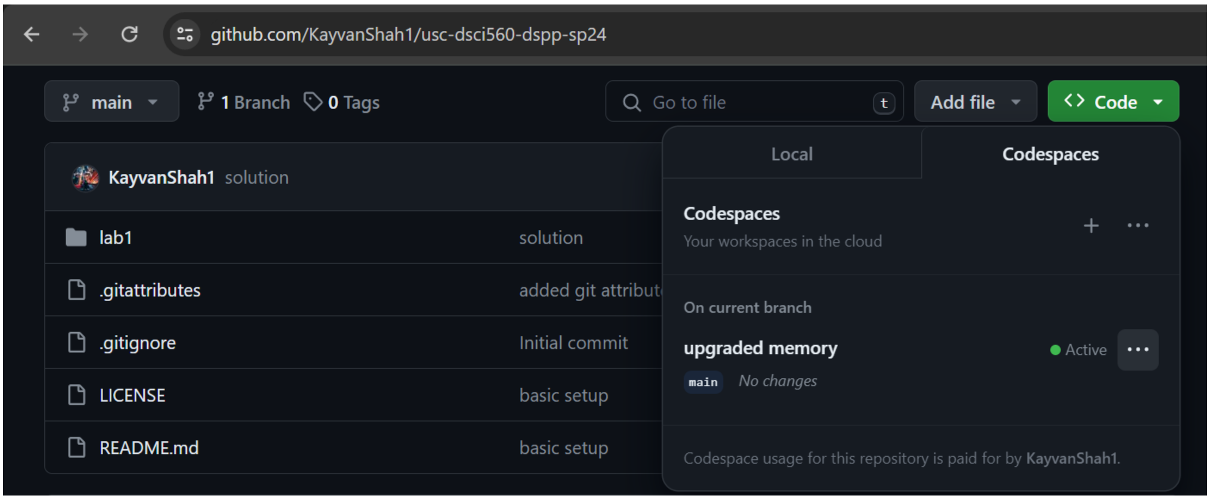
Lab 1 Solution

Installation & Setup

- Development done locally in a local git repository on a windows machine.
- The code is synced with a remote git repository.

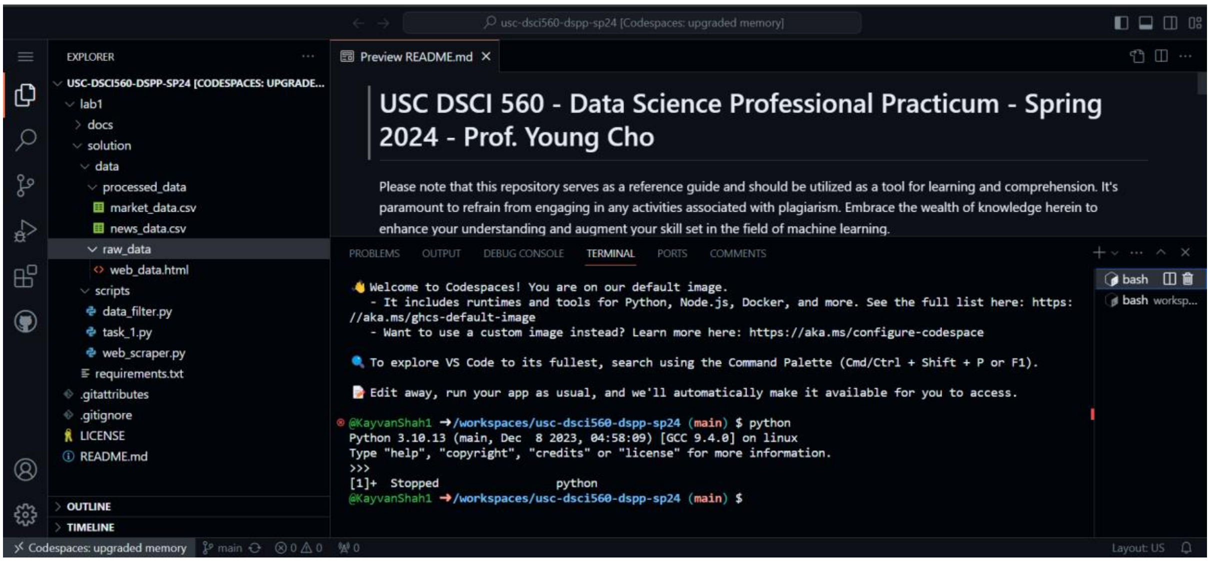


github.com/KayvanShah1/usc-dsci560-dspp-sp24
main
1 Branch
0 Tags
Go to file
t
Add file
Code
Local
Codespaces
Codespaces
Your workspaces in the cloud
On current branch
upgraded memory
Active
main
No changes
Codespace usage for this repository is paid for by KayvanShah1.

KayvanShah1 solution

lab1	solution
.gitattributes	added git attribut
.gitignore	Initial commit
LICENSE	basic setup
README.md	basic setup

- A Linux Ubuntu instance is spin up using GitHub Codespace for the repository.



