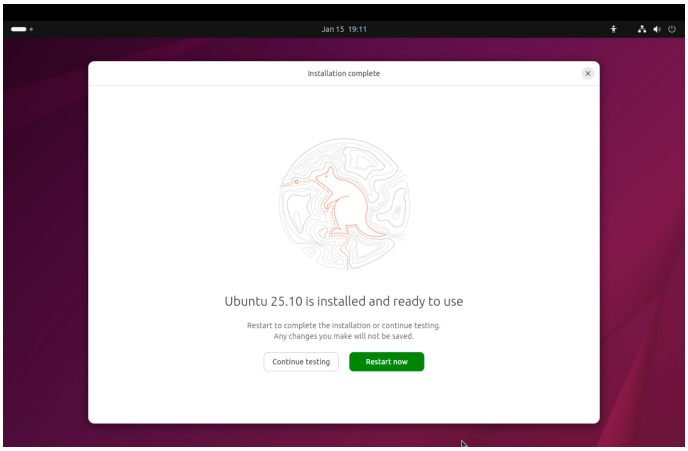
DSCI 560 – Laboratory Assignment 1

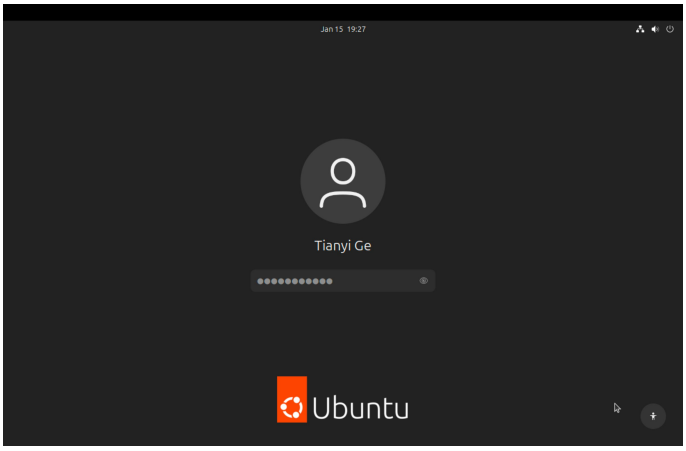
Name: Tianyi Ge
USC ID: 5804514679
Github Link:https://github.com/Mia0403/DSCI560-Lab1.git

1. Installation and Setup

- VMware & Ubuntu Installation

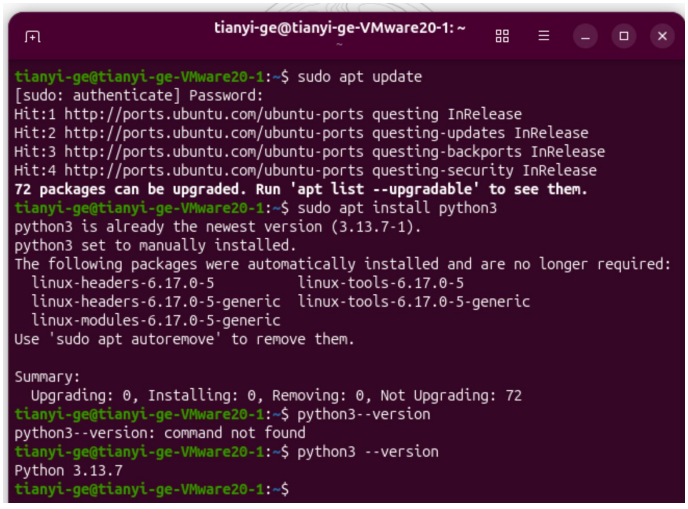
I installed VMware Fusion on macOS and created an Ubuntu 64-bit virtual machine. The virtual machine was successfully launched and used throughout this lab.

