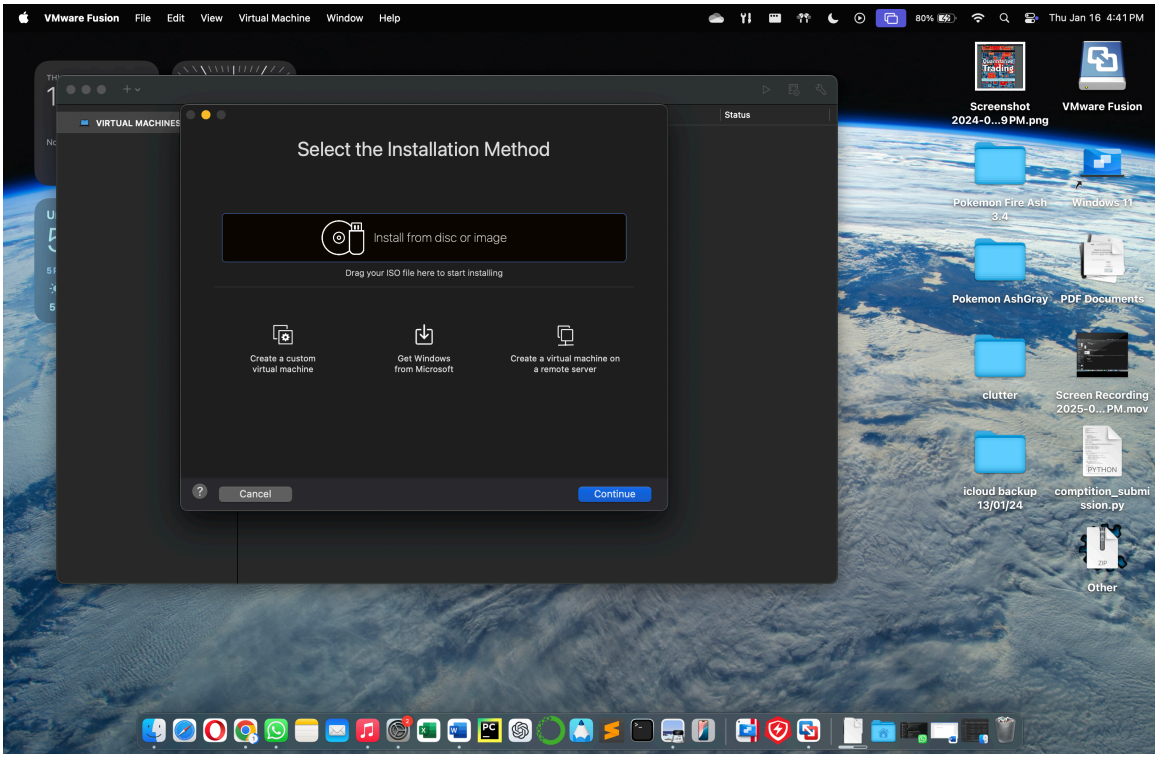
DSCI 560 - Lab 1

Name	Pratham Solanki
USC ID	3242692358
Email	pdsolank@usc.edu

1. **Installation and Setup**

1.1 **VMware installed**



1.2 **Virtual machine (Ubuntu) installed**

- Memory : 4 GB
- Disk space : 20 GB

1.3 **Python Installed on Linux**

- Updated the local packages using "sudo apt update"
- Installed python version 3.12.3

```
0 upgraded, 0 newly installed, 0 to remove and 56 not upgraded.
prathamuser@prathamserver:~/Desktop$ python3 --version
Python 3.12.3
prathamuser@prathamserver:~/Desktop$
```

- Installed pip version 24.0

```
prathamuser@prathamserver:~/Desktop$ pip3 --version
pip 24.0 from /usr/lib/python3/dist-packages/pip (python 3.12)
prathamuser@prathamserver:~/Desktop$
```

