

DSCI 560 - Lab 1

Name: Angela Kang

Student ID: 895777203

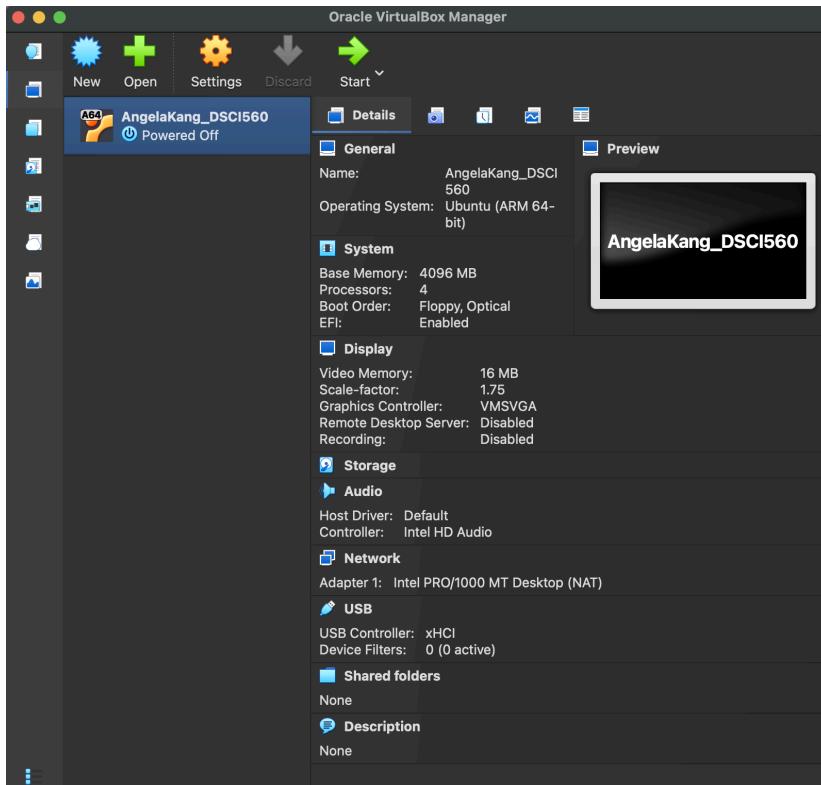
Date: January 16, 2026

Github: angelaykang

Github link: <https://github.com/angelaykang/DSCI560-Labs/tree/main/Lab1>

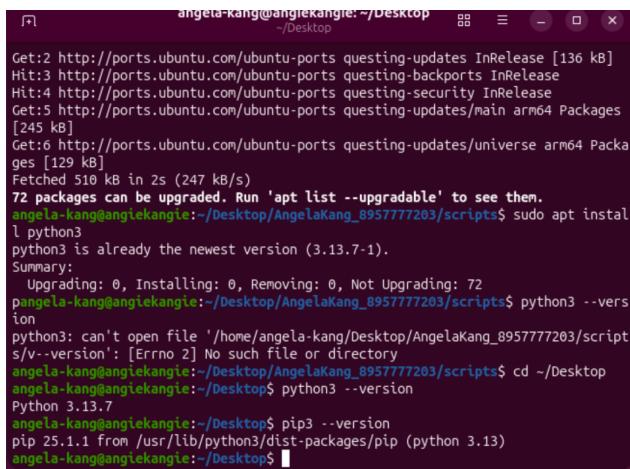
Installation and Setup

Downloaded Virtual Box and Ubuntu ISO image and created a virtual machine with these settings:



Linux and Python

Successfully installed python and pip package manager.