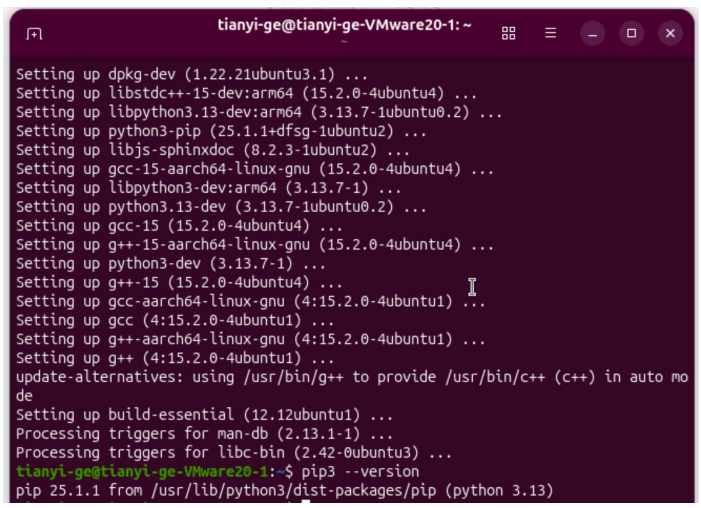




- Python Installation

Python 3 and pip were installed and verified in the Ubuntu virtual machine.





2. Get Familiar with Linux and Python

- Playing around with Linux Terminal

A new directory named d "<your name>_<your USC ID>" (A new directory named "TianyiGe_5804514679" was created on the desktop.) was created on the desktop.Inside the directory, two subdirectories named data and scripts were created, and an empty Python file named task_1.py was added to the scripts folder.

```
tianyi-ge@tianyi-ge-VMware20-1:~$ cd Desktop
tianyi-ge@tianyi-ge-VMware20-1:~/Desktop$ mkdir Yourname_YourUSCID
tianyi-ge@tianyi-ge-VMware20-1:~/Desktop$ cd ^C
tianyi-ge@tianyi-ge-VMware20-1:~/Desktop$ cd Yourname_YourUSCID
tianyi-ge@tianyi-ge-VMware20-1:~/Desktop/Yourname_YourUSCID$ mkdir data scripts
tianyi-ge@tianyi-ge-VMware20-1:~/Desktop/Yourname_YourUSCID$ cd scripts
tianyi-ge@tianyi-ge-VMware20-1:~/Desktop/Yourname_YourUSCID/scripts$ touch task_
1.py
tianyi-ge@tianyi-ge-VMware20-1:~/Desktop/Yourname_YourUSCID/scripts$ ls
task_1.py
```

```
tianyi-ge@tianyi-ge-VMware20-1:~/Desktop/Yourname_YourUSCID/scripts$ cd ~/Desktop
tianyi-ge@tianyi-ge-VMware20-1:~/Desktop$ mv Yourname_YourUSCID TianyiGe_5804514679
```

- ## A basic Python Script

The task_1.py file was edited using a terminal text editor.The script prompts the user to enter their name and prints a greeting message in the format “Hello, [name]!”.
The script was executed successfully in the terminal.

```
tianyi-ge@tianyi-ge-VMware20-1:~/Desktop/Yourname_YourUSCID$ mkdir data scripts
tianyi-ge@tianyi-ge-VMware20-1:~/Desktop/Yourname_YourUSCID$ cd scripts
tianyi-ge@tianyi-ge-VMware20-1:~/Desktop/Yourname_YourUSCID/scripts$ touch task_
1.py
tianyi-ge@tianyi-ge-VMware20-1:~/Desktop/Yourname_YourUSCID/scripts$ ls
task_1.py
tianyi-ge@tianyi-ge-VMware20-1:~/Desktop/Yourname_YourUSCID/scripts$ nano task_1
.py
tianyi-ge@tianyi-ge-VMware20-1:~/Desktop/Yourname_YourUSCID/scripts$ python3 tas
k_1.py
Please enter your name: Mia
Hello,  Mia!
```

```
tianyi-ge@tianyi-ge-VMware20-1: ~/Desktop/Yourna...
~/Desktop/Yourname_YourUSCID/scripts
 GNU nano 8.4                          task_1.py *
name=input("Please enter your name: ")
print("Hello, ",name + "!")
```