2. **Get Familiar with Linux and Python**

2.1 **Playing around with Linux Terminal**

- Created directory "prathamsolanki-3242692358" and subdirectories "data" and "scripts" using mkdir command
- Created empty python file "task1.py" using touch command

```
prathamuser@prathamserver:~/Desktop$ mkdir prathamsolanki_3242692358
prathamuser@prathamserver:~/Desktop$ cd prathamsolanki_3242692358/
prathamuser@prathamserver:~/Desktop/prathamsolanki_3242692358$ mkdir data
prathamuser@prathamserver:~/Desktop/prathamsolanki_3242692358$ mkdir scripts
prathamuser@prathamserver:~/Desktop/prathamsolanki_3242692358$ cd data/scripts
bash: cd: data/scripts: No such file or directory
prathamuser@prathamserver:~/Desktop/prathamsolanki_3242692358$ cd scripts
prathamuser@prathamserver:~/Desktop/prathamsolanki_3242692358/scripts$ touch tas
k1.py
prathamuser@prathamserver:~/Desktop/prathamsolanki_3242692358/scripts$ ls
task1.py
prathamuser@prathamserver:~/Desktop/prathamsolanki_3242692358/scripts$
```

2.2 **A basic Python Script**

- Created the basic python script which greets the user with “Hello,<name>”

```
prathamuser@prathamserver:~/Desktop/prathamsolanki_3242692358/scripts$ nano task
_1.py
prathamuser@prathamserver:~/Desktop/prathamsolanki_3242692358/scripts$ python3 t
ask_1.py Pratham
Hello,Pratham
prathamuser@prathamserver:~/Desktop/prathamsolanki_3242692358/scripts$
```

2.3 Python Web scraping Task

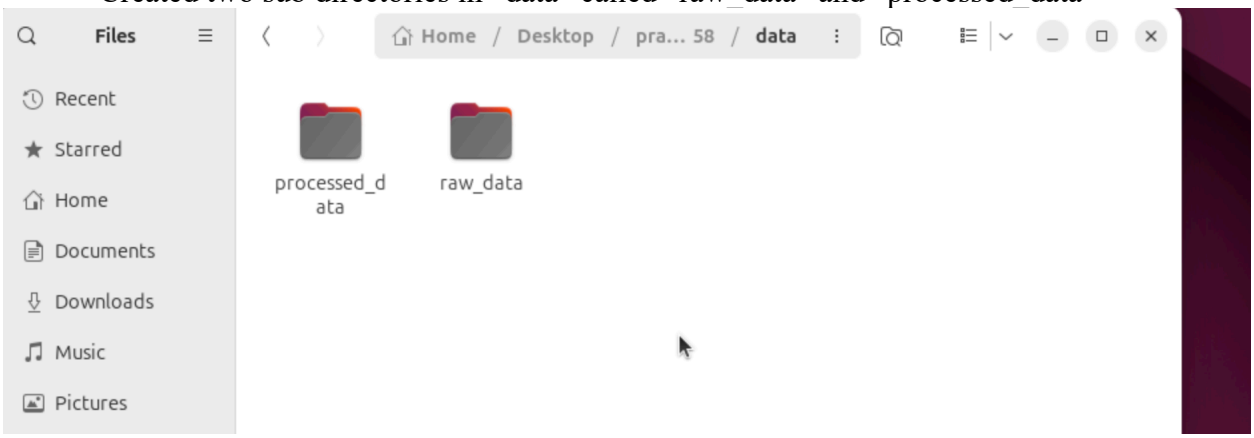
- Created web_scraper.py file

```
(my_virtual_env) prathamuser@prathamserver:~/Desktop/prathamsolanki_3242692358/s
cripts$ touch web_scraper.py
```

