. This confirmed that the webpage HTML was successfully downloaded and stored locally. The output shows the document type declaration and metadata elements of the CNBC webpage.



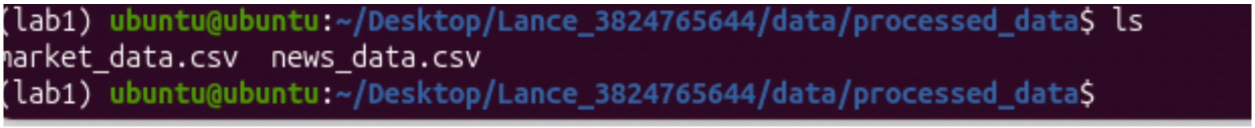
```
(lab1) ubuntu@ubuntu:~/Desktop/Lance_3824765644/scripts$ cat -n ../data/raw_data
/web_data.html | head -n 10
     1  <!DOCTYPE html>
     2  <html itemscope="" itemtype="https://schema.org/WebPage" lang="en" prefi
x="og=https://ogp.me/ns#">
     3   <head>
     4    <meta content="website" property="og:type"/>
     5    <meta content="International: Top News And Analysis" property="og:titl
e"/>
     6    <meta content="CNBC International is the world leader for news on busi
ness, technology, China, trade, oil prices, the Middle East and markets." proper
ty="og:description"/>
     7    <meta content="https://www.cnbc.com/world/" property="og:url"/>
     8    <meta content="CNBC" property="og:site_name"/>
     9    <meta content="max-image-preview:large" name="robots"/>
    10    <meta content="telephone=no" name="format-detection"/>
```

Data Filtering Script Execution

The `data_filter.py` script was executed to parse the saved HTML file and extract relevant information. The script successfully identified and extracted 30 items from the “Latest News” section, including timestamps, titles, and links. These entries were stored in `news_data.csv` inside the `processed_data` directory. The `news_data.csv` file was opened using LibreOffice Calc to visually verify the extracted data. The spreadsheet displays timestamps, article titles, and corresponding links, confirming that the Latest News data was correctly parsed and stored in a structured format.