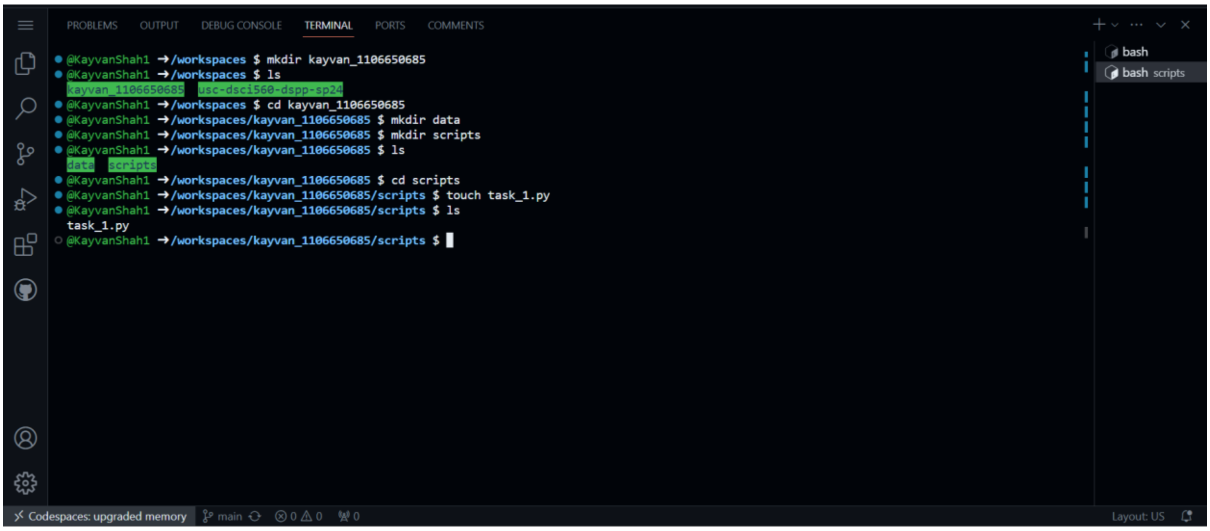
USC DSCI 560 - Data Science Professional Practicum - Spring 2024 - Prof. Young Cho

Please note that this repository serves as a reference guide and should be utilized as a tool for learning and comprehension. It's paramount to refrain from engaging in any activities associated with plagiarism. Embrace the wealth of knowledge herein to enhance your understanding and augment your skill set in the field of machine learning.

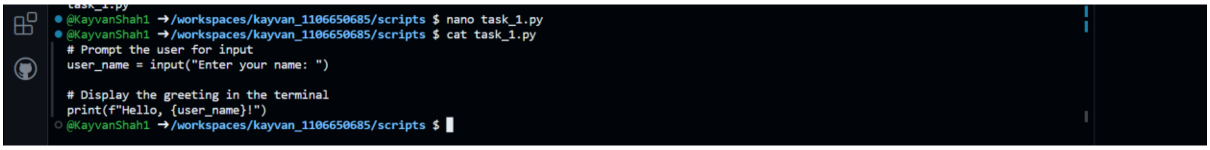
```
👋 Welcome to Codespaces! You are on our default image.
   - It includes runtimes and tools for Python, Node.js, Docker, and more. See the full list here: https:
//aka.ms/ghcs-default-image
   - Want to use a custom image instead? Learn more here: https://aka.ms/configure-codespace

🔍 To explore VS Code to its fullest, search using the Command Palette (Cmd/Ctrl + Shift + P or F1).

📝 Edit away, run your app as usual, and we'll automatically make it available for you to access.

@KayvanShah1 ➜ /workspaces/usc-dsci560-dspp-sp24 (main) $ python
Python 3.10.13 (main, Dec  8 2023, 04:58:09) [GCC 9.4.0] on linux
Type "help", "copyright", "credits" or "license" for more information.
>>>
[1]+  Stopped                 python
@KayvanShah1 ➜ /workspaces/usc-dsci560-dspp-sp24 (main) $
```

Get Familiar with Linux and Python

Playing around with Linux Terminal

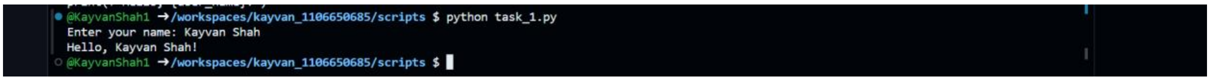


```
● @KayvanShah1 →/workspaces $ mkdir kayvan_1106650685
● @KayvanShah1 →/workspaces $ ls
  kayvan_1106650685  usc-dsci560-dspp-sp24
● @KayvanShah1 →/workspaces $ cd kayvan_1106650685
● @KayvanShah1 →/workspaces/kayvan_1106650685 $ mkdir data
● @KayvanShah1 →/workspaces/kayvan_1106650685 $ mkdir scripts
● @KayvanShah1 →/workspaces/kayvan_1106650685 $ ls
  data  scripts
● @KayvanShah1 →/workspaces/kayvan_1106650685 $ cd scripts
● @KayvanShah1 →/workspaces/kayvan_1106650685/scripts $ touch task_1.py
● @KayvanShah1 →/workspaces/kayvan_1106650685/scripts $ ls
  task_1.py
○ @KayvanShah1 →/workspaces/kayvan_1106650685/scripts $ 
```



```
● @KayvanShah1 →/workspaces/kayvan_1106650685/scripts $ nano task_1.py
● @KayvanShah1 →/workspaces/kayvan_1106650685/scripts $ cat task_1.py
  # Prompt the user for input
  user_name = input("Enter your name: ")

  # Display the greeting in the terminal
  print(f"Hello, {user_name}!")
○ @KayvanShah1 →/workspaces/kayvan_1106650685/scripts $ 
```