- Installed requests, beautifulsoup4

```
cripts$ pip list
Package            Version
------------------ ----------
beautifulsoup4     4.12.3
certifi            2024.12.14
charset-normalizer 3.4.1
idna               3.10
pip                24.0
requests           2.32.3
soupsieve          2.6
urllib3            2.3.0
(my_virtual_env) prathamuser@prathamserver:~/Desktop/prathamsolanki_32426923
```

- Created two sub directories in “data” called “raw_data” and “processed_data”



- Printed first 10 lines of the html file created using head command

- Web_scraper.py Code:

```
from selenium import webdriver
import time
from selenium.webdriver.firefox.service import Service
```

```
from selenium.webdriver.firefox.options import Options
from bs4 import BeautifulSoup


destination_file_path =
'/home/prathamuser/Desktop/prathamsolanki_3242692358/data/raw_data/web_dat
a.html'
gecko_driver_path = '/snap/bin/firefox.geckodriver'


options = Options()
#options.add_argument("--headless")
#options.add_argument("--disable-gpu")
options.add_argument("--window-size=1920,1080")


service = Service(executable_path=gecko_driver_path)
driver = webdriver.Firefox(service=service, options=options)

try:
    driver.get("https://www.cnbc.com/world/?region=world")

    driver.implicitly_wait(20)
    time.sleep(10)
    html_content = driver.page_source
    html_parsed = BeautifulSoup(html_content,'html.parser')

    with open(destination_file_path, "w", encoding="utf-8") as file:
        file.write(html_parsed.prettify())
    print("Successfully written HTML contents of CNBC home page to
web_data.html")
finally:

    driver.quit()
```

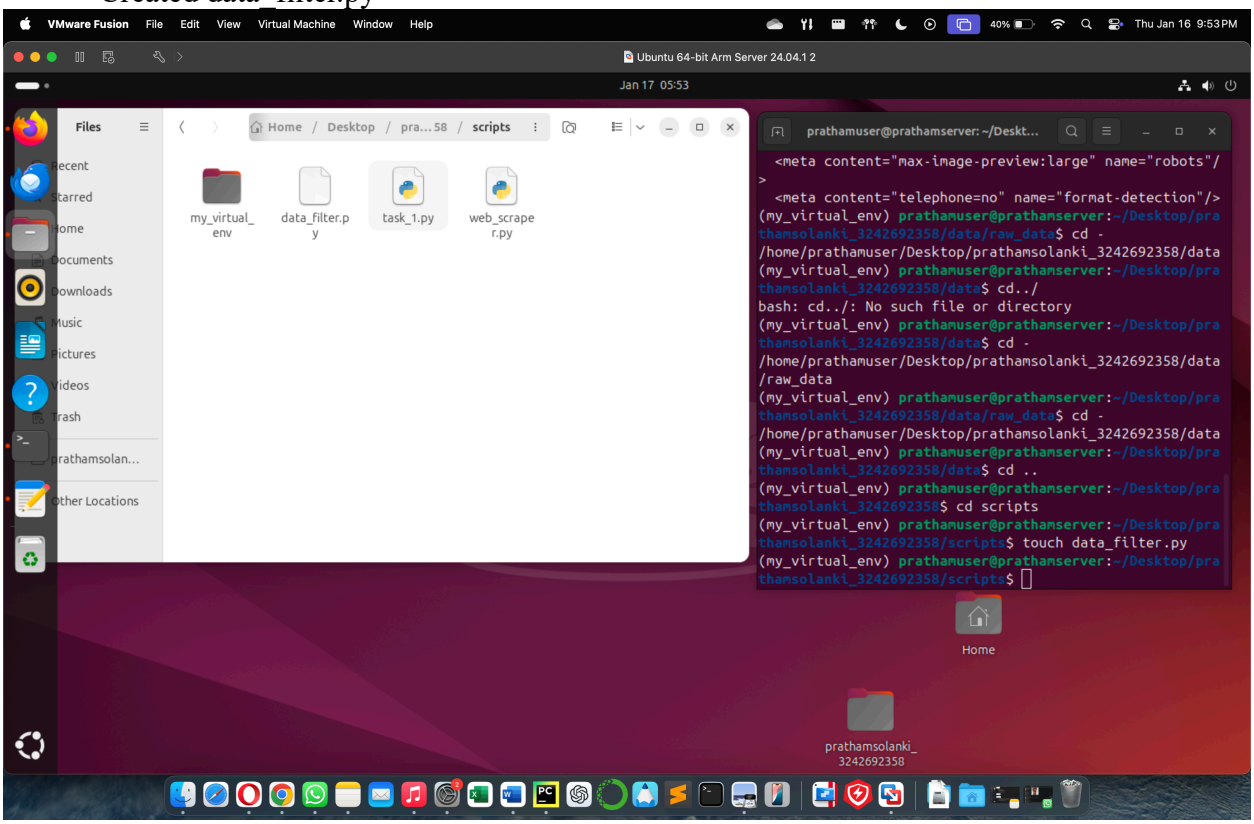
- Output:

```
(my_virtual_env) prathamuser@prathamserver:~/Desktop/prathamsolanki_3242692358/scripts$ python3 web_scraper.py
Successfully written HTML contents of CNBC home page to web_data.html
```

