

To download Ubuntu on my M1 Mac I downloaded VMWare 13.5.2

VMware Fusion (for Intel-based and Apple silicon Macs)	Release 13.5.2	Release Level Info 520443		
File Name	Release Date	Last Updated	SHA2	MD5
VMware Fusion 13.5.2 (for Intel-based and Apple silicon Macs) VMware-Fusion-13.5.2-23775688_universal.dmg(750.23 MB) Build Number: 23775688	May 14, 2024	May 10, 2024	baaa201c797af8e32a2ec3ae78c69bfe dbe5c5c7960c3673885bd84e42ddfb9	0601c5b92139b3cbab3eba68217f6576

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

Ubuntu 25.10



The latest version of the Ubuntu operating system for desktop PCs and laptops, Ubuntu 25.10 comes with nine months of security and maintenance updates, until July 2026.

Intel or AMD 64-bit architecture	Download	5.3GB
ARM 64-bit architecture	Download	3.6GB

For other versions of Ubuntu Desktop including torrents, the network installer, a list of local mirrors and past releases [check out our alternative downloads](#).

[What's new](#) [System requirements](#) [How to install](#)

- ⊖ Linux Kernel 6.17 with support for the latest hardware
- ⊖ GNOME 49 desktop environment with new accessibility menu, lock screen media control and reboot/shutdown, HDR brightness controls

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

```
(lab_env) ubuntu@ubuntu:~/Desktop/AryaMiryala_9567678120/scripts$ python3 task_1.py
Please enter your name: Arya
Hello, Arya!
(lab_env) ubuntu@ubuntu:~/Desktop/AryaMiryala_9567678120/scripts$
```

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`.

```
(lab_env) ubuntu@ubuntu:~/Desktop/AryaMiryala_9567678120/scripts$ nano web_scraper.py
(lab_env) ubuntu@ubuntu:~/Desktop/AryaMiryala_9567678120/scripts$ python3 web_scraper.py
Successfully saved data to ..../data/raw_data/web_data.html
(lab_env) ubuntu@ubuntu:~/Desktop/AryaMiryala_9567678120/scripts$ head -n 10 ~/Desktop/AryaMiryala_9567678120/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"/><meta content="https://fm.cnbc.com/applications/cnbc.com/staticcontent/img/versant/cnbc_share_versant.png?v=1524171804&w=1920&h=1080" property="image" type="image"/><meta content="https://fm.cnbc.com/applications/cnbc.com/staticcontent/img/versant/cnbc_share_versant.png?v=1524171804&w=1920&h=1080" name="twitter:image:src"/><meta content="https://fm.cnbc.com/applications/cnbc.com/staticcontent/img/versant/cnbc_share_versant.png?v=1524171804&w=1920&h=1080" name="twitter:image" type="image"/><meta content="text/html" type="text/html"/><meta content="summary_large_image" name="twitter:card"/><meta content="@CNBC" name="twitter:site"/><meta content="https://www.cnbc.com/world/" name="twitter:url"/><meta content="International: Top News And Analysis" name="twitter:title"/><meta content="CNBC International is the world leader for news on business, technology, China, trade, oil prices, the Middle East and markets." name="twitter:description"/><meta content="CNBC" name="twitter:creator"/><link href="https://fm.cnbc.com/applications/cnbc.com/resources/img/CNBC_LOGO_FAVICON_IC_KO_RGB.ico" rel="icon" type="image/x-icon"/><style type="text/css">@charset "UTF-8"; .Modal-modalBackground{background:#000000b3;height:100%;left:0;overflow-y:auto;position:fixed;top:0;transition:background-color .4s;width:100%;z-index:100001}.Modal-bottomModal.Modal-modal[background:#f0f0f0;border-radius:3px;bottom:0;box-shadow:5px 5px 20px #1717171a;display:inline-block;height:528px;left:0;margin-top:0!important;max-width:100%;position:fixed;top:0;transform:none;width:100%}{.Modal-bottomModal.Modal-modal[height:100%;position:relative;top:0]}.Modal-modal[background-color:#fff;border-radius:3px;box-shadow:5px 5px 20px #1717171a;display:inline-block;left:50%;margin-top:5vh;max-width:100%;overflow:auto;position:relative;transform:translateY(-50%)</style>
```

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.

```
(lab_env) ubuntu@ubuntu:~/Desktop/DSCI560/AryaMiryala_9567678120/scripts$ cat ../data/processed_data/market_data.csv
Symbol,Position,ChangePct
(lab_env) ubuntu@ubuntu:~/Desktop/DSCI560/AryaMiryala_9567678120/scripts$ cat ../data/processed_data/news_data.csv
Timestamp,Title,Link
7 Hours Ago,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-memphis-area-after-epa-update.html
8 Hours Ago,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-can-go-higher.html
9 Hours Ago,Cramer's Lightning Round: Sell Super Micro Computer,https://www.cnbc.com/2026/01/16/cramers-lightning-round-sell-super-micro-computer.html
9 Hours Ago,"Cramer's week ahead: Earnings from Netflix, Intel, Capital One, McCormick",https://www.cnbc.com/2026/01/16/cramers-week-ahead-earnings-from-netflix-intel-capital-one-mccormick.html
9 Hours Ago,Google files to appeal search monopoly case,https://www.cnbc.com/2026/01/16/google-files-to-appeal-search-monopoly-case.html
11 Hours Ago,"More employers worry about workers' financial well-being, research shows",https://www.cnbc.com/2026/01/16/employers-focusing-more-on-employee-financial-wellbeing-study-shows.html
11 Hours Ago,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
11 Hours Ago,Labor Dept. accused of echoing Nazi slogan in social media post,https://www.cnbc.com/2026/01/16/trump-labor-nazi-slogan-social-media.html
12 Hours Ago,Education Department to delay collections on defaulted student loans,https://www.cnbc.com/2026/01/16/student-loan-collections-paused.html
12 Hours Ago,"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-19-23-2026.html
12 Hours Ago,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-roster-of-nvidia-amd-broadcom.html
13 Hours Ago,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-its-cfo-what-it-means-.html
14 Hours Ago,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-card-battle.html
14 Hours Ago,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-court-pros-move-the-needle-in-its-own-business.html
14 Hours Ago,These stocks reporting earnings next week have a history of beating expectations,https://www.cnbc.com/2026/01/16/these-stocks-reporting-earnings-next-week-have-a-history-of-beating-expectations.html
14 Hours Ago,Coastal Virginia Offshore Wind to restart work after judge lifts Trump suspension,https://www.cnbc.com/2026/01/16/biggest-offshore-wind-
```

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