- ## Python Web-scraping Task

(1)Installed required Python libraries (requests and beautifulsoup4) using the Ubuntu package manager due to system-managed Python environment.

```
error: externally-managed-environment

× This environment is externally managed
╰─> To install Python packages system-wide, try apt install
    python3-xyz, where xyz is the package you are trying to
    install.

    If you wish to install a non-Debian-packaged Python package,
    create a virtual environment using python3 -m venv path/to/venv.
    Then use path/to/venv/bin/python and path/to/venv/bin/pip. Make
    sure you have python3-full installed.

    If you wish to install a non-Debian packaged Python application,
    it may be easiest to use pipx install xyz, which will manage a
    virtual environment for you. Make sure you have pipx installed.

    See /usr/share/doc/python3.13/README.venv for more information.

note: If you believe this is a mistake, please contact your Python installation
or OS distribution provider. You can override this, at the risk of breaking your
 Python installation or OS, by passing --break-system-packages.
hint: See PEP 668 for the detailed specification.
```

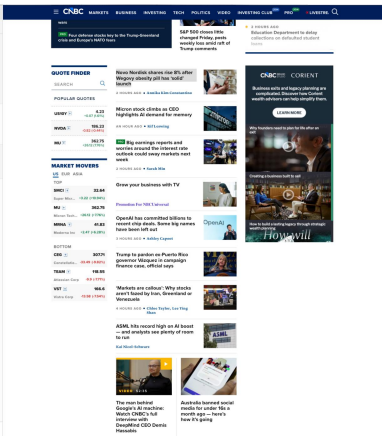
```
tianyi-ge@tianyi-ge-VMware20-1:~/Desktop/Yourname_YourUSCID/scripts
stall python3-requests python3-bs4
[sudo: authenticate] Password:
python3-requests is already the newest version (2.32.3+dfsg-5ubuntu
python3-requests set to manually installed.
The following packages were automatically installed and are no long
  linux-headers-6.17.0-5          linux-tools-6.17.0-5
  linux-headers-6.17.0-5-generic  linux-tools-6.17.0-5-generic
  linux-modules-6.17.0-5-generic
Use 'sudo apt autoremove' to remove them.

Installing:
  python3-bs4

Installing dependencies:
  python3-cssselect  python3-lxml       python3-webencodings
  python3-html5lib   python3-soupsieve

Suggested packages:
  python3-genshi  python-lxml-doc
```

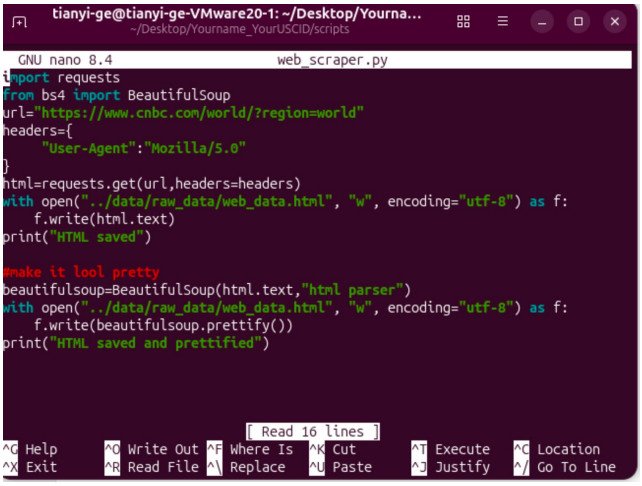
(2)A web scraping script was written to collect HTML data from the CNBC World News webpage.