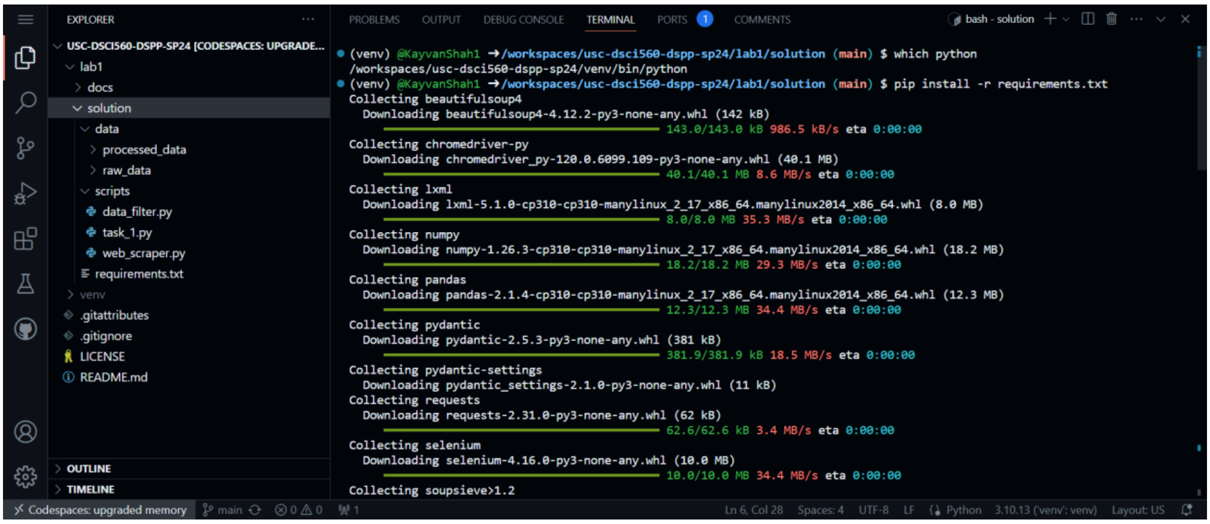
A basic Python Script



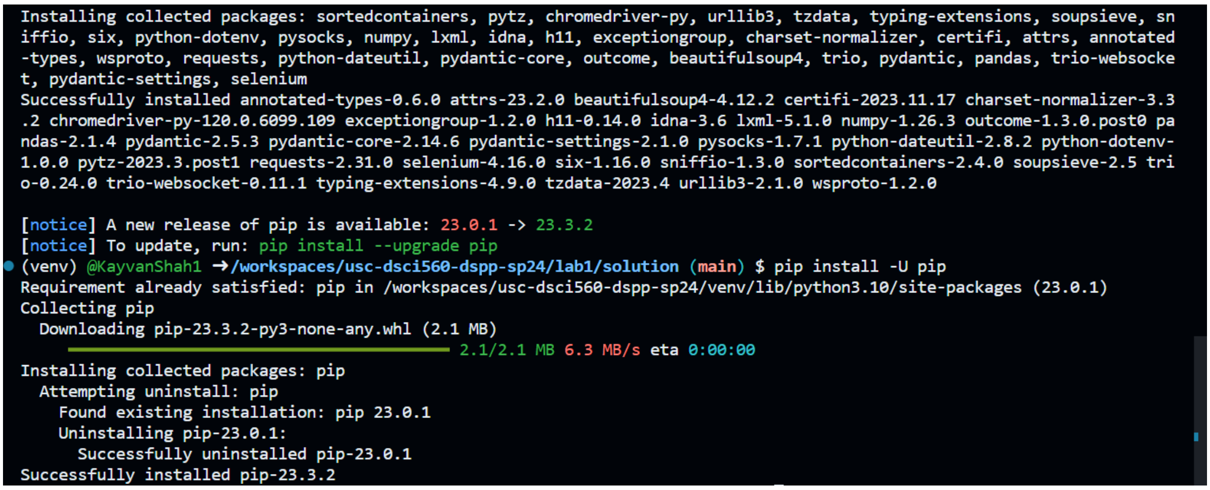
```
● @KayvanShah1 →/workspaces/kayvan_1106650685/scripts $ python task_1.py
  Enter your name: Kayvan Shah
  Hello, Kayvan Shah!
○ @KayvanShah1 →/workspaces/kayvan_1106650685/scripts $ 
```

Web Scraping Task

- Create a virtual environment and install the dependencies using the ***requirements.txt***.
  - *venv* – Created a virtual environment using command $python - m\ venv\ venv$
  - Linux has a built-in support for Chrome and using a library **chromedriver-py** we automatically manage the webdriver downloads and retrieval of the executable path while initializing the web driver.