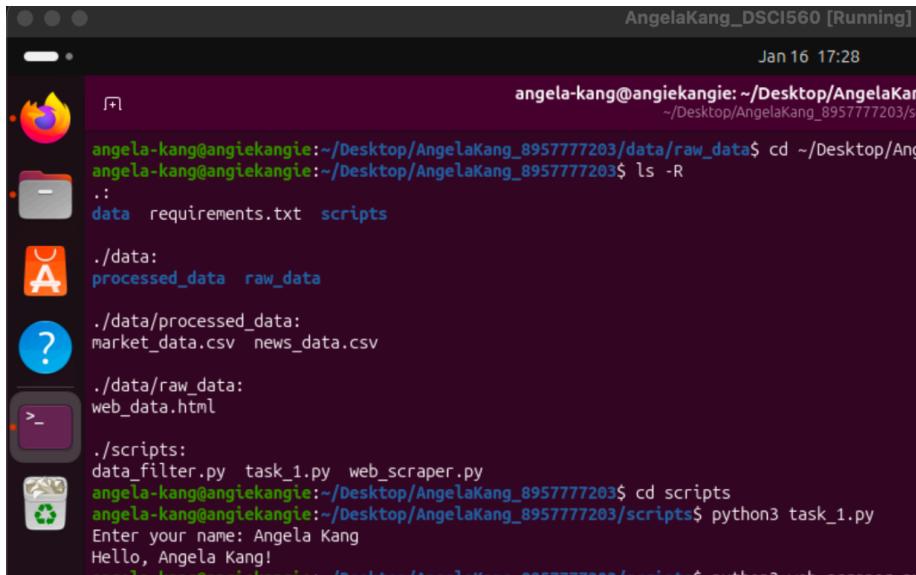


```
angela-kang@angiekangie:~/Desktop
```

```
Get:2 http://ports.ubuntu.com/ubuntu-ports questing-updates InRelease [136 kB]
Hit:3 http://ports.ubuntu.com/ubuntu-ports questing-backports InRelease
Hit:4 http://ports.ubuntu.com/ubuntu-ports questing-security InRelease
Get:5 http://ports.ubuntu.com/ubuntu-ports questing-updates/main arm64 Packages [245 kB]
Get:6 http://ports.ubuntu.com/ubuntu-ports questing-updates/universe arm64 Packages [129 kB]
Fetched 510 kB in 2s (247 kB/s)
72 packages can be upgraded. Run 'apt list --upgradable' to see them.
angela-kang@angiekangie:~/Desktop/AngelaKang_8957777203/scripts$ sudo apt install python3
python3 is already the newest version (3.13.7-1).
Summary:
Upgrading: 0, Installing: 0, Removing: 0, Not Upgrading: 72
pangela-kang@angiekangie:~/Desktop/AngelaKang_8957777203/scripts$ python3 --version
python3: can't open file '/home/angela-kang/Desktop/AngelaKang_8957777203/scripts/v--version': [Errno 2] No such file or directory
angela-kang@angiekangie:~/Desktop/AngelaKang_8957777203/scripts$ cd ~/Desktop
angela-kang@angiekangie:~/Desktop$ python3 --version
Python 3.13.7
angela-kang@angiekangie:~/Desktop$ pip3 --version
pip 25.1.1 from /usr/lib/python3/dist-packages/pip (python 3.13)
angela-kang@angiekangie:~/Desktop$
```

Created the required directory structure on the desktop:

