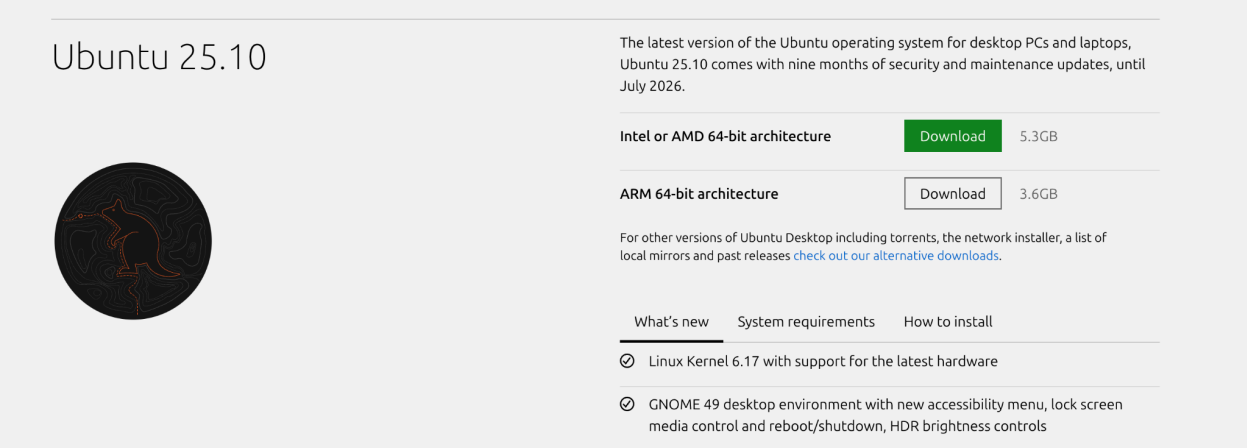
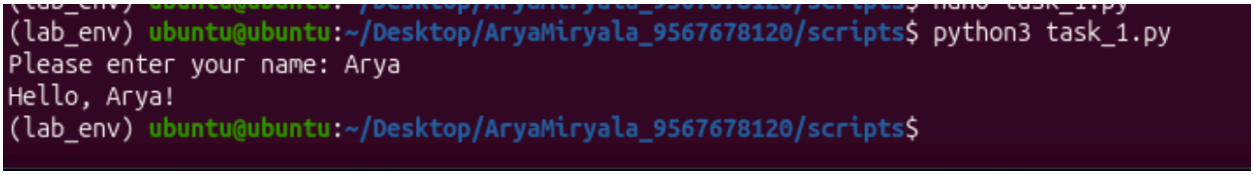
Name: Arya Miryala Lab 1 Report USC ID: 9567678120

To download Ubuntu on my M1 Mac I downloaded VMWare 13.5.2

After VMware was installed, I downloaded the Ubuntu Desktop ISO image. I installed Ubuntu 25.10 and the ARM 64-bit architecture.



Next I went on terminal and created my project folder. I first created a Python virtual environment to install all the libraries needed and worked on task_1.py



Task 2.3: Web Scraper Implementation

Objective: To programmatically retrieve the raw HTML source code from the CNBC World News website.

Implementation (`web_scraper.py`): I used the `requests` library to send an HTTP GET request to the target URL and `BeautifulSoup` to handle the content.

Challenges & Solutions:

- **The 403 Forbidden Error:** Initially, the CNBC server rejected the request with a 403 status code. This is a security measure to block automated bots.
- **The Fix:** I inspected the request headers and added a custom `User-Agent` string mimicking a Google Chrome browser (`Mozilla/5.0...`). This successfully bypassed the bot detection, resulting in a **200 OK** status code and allowing me to save the HTML to `data/raw_data/web_data.html`.

In the screenshot above, you can also see that I printed the first 10 lines of the created html file.