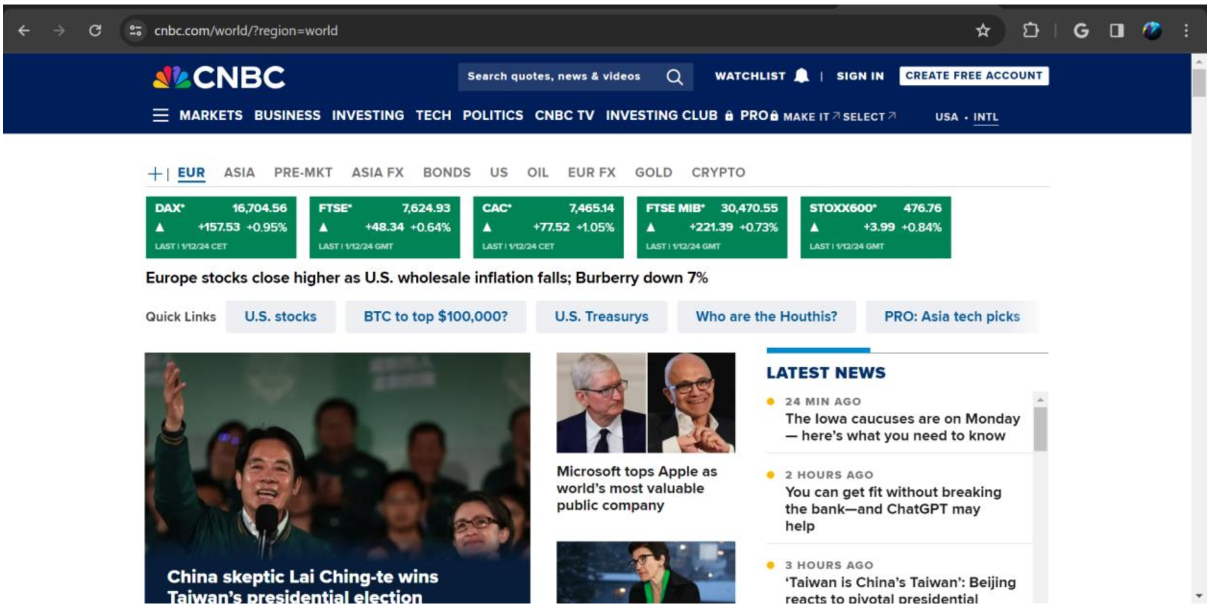


```
● (venv) @KayvanShah1 →/workspaces/usc-dsci560-dspp-sp24/lab1/solution (main) $ which python
  /workspaces/usc-dsci560-dspp-sp24/venv/bin/python
● (venv) @KayvanShah1 →/workspaces/usc-dsci560-dspp-sp24/lab1/solution (main) $ pip install -r requirements.txt
  Collecting beautifulsoup4
    Downloading beautifulsoup4-4.12.2-py3-none-any.whl (142 kB)
       ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 143.0/143.0 kB 986.5 kB/s eta 0:00:00
  Collecting chromedriver-py
    Downloading chromedriver_py-120.0.6099.109-py3-none-any.whl (40.1 MB)
       ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 40.1/40.1 MB 8.6 MB/s eta 0:00:00
  Collecting lxml
    Downloading lxml-5.1.0-cp310-cp310-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (8.0 MB)
       ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 8.0/8.0 MB 35.3 MB/s eta 0:00:00
  Collecting numpy
    Downloading numpy-1.26.3-cp310-cp310-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (18.2 MB)
       ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 18.2/18.2 MB 29.3 MB/s eta 0:00:00
  Collecting pandas
    Downloading pandas-2.1.4-cp310-cp310-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (12.3 MB)
       ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 12.3/12.3 MB 34.4 MB/s eta 0:00:00
  Collecting pydantic
    Downloading pydantic-2.5.3-py3-none-any.whl (381 kB)
       ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 381.9/381.9 kB 18.5 MB/s eta 0:00:00
  Collecting pydantic-settings
    Downloading pydantic_settings-2.1.0-py3-none-any.whl (11 kB)
  Collecting requests
    Downloading requests-2.31.0-py3-none-any.whl (62 kB)
       ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 62.6/62.6 kB 3.4 MB/s eta 0:00:00
  Collecting selenium
    Downloading selenium-4.16.0-py3-none-any.whl (10.0 MB)
       ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 10.0/10.0 MB 34.4 MB/s eta 0:00:00
  Collecting soupsieve>1.2
```



```
Installing collected packages: sortedcontainers, pytz, chromedriver-py, urllib3, tzdata, typing-extensions, soupsieve, sn
iffio, six, python-dotenv, pysocks, numpy, lxml, idna, h11, exceptiongroup, charset-normalizer, certifi, attrs, annotated
-types, wsproto, requests, python-dateutil, pydantic-core, outcome, beautifulsoup4, trio, pydantic, pandas, trio-websocke
t, pydantic-settings, selenium
Successfully installed annotated-types-0.6.0 attrs-23.2.0 beautifulsoup4-4.12.2 certifi-2023.11.17 charset-normalizer-3.3
.2 chromedriver-py-120.0.6099.109 exceptiongroup-1.2.0 h11-0.14.0 idna-3.6 lxml-5.1.0 numpy-1.26.3 outcome-1.3.0.post0 pa
ndas-2.1.4 pydantic-2.5.3 pydantic-core-2.14.6 pydantic-settings-2.1.0 pysocks-1.7.1 python-dateutil-2.8.2 python-dotenv-
1.0.0 pytz-2023.3.post1 requests-2.31.0 selenium-4.16.0 six-1.16.0 sniffio-1.3.0 sortedcontainers-2.4.0 soupsieve-2.5 tri
o-0.24.0 trio-websocket-0.11.1 typing-extensions-4.9.0 tzdata-2023.4 urllib3-2.1.0 wsproto-1.2.0

[notice] A new release of pip is available: 23.0.1 -> 23.3.2
[notice] To update, run: pip install --upgrade pip
(venv) @KayvanShah1 ➜/workspaces/usc-dsci560-dspp-sp24/lab1/solution (main) $ pip install -U pip
Requirement already satisfied: pip in /workspaces/usc-dsci560-dspp-sp24/venv/lib/python3.10/site-packages (23.0.1)
Collecting pip
  Downloading pip-23.3.2-py3-none-any.whl (2.1 MB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 2.1/2.1 MB 6.3 MB/s eta 0:00:00
Installing collected packages: pip
  Attempting uninstall: pip
    Found existing installation: pip 23.0.1
    Uninstalling pip-23.0.1:
      Successfully uninstalled pip-23.0.1
Successfully installed pip-23.3.2
```