- Screenshot of output file web_data.html:

2.4 **Data Filtering Task**

- Created data_filter.py



- data_filter.py Code:

```
from selenium import webdriver
import time
```

```
from selenium.webdriver.firefox.service import Service
from selenium.webdriver.firefox.options import Options
from bs4 import BeautifulSoup


destination_file_path =
'/home/prathamuser/Desktop/prathamsolanki_3242692358/data/raw_data/web_dat
a.html'
gecko_driver_path = '/snap/bin/firefox.geckodriver'


options = Options()
#options.add_argument("--headless")
#options.add_argument("--disable-gpu")
options.add_argument("--window-size=1920,1080")


service = Service(executable_path=gecko_driver_path)
driver = webdriver.Firefox(service=service, options=options)

try:
    driver.get("https://www.cnbc.com/world/?region=world")

    driver.implicitly_wait(20)
    time.sleep(10)
    html_content = driver.page_source
    html_parsed = BeautifulSoup(html_content,'html.parser')

    with open(destination_file_path, "w", encoding="utf-8") as file:
        file.write(html_parsed.prettify())
    print("Successfully written HTML contents of CNBC home page to
web_data.html")
finally:

    driver.quit()
```

- Output:

```
(my_virtual_env) prathamuser@prathamserver:~/Desktop/prathamsolanki_3242692358/scripts$ python3 data_filter.py
successfully Read HTML code from web_data.html
successfully Fetched Market data
successfully Fetched news data
Data successfully written to market_data.csv and news_data.csv
```