Created a python script (task_1.py) that prompts for user input and displays a greeting message:



```
AngelaKang_DSCI560 [Running]
```

```
Jan 16 17:28
```

```
angela-kang@angiekangie:~/Desktop/AngelaKang_8957777203/data/raw_data$ cd ~/Desktop/AngelaKang_8957777203
```

```
angela-kang@angiekangie:~/Desktop/AngelaKang_8957777203$ ls -R
.:
data requirements.txt scripts

./data:
processed_data raw_data

./data/processed_data:
market_data.csv news_data.csv

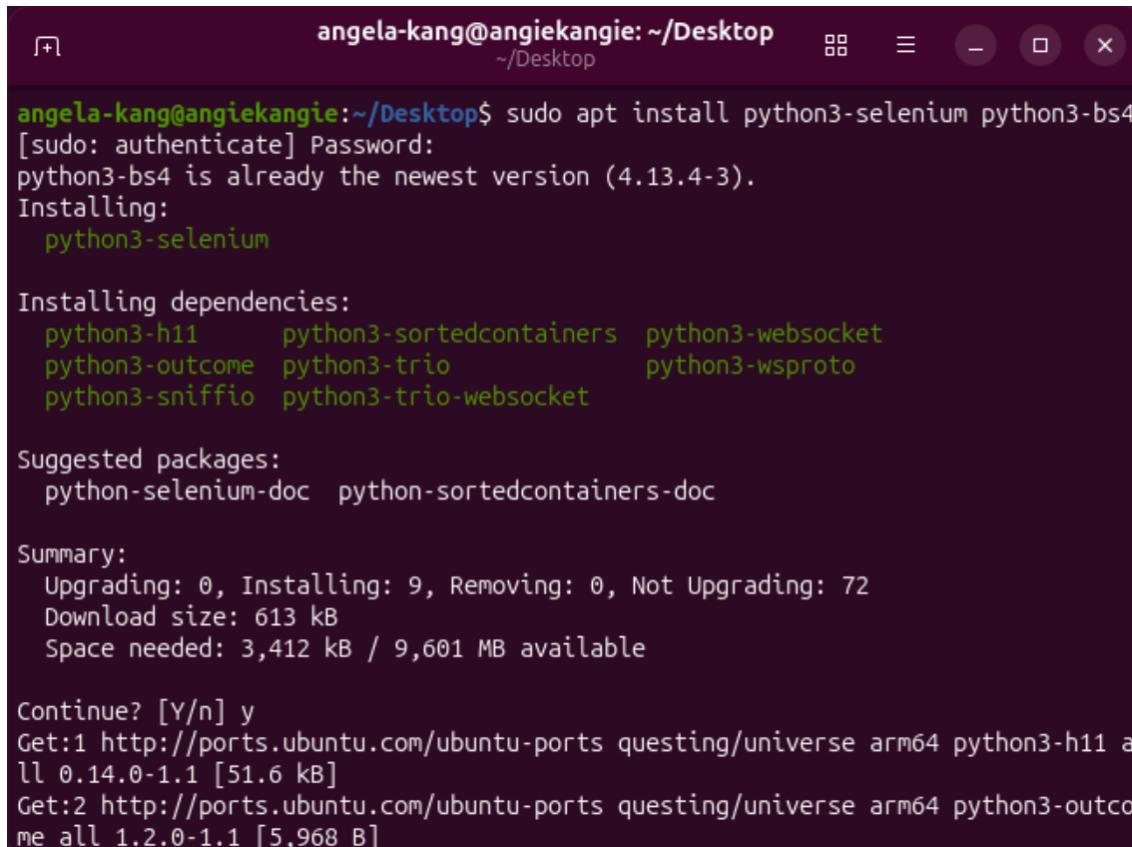
./data/raw_data:
web_data.html

./scripts:
data_filter.py task_1.py web_scraper.py
```

```
angela-kang@angiekangie:~/Desktop/AngelaKang_8957777203$ cd scripts
angela-kang@angiekangie:~/Desktop/AngelaKang_8957777203/scripts$ python3 task_1.py
Enter your name: Angela Kang
Hello, Angela Kang!
```

Web Scraping

Install Selenium and BeautifulSoup4 for web scraping:



```
angela-kang@angiekangie:~/Desktop$ sudo apt install python3-selenium python3-bs4
[sudo: authenticate] Password:
python3-bs4 is already the newest version (4.13.4-3).
Installing:
  python3-selenium

Installing dependencies:
  python3-h11      python3-sortedcontainers  python3-websocket
  python3-outcome   python3-trio            python3-wsproto
  python3-sniffio   python3-trio-websocket

Suggested packages:
  python-selenium-doc  python-sortedcontainers-doc

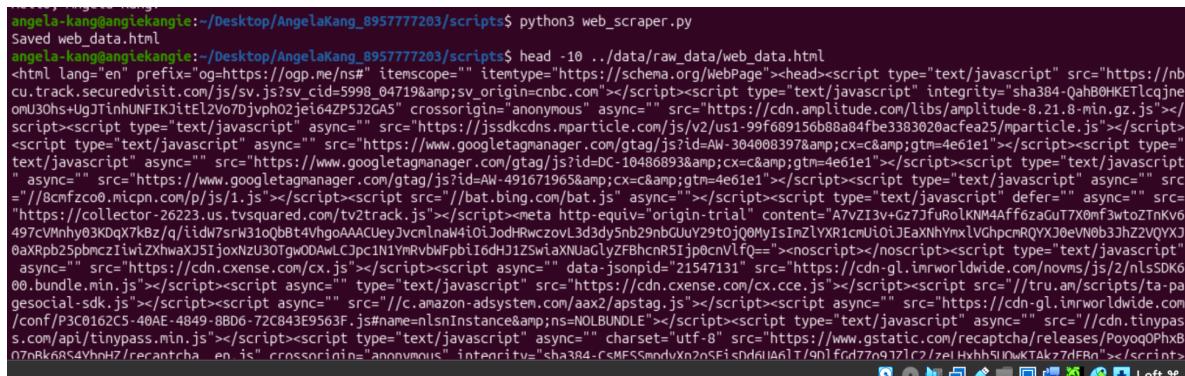
Summary:
Upgrading: 0, Installing: 9, Removing: 0, Not Upgrading: 72
Download size: 613 kB
Space needed: 3,412 kB / 9,601 MB available

Continue? [Y/n] y
Get:1 http://ports.ubuntu.com/ubuntu-ports questing/universe arm64 python3-h11 all 0.14.0-1.1 [51.6 kB]
Get:2 http://ports.ubuntu.com/ubuntu-ports questing/universe arm64 python3-outcome all 1.2.0-1.1 [5,968 B]
```

Created web_scraping.py to scrape CNBC World News page

(<https://www.cnbc.com/world/?region=world>). It uses Selenium with Chrome to load the page dynamically and saves the HTML content to data/raw_data/web_data.html.