- A snapshot of the webpage from where the data is to be scraped.
  - The webpage loads a few fields dynamically. In this case the market banner is loaded using a JavaScript.
  - To capture the data from dynamically populated HTML elements we must use selenium with a web driver suited for the installed browser on the system.
  - While the latest news isn't loaded dynamically hence, we can get the static snapshot of the HTML tree of the scrape that its section's data.



cnbc.com/world/?region=world
CNBC
Search quotes, news & videos
WATCHLIST
SIGN IN
CREATE FREE ACCOUNT
MARKETS BUSINESS INVESTING TECH POLITICS CNBC TV INVESTING CLUB PRO MAKE IT SELECT
USA · INTL
EUR ASIA PRE-MKT ASIA FX BONDS US OIL EUR FX GOLD CRYPTO
Quick Links

DAX* 16,704.56 ▲ +157.53 +0.95% LAST | 1/12/24 CET

FTSE* 7,624.93 ▲ +48.34 +0.64% LAST | 1/12/24 GMT

CAC* 7,465.14 ▲ +77.52 +1.05% LAST | 1/12/24 CET

FTSE MIB* 30,470.55 ▲ +221.39 +0.73% LAST | 1/12/24 GMT

STOXX600* 476.76 ▲ +3.99 +0.84% LAST | 1/12/24 GMT

Europe stocks close higher as U.S. wholesale inflation falls; Burberry down 7%

U.S. stocks | BTC to top $100,000? | U.S. Treasurys | Who are the Houthis? | PRO: Asia tech picks

China skeptic Lai Ching-te wins Taiwan's presidential election

Microsoft tops Apple as world's most valuable public company

LATEST NEWS

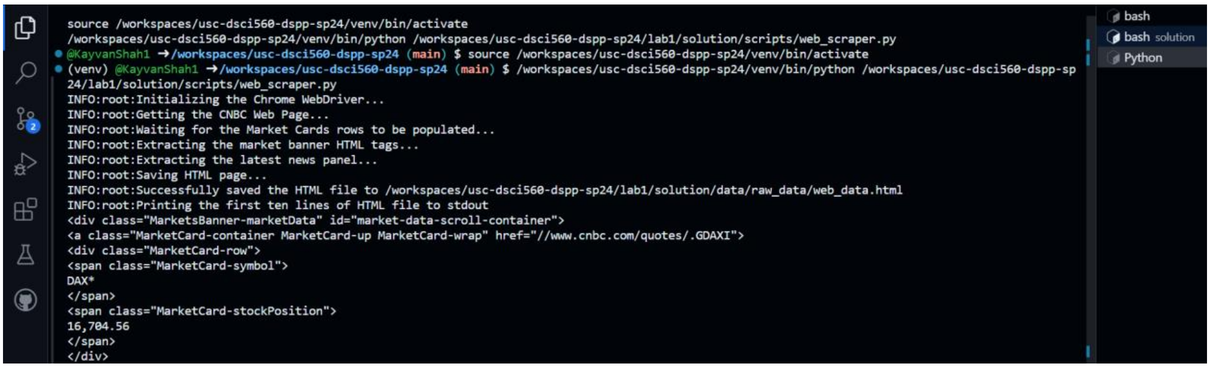
24 MIN AGO
The Iowa caucuses are on Monday — here's what you need to know

2 HOURS AGO
You can get fit without breaking the bank—and ChatGPT may help

3 HOURS AGO
'Taiwan is China's Taiwan': Beijing reacts to pivotal presidential

- Scraping the raw data
  - This Python script is designed to scrape data from the CNBC website related to global markets. It begins by configuring logging for informative messages and defining file and directory paths.
  - The script utilizes the `pydantic_settings` module to manage configuration settings, with a default base URL for CNBC. It then initializes a headless Chrome WebDriver using `chromedriver_py` and configurable options.
  - The main functionality involves loading the CNBC webpage, waiting for the visibility of market card rows, and extracting HTML content for the market banner and the latest news panel using BeautifulSoup.