- Screenshot of market_data.csv

market_data.csv — LibreOffice Calc

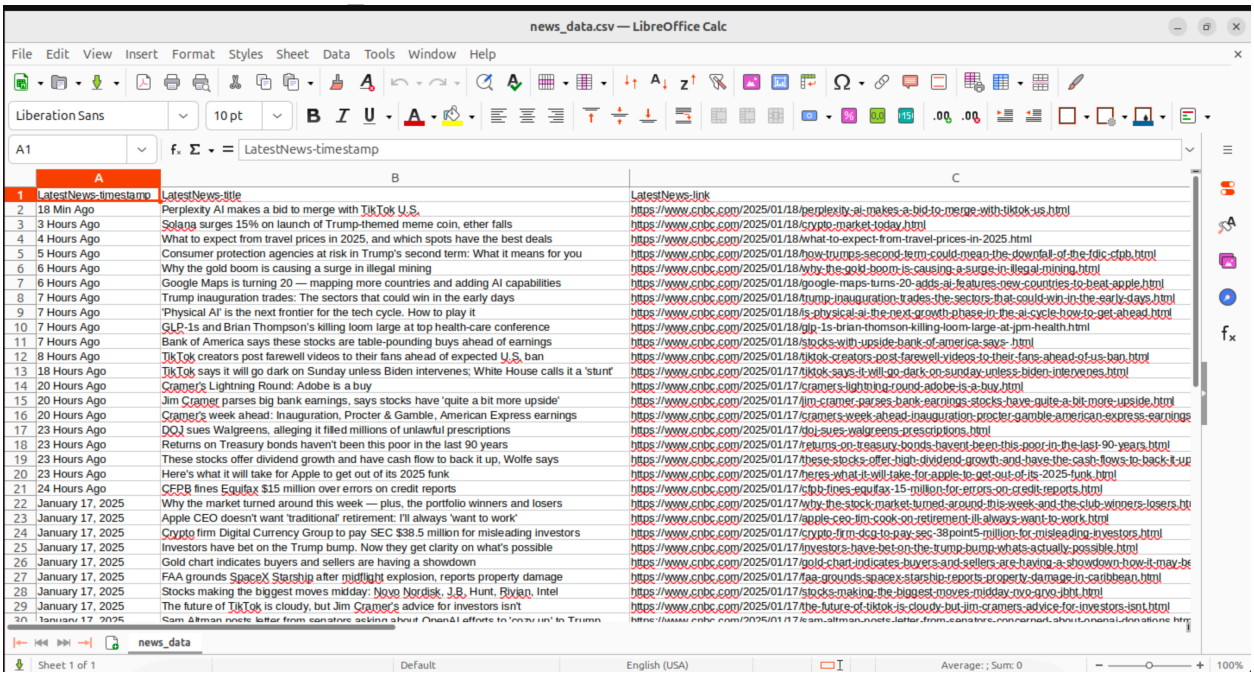
File Edit View Insert Format Styles Sheet Data Tools Window Help

Liberation Sans 10 pt

A1 marketCard_symbol

	A	B	C	D	E	F	G	H	I
1	marketCard_symbol	marketCard_stockPosition	marketCard-changePct						
2	DJIA	43487.83	+0.78%						
3	S&P 500	5996.66	+1.00%						
4	NASDAQ	19630.2	+1.51%						
5	RUSS 2K*	2275.88	+0.40%						
6	VIX	15.97	-3.80%						
7									
8									
9									
10									