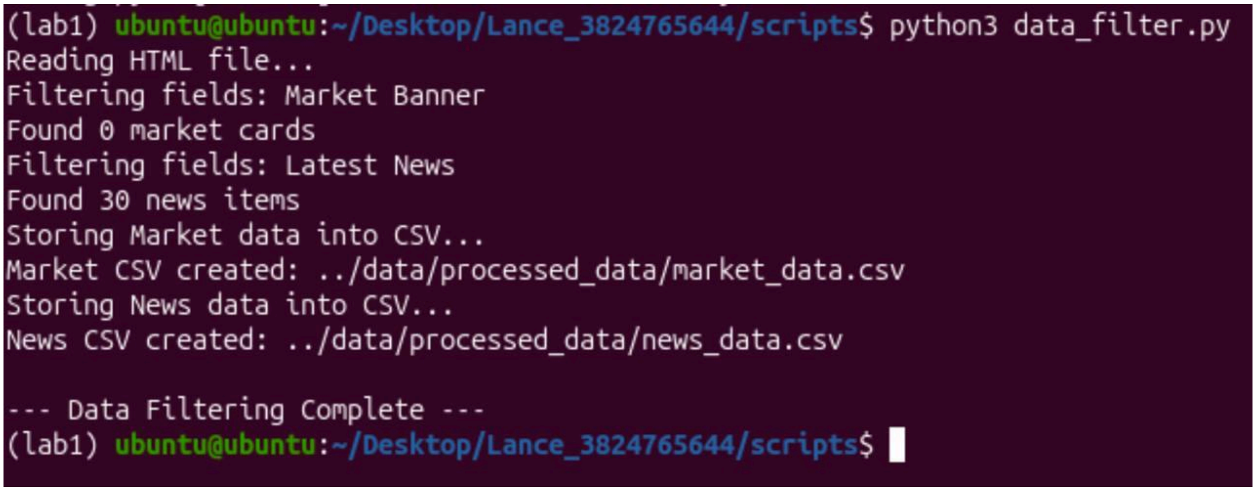
	A	B	C
1	LatestNews-timestamp	title	link
2	6 Hours Ago	Week in review: Stocks battled a flood of news and we booked some profits	https://www.cnbc.com/2026/01/17/week-in-review-stocks-battled-a-flood-of-n
3	7 Hours Ago	Trump threatens to sue JPMorgan Chase for 'debanking' him	https://www.cnbc.com/2026/01/17/trump-jpmorgan-chase-debanking.html
4	9 Hours Ago	Trump: NATO members to face tariffs up to 25% until a Greenland deal is struck	https://www.cnbc.com/2026/01/17/trump-greenland-tariffs-nato.html
5	10 Hours Ago	Led by Texas, states race to prove they can put bitcoin on public balance sheet	https://www.cnbc.com/2026/01/17/texas-us-states-budgets-bitcoin-crypto-str
6	12 Hours Ago	Unshaken: Why Brazilian stocks have looked past the Venezuela attack	https://www.cnbc.com/2026/01/17/unshaken-why-brazilian-stocks-have-look
7	12 Hours Ago	Bestselling author: How to create better habits without relying on discipline	https://www.cnbc.com/2026/01/17/james-clear-how-to-create-better-habits-w
8	12 Hours Ago	Warren Buffett: To maximize your potential, ask yourself this question	https://www.cnbc.com/2026/01/17/warren-buffett-to-maximize-your-potential-
9	12 Hours Ago	Buy these five stocks ahead of earnings, Bank of America says	https://www.cnbc.com/2026/01/17/stocks-to-buy-ahead-of-earnings-bank-of-
10	12 Hours Ago	This week's most overbought names include Darden Restaurants and Target	https://www.cnbc.com/2026/01/17/this-weeks-most-overbought-names-inclu
11	12 Hours Ago	Buffett on parenting, giving up horse betting and why he stopped talking politics	https://www.cnbc.com/2026/01/17/warren-buffett-on-parenting-horse-betting-
12	12 Hours Ago	Unexpected expenses take 10% of retirees' income, on average, research shows	https://www.cnbc.com/2026/01/17/retirees-emergency-savings.html
13	13 Hours Ago	Disney dominated the 2025 box office. Here's how it could keep the crown in 2026	https://www.cnbc.com/2026/01/17/disney-dominated-2025-box-office.html
14	16 Hours Ago	The founders of billion-dollar AI startups are getting younger — here's why	https://www.cnbc.com/2026/01/17/billion-dollar-ai-startup-founders-are-gettin
15	January 16, 2026	Elon Musk's xAI faces tougher road building data centers after EPA rule update	https://www.cnbc.com/2026/01/16/musks-xai-faces-tougher-road-expanding-
16	January 16, 2026	Here's why Jim Cramer thinks chip stocks can go higher	https://www.cnbc.com/2026/01/16/heres-why-jim-cramer-thinks-chip-stocks-
17	January 16, 2026	Cramer's Lightning Round: Sell Super Micro Computer	https://www.cnbc.com/2026/01/16/cramers-lightning-round-sell-super-micro-
18	January 16, 2026	Cramer's week ahead: Earnings from Netflix, Intel, Capital One, McCormick	https://www.cnbc.com/2026/01/16/cramers-week-ahead-earnings-from-netfli
19	January 16, 2026	Google files to appeal search monopoly case	https://www.cnbc.com/2026/01/16/google-files-to-appeal-search-monopoly-c
20	January 16, 2026	More employers worry about workers' financial well-being, research shows	https://www.cnbc.com/2026/01/16/employers-focusing-more-on-employee-fi
21	January 16, 2026	Republicans want to end the ‘marriage penalty’ for this childcare tax credit	https://www.cnbc.com/2026/01/16/child-and-dependent-care-tax-credit.html
22	January 16, 2026	Labor Department accused of echoing Nazi slogan in social media post	https://www.cnbc.com/2026/01/16/trump-labor-nazi-slogan-social-media.html
23	January 16, 2026	Education Department to delay collections on defaulted student loans	https://www.cnbc.com/2026/01/16/student-loan-collections-paused.html
24	January 16, 2026	Earnings reports, fears around interest rate outlook may sway markets next week	https://www.cnbc.com/2026/01/16/stock-market-next-week-outlook-for-jan-19
25	January 16, 2026	OpenAI has committed billions to recent chip deals. Some big names have been left out	https://www.cnbc.com/2026/01/16/openai-chip-deal-with-cerebras-adds-to-ro
26	January 16, 2026	One of our top stocks this week just lost its CFO — what it means for investors	https://www.cnbc.com/2026/01/16/one-of-our-top-stocks-this-week-just-lost-i
27	January 16, 2026	Hassett pivots to possible ‘Trump cards’ amid credit card battle with banks	https://www.cnbc.com/2026/01/16/white-house-hassett-trump-cards-credit-ca
28	January 16, 2026	Can Home Depot's AI-powered push to court pros move the needle in its own business?	https://www.cnbc.com/2026/01/16/can-home-depots-ai-powered-push-to-cou
29	January 16, 2026	These stocks reporting earnings next week have a history of beating expectations	https://www.cnbc.com/2026/01/16/these-stocks-reporting-earnings-next-wee
30	January 16, 2026	Coastal Virginia Offshore Wind to restart work after judge lifts Trump suspension	https://www.cnbc.com/2026/01/16/biggest-offshore-wind-project-in-us-to-res



```
(lab1) ubuntu@ubuntu:~/Desktop/Lance_3824765644/data/processed_data$ ls
market_data.csv  news_data.csv
(lab1) ubuntu@ubuntu:~/Desktop/Lance_3824765644/data/processed_data$
```