  - The extracted HTML is saved to a file in the "raw_data" directory. As part of error handling, the script logs any exceptions that might occur during data fetching.
  - Finally, it prints the first ten lines of the saved HTML file to the standard output for quick verification.
  - Overall, this script demonstrates the use of web scraping tools like Selenium and BeautifulSoup, with logging providing insights into the different stages of the data retrieval process.



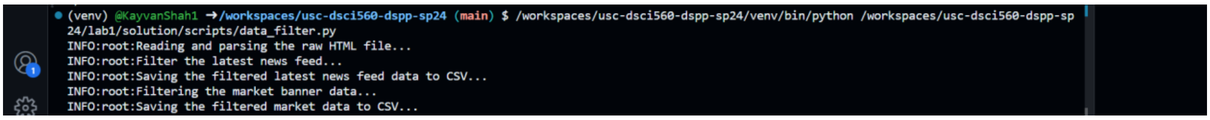
```
source /workspaces/usc-dsci560-dspp-sp24/venv/bin/activate
/workspaces/usc-dsci560-dspp-sp24/venv/bin/python /workspaces/usc-dsci560-dspp-sp24/lab1/solution/scripts/web_scraper.py
● @KayvanShah1 →/workspaces/usc-dsci560-dspp-sp24 (main) $ source /workspaces/usc-dsci560-dspp-sp24/venv/bin/activate
● (venv) @KayvanShah1 →/workspaces/usc-dsci560-dspp-sp24 (main) $ /workspaces/usc-dsci560-dspp-sp24/venv/bin/python /workspaces/usc-dsci560-dspp-sp
24/lab1/solution/scripts/web_scraper.py
INFO:root:Initializing the Chrome WebDriver...
INFO:root:Getting the CNBC Web Page...
INFO:root:Waiting for the Market Cards rows to be populated...
INFO:root:Extracting the market banner HTML tags...
INFO:root:Extracting the latest news panel...
INFO:root:Saving HTML page...
INFO:root:Successfully saved the HTML file to /workspaces/usc-dsci560-dspp-sp24/lab1/solution/data/raw_data/web_data.html
INFO:root:Printing the first ten lines of HTML file to stdout
<div class="MarketsBanner-marketData" id="market-data-scroll-container">
<a class="MarketCard-container MarketCard-up MarketCard-wrap" href="//www.cnbc.com/quotes/.GDAXI">
<div class="MarketCard-row">
<span class="MarketCard-symbol">
DAX*
</span>
<span class="MarketCard-stockPosition">
16,704.56
</span>
</div>
```

- Filtering the data to extract the fields of interest.

  This Python script performs the following tasks:

  1. Logging Setup: Configures logging to display INFO level messages.
  2. Path and Directory Definitions: Defines the file and directory paths using the `Path` class.
  3. HTML Parsing and Data Extraction:
     a. “read_parse_raw_data”: Reads and parses an HTML file using BeautifulSoup.
     b. “NewsItem” and “MarketCard” Pydantic models are defined for structured data representation.
  4. Data Filtering Functions:
     a. “filter_latest_news”: Extracts the latest news feed from the HTML, converts it into “NewsItem” Pydantic model instances, and appends them to a list.
     b. “filter_market_banner”: Filters market banner data from the HTML, converts it into “MarketCard” Pydantic model instances, and appends them to a list.
  5. Main Execution:
     a. Reads and parses an HTML file (“web_data.html”).
     b. Filters the latest news and market banner data using the defined functions.
     c. Logs the process at different stages (reading, filtering, and saving).
     d. Saves the filtered data into CSV files (“news_data.csv” and “market_data.csv”) in the "processed_data" directory.