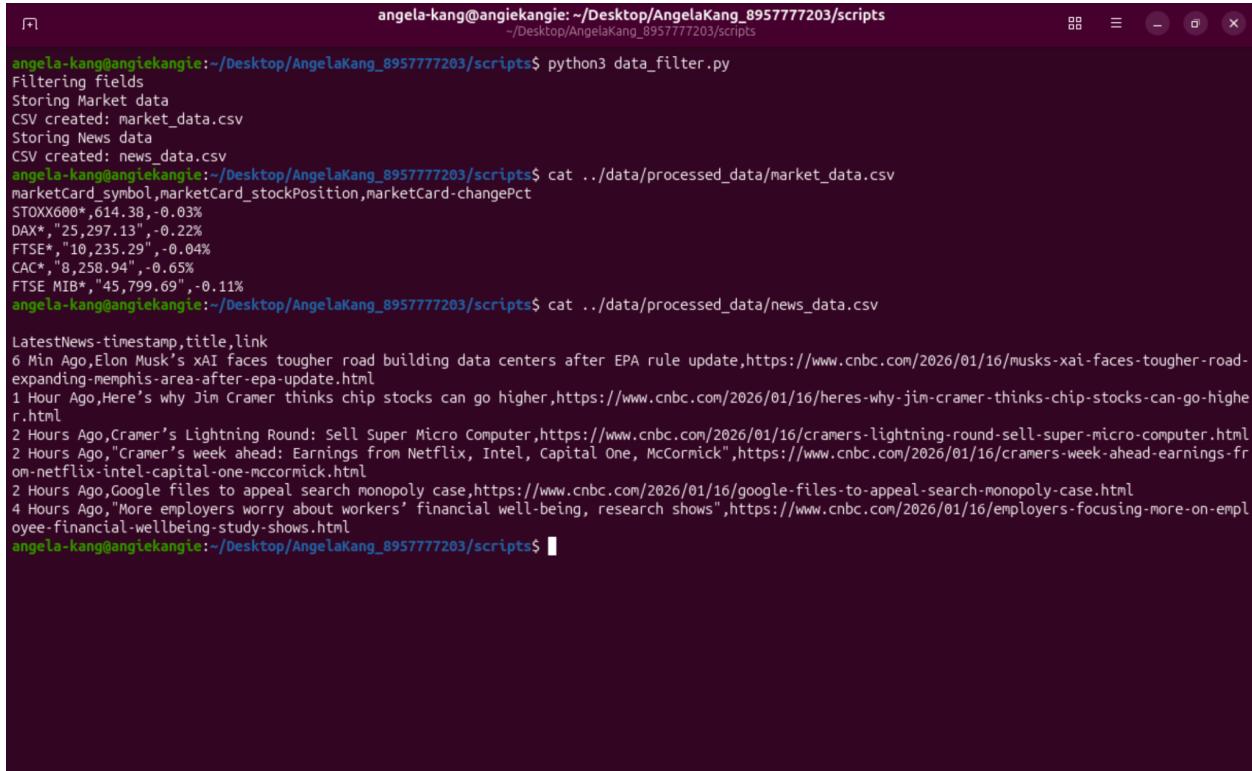
First 10 lines of web_data.html



```
angela-kang@angiekangie:~/Desktop/AngelaKang_8957777203/scripts$ python3 web_scraping.py
Saved web_data.html
angela-kang@angiekangie:~/Desktop/AngelaKang_8957777203/scripts$ head -10 ./data/raw_data/web_data.html
<html lang="en" prefix="og:https://ogp.me/ns#"
  itemscope="" itemtype="https://schema.org/WebPage"><head><script type="text/javascript" src="https://nb
  cu.track.secured.cdn.mparticle.com/ja/sv.js?sv_cid=5998_04719&amp;sv_origin=cnbc.com"></script><script type="text/javascript" integrity="sha384-QahB0HKEtIcqjne
  onU30hs+UgJTinhUNFIKjiteL2vo70jwph02jei64ZP5J2GAS" crossorigin="anonymous" async="" src="https://cdn.amplitude.com/libs/amplitude-8.21.8-min.gz.js"></script>
<script type="text/javascript" async="" src="https://jssdkcdns.mparticle.com/js/v2/us1-99f68915688a04fbe3383020acfca25/mparticle.js"></script>
<script type="text/javascript" async="" src="https://www.googletagmanager.com/gtag/js?id=AW-304008397&amp;cx=&gtm=4e61e1"></script><script type="text/javascript" async="" src="https://www.googletagmanager.com/gtag/js?id=DC-10486893&amp;cx=&gtm=4e61e1"></script><script type="text/javascript" async="" src="https://www.googletagmanager.com/gtag/js?id=AW-491671965&amp;cx=&gtm=4e61e1"></script><script type="text/javascript" async="" src="/8cmfcz0o.micpnm.com/p/ja/1.js"></script><script src="//bat.bing.com/bat.js" async=""></script><script type="text/javascript" defer="" async="" src="https://collector-26223.us.tvsquared.com/tvztrack.js"></script><meta http-equiv="origin-trial" content="A7VZ13v+Gz77fuRoIuKMM4Aff6zaQut7X0mf3wtoZTnKv6
  497cMhy03KQdX7kBz/q/iidW7rsrW31oqbbit4VhgoAAACUeyJvcmLnw4i0iJodhRwcZovl3d3dy5nb29nbGUy29t0j0MyIsInzlyXR1cmUi0iJeaXNhymxlvGhpcmRQYXJ0eVN0b33jhZ2VQYXJ
  0axRpzb25pbmczIwiZxhwaXJSIjoXNu130Tgw0AwLCJpcIN1YMrVbWFpbIIdhJ1ZSwiaXNuAgLyZFbhcnR51jp0cnevLvf0=="><noscript><script type="text/javascript" async="" src="https://cdn.cxense.com/cx.js"></script><script async="" data-jsonpid="21547131" src="https://cdn-gl.lmrworldwide.com/novms/js/2/nlsSDK6
  00.bundle.min.js"></script><script async="" type="text/javascript" src="https://cdn.cxense.com/cx.cce.js"></script><script src="/tru.am/scripts/ta-pa
  gesocial-sdk.js"></script><script async="" src="//c.amazon-adsystem.com/aa2x/apstag.js"></script><script async="" src="https://cdn-gl.lmrworldwide.com
  /conf/P3C0162C5-40AE-4849-8B06-72C843E9563F.js?name=nlsInstance&amp;n=NOLBUNDLE"></script><script type="text/javascript" async="" src="/cdn.tinypass
  .com/api/tinypass.min.js"></script><script type="text/javascript" async="" charset="utf-8" src="https://www.gstatic.com/recaptcha/releases/PoyqOPhx8
  07nRk6RS4Vh0H7/recaptcha_en_is"><script type="text/javascript" async="" crossorigin="anonymous" integrity="sha384-CsMFSSmnDvXn2oF1s1nd6IA61T/qn1frd70q9171C7/7el1Hxhhs1uWKTAKz7dFRN"></script>
```