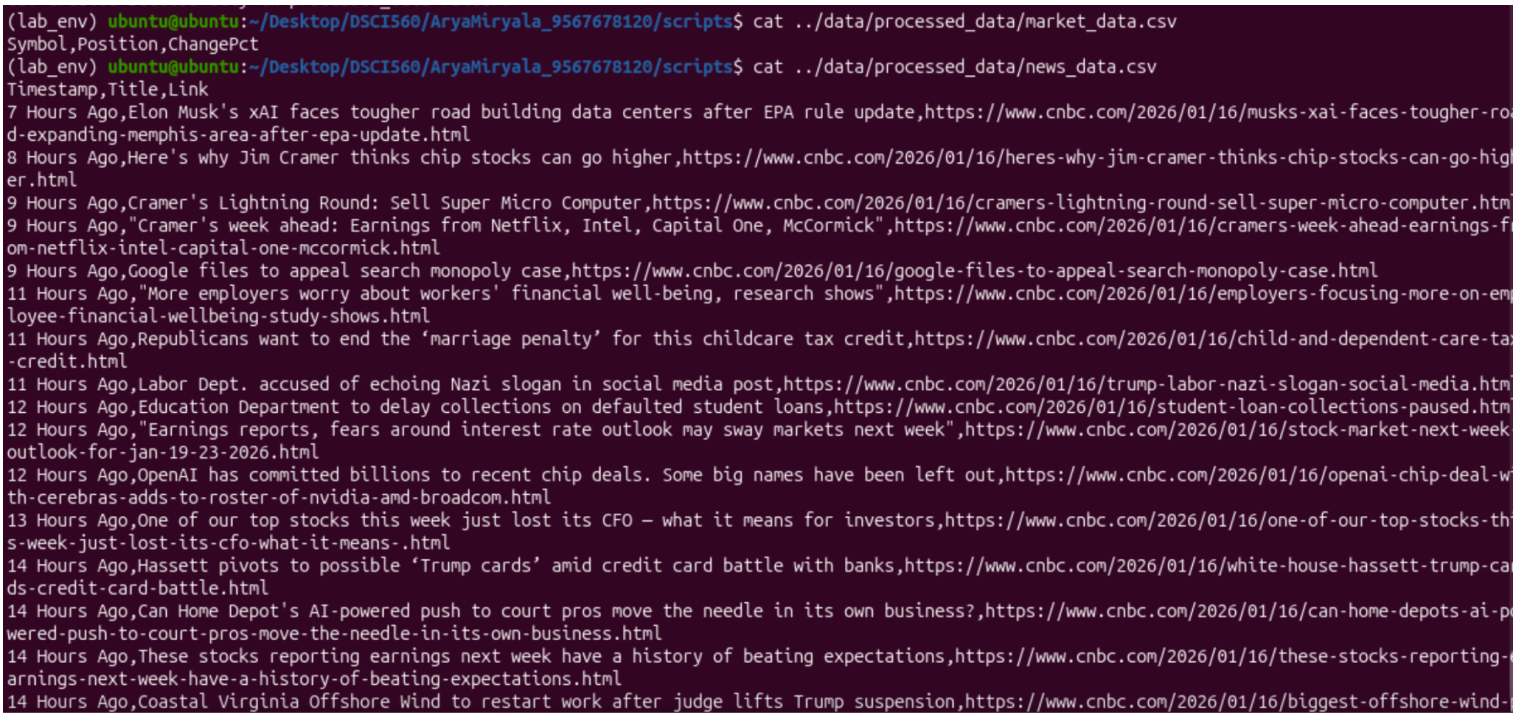
Task 2.4: Data Parsing & Results

Objective: Parse the raw HTML to extract "Latest News" and "Market Data" into CSV files.
Implementation: I used `BeautifulSoup` to parse the HTML.

- **News Data:** Successfully extracted timestamps, headlines, and links into `news_data.csv`.
- **Market Data:** Targeted the market card containers, but the resulting `market_data.csv` did not contain numerical values.

Why the Market Data is Missing: The scraper uses the `requests` library, which downloads the **static HTML** source code exactly as it exists on the server. However, CNBC loads its financial data (stock prices and changes) **dynamically** using JavaScript that runs *after* the page loads in a browser. Because `requests` does not execute JavaScript, my script captured the empty container tags (the "skeleton" of the page) but could not retrieve the numerical data that is injected later.

Name: Arya Miryala Lab 1 Report USC ID: 9567678120



Github repo to find source code: https://github.com/aryamiryala/DSCI560/tree/main