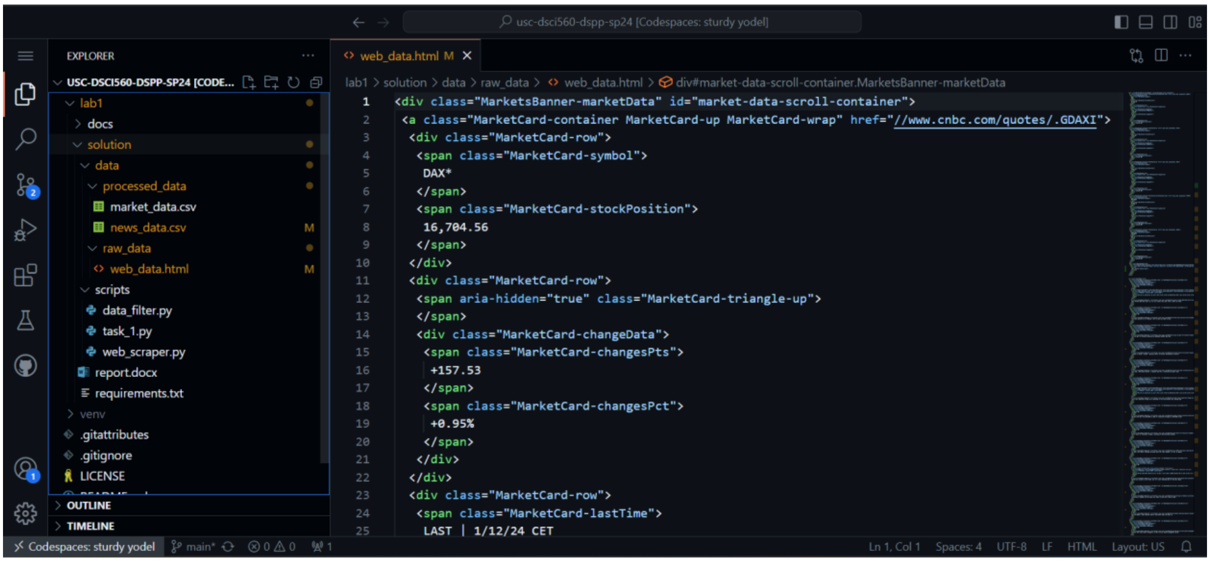
  Overall, the script demonstrates the use of BeautifulSoup for HTML parsing, Pydantic for structured data representation, and Pandas for data manipulation. Logging is employed to provide information about the different stages of the process, making it easier to understand and debug the execution flow.



```
● (venv) @KayvanShah1 →/workspaces/usc-dsci560-dspp-sp24 (main) $ /workspaces/usc-dsci560-dspp-sp24/venv/bin/python /workspaces/usc-dsci560-dspp-sp
24/lab1/solution/scripts/data_filter.py
INFO:root:Reading and parsing the raw HTML file...
INFO:root:Filter the latest news feed...
INFO:root:Saving the filtered latest news feed data to CSV...
INFO:root:Filtering the market banner data...
INFO:root:Saving the filtered market data to CSV...
```