Data Filtering

Created data_filter.py to parse the HTML file and extract market data (symbol, stock position, and percentage change) into market_data.csv and news data (timestamp, title, and link) into news_data.csv.



```
angela-kang@angiekangie:~/Desktop/AngelaKang_8957777203/scripts
angela-kang@angiekangie:~/Desktop/AngelaKang_8957777203/scripts$ python3 data_filter.py
Filtering fields
Storing Market data
CSV created: market_data.csv
Storing News data
CSV created: news_data.csv
angela-kang@angiekangie:~/Desktop/AngelaKang_8957777203/scripts$ cat ../data/processed_data/market_data.csv
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%
angela-kang@angiekangie:~/Desktop/AngelaKang_8957777203/scripts$ cat ../data/processed_data/news_data.csv
LatestNews-timestamp,title,link
6 Min 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
1 Hour 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
2 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
2 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
2 Hours Ago,Google files to appeal search monopoly case,https://www.cnbc.com/2026/01/16/google-files-to-appeal-search-monopoly-case.html
4 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
angela-kang@angiekangie:~/Desktop/AngelaKang_8957777203/scripts$
```

Conclusion

1. I have successfully installed Virtual Box and Ubuntu virtual machine.
2. Verified python installation and pip package manager.
3. Created directory structure and python greeting script.
4. Implemented web scraper using Selenium to collect CNBC world news data.
5. Parsed HTML data and extracted market and news information into CSV files.