Market Banner Extraction Issue (JavaScript-rendered Content)

While attempting to extract Market Banner data (marketCard_symbol, marketCard_stockPosition, marketCard_changePct), zero market cards were found. This occurs because the Market Banner section on the CNBC website is dynamically rendered using JavaScript. Since `requests` and BeautifulSoup only retrieve static HTML, the Market Banner elements are not present in the saved source code. As a result, this data cannot be extracted without a JavaScript rendering engine such as Selenium with a configured browser driver.



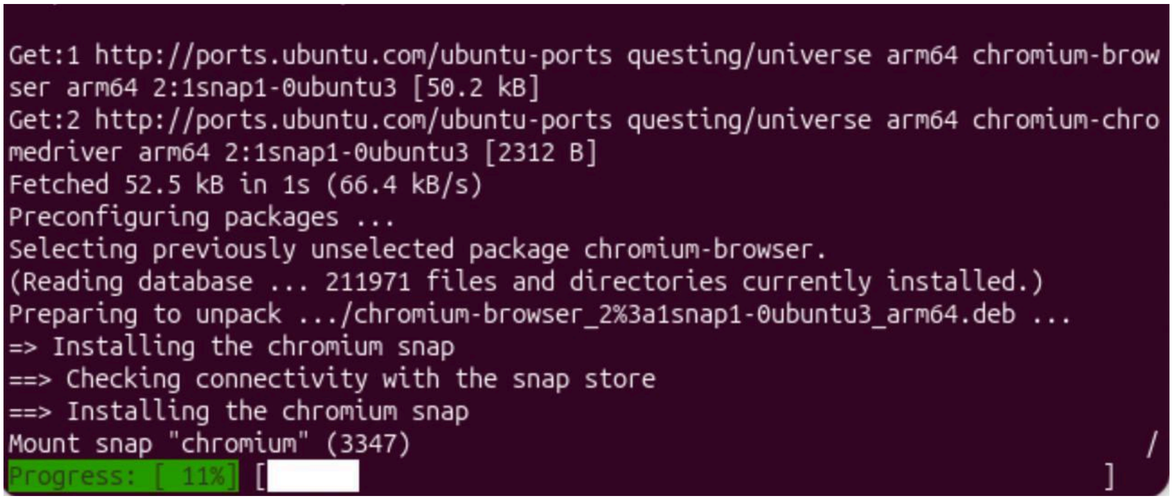
```
(lab1) ubuntu@ubuntu:~/Desktop/Lance_3824765644/scripts$ python3 data_filter.py
Reading HTML file...
Filtering fields: Market Banner
Found 0 market cards
Filtering fields: Latest News
Found 30 news items
Storing Market data into CSV...
Market CSV created: ../data/processed_data/market_data.csv
Storing News data into CSV...
News CSV created: ../data/processed_data/news_data.csv

--- Data Filtering Complete ---
(lab1) ubuntu@ubuntu:~/Desktop/Lance_3824765644/scripts$ █
```

Selenium-Based Market Data Extraction

The Market Banner section on the CNBC website is dynamically rendered using JavaScript and is not present in the static HTML retrieved by the Requests library. Selenium is required because it executes JavaScript in a real browser environment, allowing access to dynamically loaded elements such as market cards.