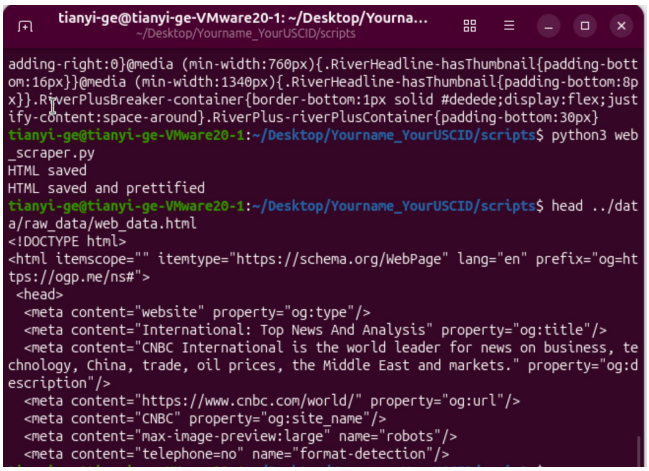


(3) Two directories named raw_data and processed_data were created inside the data folder to store raw and processed data separately

```
tianyi-ge@tianyi-ge-VMware20-1:~/Desktop/Yourname_YourUSCID/scripts$ cd ..
tianyi-ge@tianyi-ge-VMware20-1:~/Desktop/Yourname_YourUSCID$ cd data
tianyi-ge@tianyi-ge-VMware20-1:~/Desktop/Yourname_YourUSCID/data$ mkdir raw_data
 processed_data
tianyi-ge@tianyi-ge-VMware20-1:~/Desktop/Yourname_YourUSCID/data$ ls
processed_data  raw_data
```

(4) A Python script named web_scraper.py was written using Requests and BeautifulSoup to send an HTTP request to the target URL and retrieve the HTML content of the webpage. The collected HTML content was saved into a file named web_data.html inside the data/raw_data directory for further processing.





- Data Filtering Task

A Python script named data_filter.py was written to read the saved HTML file and extract relevant information using BeautifulSoup.Market data and latest news data were parsed from the webpage and stored into two CSV files named market_data.csv and news_data.csv inside the data/processed_data directory.The script was executed successfully and printed messages to the terminal during processing.This screenshot shows the successful execution of the data filtering script.

```
tianyi-ge@tianyi-ge-VMware20-1:~/Desktop/TianyiGe_5804514679/scripts$ python3 data_filter.py
Filtering field
Store the market data
CSV created
```

```
tianyi-ge@tianyi-ge-VMware20-1:~/Desktop/TianyiGe_5804514679/data$ ls processed_data
market_data.csv  news_data.csv
```

```
tianyi-ge@tianyi-ge-VMware20-1:~/Desktop/TianyiGe_5804514679/scripts$ head ../data/processed_data/market_data.csv
symbol,stockposition,pct
tianyi-ge@tianyi-ge-VMware20-1:~/Desktop/TianyiGe_5804514679/scripts$ head ../data/processed_data/news_data.csv
timestamp,title,link
2 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-news-and-we-booked-some-profits.html
3 Hours Ago,Trump threatens to sue JPMorgan Chase for 'debanking' him,https://www.cnbc.com/2026/01/17/trump-jpmorgan-chase-debanking.html
5 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
6 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-strategic-reserve.html
7 Hours Ago,Unshaken: Why Brazilian stocks have looked past the Venezuela attack,https://www.cnbc.com/2026/01/17/unshaken-why-brazilian-stocks-have-looked-past-the-venezuela-attack.html
7 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-without-relying-on-discipline.html
8 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-ask-yourself-this-question.html
8 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-america-says.html
8 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-include-darden-restaurants-and-target.html
```

！！！！ **The CNBC webpage contains certain elements, such as the market banner at the top of the page, that are dynamically rendered using JavaScript. Since this script uses the Requests library, which only retrieves the initial HTML response from the server and does not execute JavaScript, these dynamically generated elements are not present in the saved HTML file. The script therefore collects and saves the raw HTML content returned by the server.**