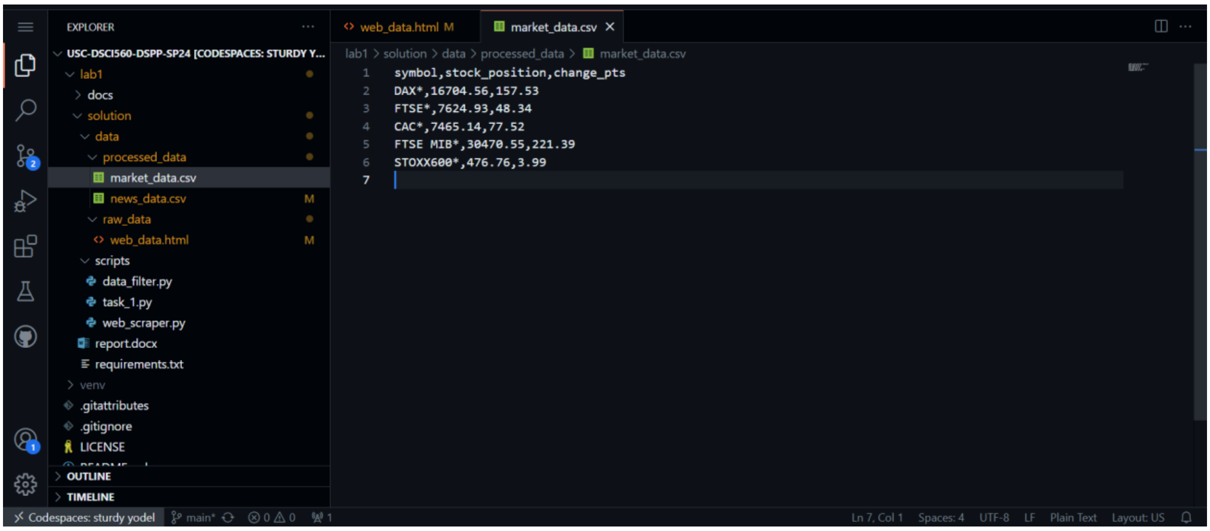
Output Files

- Raw data output HTML file



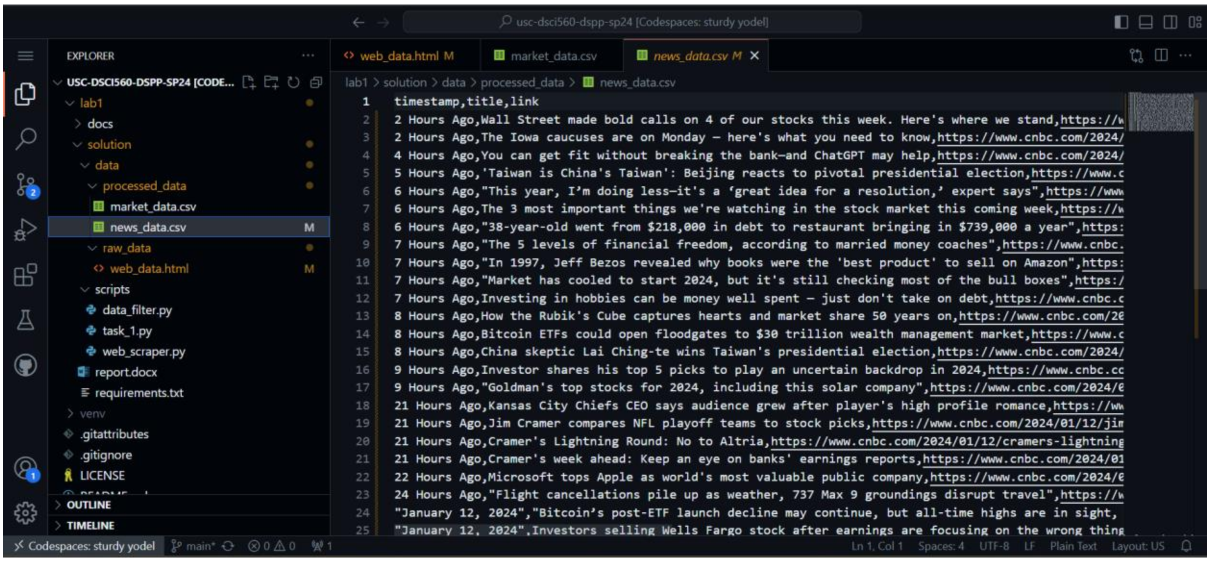
```html
<div class="MarketsBanner-marketData" id="market-data-scroll-container">
 <a class="MarketCard-container MarketCard-up MarketCard-wrap" href="//www.cnbc.com/quotes/.GDAXI">
  <div class="MarketCard-row">
   <span class="MarketCard-symbol">
    DAX*
   </span>
   <span class="MarketCard-stockPosition">
    16,704.56
   </span>
  </div>
  <div class="MarketCard-row">
   <span aria-hidden="true" class="MarketCard-triangle-up">
   </span>
   <div class="MarketCard-changeData">
    <span class="MarketCard-changesPts">
     +157.53
    </span>
    <span class="MarketCard-changesPct">
     +0.95%
    </span>
   </div>
  </div>
  <div class="MarketCard-row">
   <span class="MarketCard-lastTime">
    LAST | 1/12/24 CET
```

- Processed data CSV files
    - Market data
        - Has the following fields:
            - symbol – abbreviations for a stock ticker.
            - stock_position – current/latest price of the stock.
            - change_pts – most recent change in price observed.



```
symbol,stock_position,change_pts
DAX*,16704.56,157.53
FTSE*,7624.93,48.34
CAC*,7465.14,77.52
FTSE MIB*,30470.55,221.39
STOXX600*,476.76,3.99
```

- Latest news
    - Has the following fields:
        - timestamp – approximate time from the current time when the news was posted.
        - title – the headline of the news column
        - link – link to the official source of news on the website



```
timestamp,title,link
2 Hours Ago,Wall Street made bold calls on 4 of our stocks this week. Here's where we stand,https://w
2 Hours Ago,The Iowa caucuses are on Monday — here's what you need to know,https://www.cnbc.com/2024/
4 Hours Ago,You can get fit without breaking the bank—and ChatGPT may help,https://www.cnbc.com/2024/
5 Hours Ago,'Taiwan is China's Taiwan': Beijing reacts to pivotal presidential election,https://www.c
6 Hours Ago,"This year, I’m doing less—it's a ‘great idea for a resolution,’ expert says",https://www
6 Hours Ago,The 3 most important things we're watching in the stock market this coming week,https://w
6 Hours Ago,"38-year-old went from $218,000 in debt to restaurant bringing in $739,000 a year",https:
7 Hours Ago,"The 5 levels of financial freedom, according to married money coaches",https://www.cnbc.
7 Hours Ago,"In 1997, Jeff Bezos revealed why books were the 'best product' to sell on Amazon",https:
7 Hours Ago,"Market has cooled to start 2024, but it's still checking most of the bull boxes",https:/
7 Hours Ago,Investing in hobbies can be money well spent — just don't take on debt,https://www.cnbc.c
8 Hours Ago,How the Rubik's Cube captures hearts and market share 50 years on,https://www.cnbc.com/20
8 Hours Ago,Bitcoin ETFs could open floodgates to $30 trillion wealth management market,https://www.c
8 Hours Ago,China skeptic Lai Ching-te wins Taiwan's presidential election,https://www.cnbc.com/2024/
9 Hours Ago,Investor shares his top 5 picks to play an uncertain backdrop in 2024,https://www.cnbc.cc
9 Hours Ago,"Goldman's top stocks for 2024, including this solar company",https://www.cnbc.com/2024/0
21 Hours Ago,Kansas City Chiefs CEO says audience grew after player's high profile romance,https://ww
21 Hours Ago,Jim Cramer compares NFL playoff teams to stock picks,https://www.cnbc.com/2024/01/12/jim
21 Hours Ago,Cramer's Lightning Round: No to Altria,https://www.cnbc.com/2024/01/12/cramers-lightning
21 Hours Ago,Cramer's week ahead: Keep an eye on banks' earnings reports,https://www.cnbc.com/2024/01
22 Hours Ago,Microsoft tops Apple as world's most valuable public company,https://www.cnbc.com/2024/0
24 Hours Ago,"Flight cancellations pile up as weather, 737 Max 9 groundings disrupt travel",https://w
"January 12, 2024","Bitcoin’s post-ETF launch decline may continue, but all-time highs are in sight,
"January 12, 2024",Investors selling Wells Fargo stock after earnings are focusing on the wrong thing
```