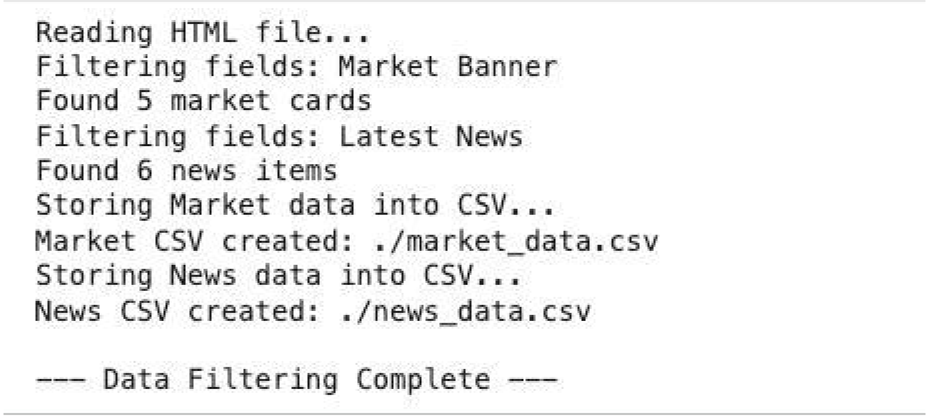
Chromium Installation Issue on Ubuntu VM (VMware)

While installing it on ubuntu in vmware, it gets stuck and having tried for 4 to 5 times, i tried on local machine

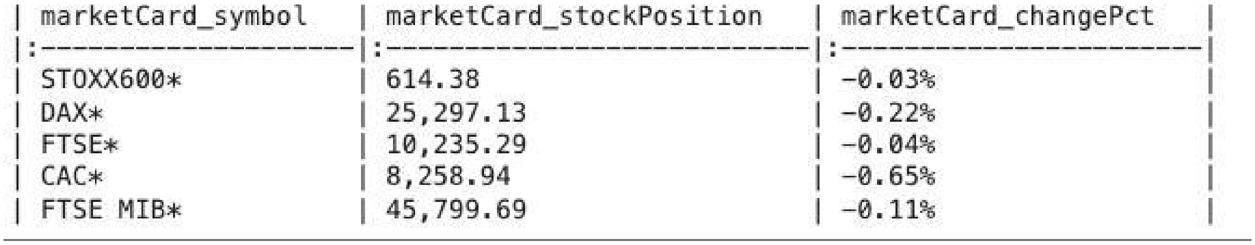


```
Get:1 http://ports.ubuntu.com/ubuntu-ports questing/universe arm64 chromium-brow
ser arm64 2:1snap1-0ubuntu3 [50.2 kB]
Get:2 http://ports.ubuntu.com/ubuntu-ports questing/universe arm64 chromium-chro
medriver arm64 2:1snap1-0ubuntu3 [2312 B]
Fetched 52.5 kB in 1s (66.4 kB/s)
Preconfiguring packages ...
Selecting previously unselected package chromium-browser.
(Reading database ... 211971 files and directories currently installed.)
Preparing to unpack .../chromium-browser_2%3a1snap1-0ubuntu3_arm64.deb ...
=> Installing the chromium snap
==> Checking connectivity with the snap store
==> Installing the chromium snap
Mount snap "chromium" (3347)                                                   /
Progress: [ 11%] [                                                           ]
```

Selenium was tested on a local machine where Chromium and ChromeDriver were properly configured. In this environment, Selenium successfully rendered the page, allowing the extraction of Market Banner elements and Latest News items. The script detected multiple market cards and news entries, confirming that Selenium can correctly retrieve JavaScript-rendered content when the browser environment is properly supported.



```
Reading HTML file...
Filtering fields: Market Banner
Found 5 market cards
Filtering fields: Latest News
Found 6 news items
Storing Market data into CSV...
Market CSV created: ./market_data.csv
Storing News data into CSV...
News CSV created: ./news_data.csv

--- Data Filtering Complete ---
```



```
| marketCard_symbol   | marketCard_stockPosition    | marketCard_changePct    |
|:--------------------|:----------------------------|:------------------------|
| STOXX600*           | 614.38                      | -0.03%                  |
| DAX*                | 25,297.13                   | -0.22%                  |
| FTSE*               | 10,235.29                   | -0.04%                  |
| CAC*                | 8,258.94                    | -0.65%                  |
| FTSE MIB*           | 45,799.69                   | -0.11%                  |
```

This screenshot displays the contents of the extracted Market Banner data after successful Selenium execution. The table includes market symbols, stock positions, and percentage changes for multiple indices. This output confirms that the Market Banner data is accessible only after JavaScript execution and validates the correctness of the scraping and filtering logic used in the Python scripts.