- Screenshot of news_data.csv



	A	B	C
1	LatestNews-timestamp	LatestNews-title	LatestNews-link
2	18 Min Ago	Perplexity AI makes a bid to merge with TikTok U.S.	https://www.cnbc.com/2025/01/18/perplexity-ai-makes-a-bid-to-merge-with-tiktok-us.html
3	3 Hours Ago	Solana surges 15% on launch of Trump-themed meme coin, ether falls	https://www.cnbc.com/2025/01/18/crypto-market-today.html
4	4 Hours Ago	What to expect from travel prices in 2025, and which spots have the best deals	https://www.cnbc.com/2025/01/18/what-to-expect-from-travel-prices-in-2025.html
5	5 Hours Ago	Consumer protection agencies at risk in Trump's second term: What it means for you	https://www.cnbc.com/2025/01/18/how-trumps-second-term-could-mean-the-downfall-of-the-fdic-cfpb.html
6	6 Hours Ago	Why the gold boom is causing a surge in illegal mining	https://www.cnbc.com/2025/01/18/why-the-gold-boom-is-causing-a-surge-in-illegal-mining.html
7	6 Hours Ago	Google Maps is turning 20 — mapping more countries and adding AI capabilities	https://www.cnbc.com/2025/01/18/google-maps-turns-20-adds-ai-features-new-countries-to-beat-apple.html
8	7 Hours Ago	Trump inauguration trades: The sectors that could win in the early days	https://www.cnbc.com/2025/01/18/trump-inauguration-trades-the-sectors-that-could-win-in-the-early-days.html
9	7 Hours Ago	'Physical AI' is the next frontier for the tech cycle. How to play it	https://www.cnbc.com/2025/01/18/is-physical-ai-the-next-growth-phase-in-the-ai-cycle-how-to-get-ahead.html
10	7 Hours Ago	GLP-1s and Brian Thompson's killing loom large at top health-care conference	https://www.cnbc.com/2025/01/18/glp-1s-brian-thomson-killing-loom-large-at-jpm-health.html
11	7 Hours Ago	Bank of America says these stocks are table-pounding buys ahead of earnings	https://www.cnbc.com/2025/01/18/stocks-with-upside-bank-of-america-says-.html
12	8 Hours Ago	TikTok creators post farewell videos to their fans ahead of expected U.S. ban	https://www.cnbc.com/2025/01/18/tiktok-creators-post-farewell-videos-to-their-fans-ahead-of-us-ban.html
13	18 Hours Ago	TikTok says it will go dark on Sunday unless Biden intervenes; White House calls it a 'stunt'	https://www.cnbc.com/2025/01/17/tiktok-says-it-will-go-dark-on-sunday-unless-biden-intervenes.html
14	20 Hours Ago	Cramer's Lightning Round: Adobe is a buy	https://www.cnbc.com/2025/01/17/cramers-lightning-round-adobe-is-a-buy.html
15	20 Hours Ago	Jim Cramer parses big bank earnings, says stocks have 'quite a bit more upside'	https://www.cnbc.com/2025/01/17/jim-cramer-parses-bank-earnings-stocks-have-quite-a-bit-more-upside.html
16	20 Hours Ago	Cramer's week ahead: Inauguration, Procter & Gamble, American Express earnings	https://www.cnbc.com/2025/01/17/cramers-week-ahead-inauguration-procter-gamble-american-express-earnings
17	23 Hours Ago	DOJ sues Walgreens, alleging it filled millions of unlawful prescriptions	https://www.cnbc.com/2025/01/17/doj-sues-walgreens-prescriptions.html
18	23 Hours Ago	Returns on Treasury bonds haven't been this poor in the last 90 years	https://www.cnbc.com/2025/01/17/returns-on-treasury-bonds-havent-been-this-poor-in-the-last-90-years.html
19	23 Hours Ago	These stocks offer dividend growth and have cash flow to back it up, Wolfe says	https://www.cnbc.com/2025/01/17/these-stocks-offer-high-dividend-growth-and-have-the-cash-flows-to-back-it-up
20	23 Hours Ago	Here's what it will take for Apple to get out of its 2025 funk	https://www.cnbc.com/2025/01/17/heres-what-it-will-take-for-apple-to-get-out-of-its-2025-funk.html
21	24 Hours Ago	CFPB fines Equifax $15 million over errors on credit reports	https://www.cnbc.com/2025/01/17/cfpb-fines-equifax-15-million-for-errors-on-credit-reports.html
22	January 17, 2025	Why the market turned around this week — plus, the portfolio winners and losers	https://www.cnbc.com/2025/01/17/why-the-stock-market-turned-around-this-week-and-the-club-winners-losers.ht
23	January 17, 2025	Apple CEO doesn't want 'traditional' retirement: I'll always 'want to work'	https://www.cnbc.com/2025/01/17/apple-ceo-tim-cook-on-retirement-ill-always-want-to-work.html
24	January 17, 2025	Crypto firm Digital Currency Group to pay SEC $38.5 million for misleading investors	https://www.cnbc.com/2025/01/17/crypto-firm-dcg-to-pay-sec-38point5-million-for-misleading-investors.html
25	January 17, 2025	Investors have bet on the Trump bump. Now they get clarity on what's possible	https://www.cnbc.com/2025/01/17/investors-have-bet-on-the-trump-bump-whats-actually-possible.html
26	January 17, 2025	Gold chart indicates buyers and sellers are having a showdown	https://www.cnbc.com/2025/01/17/gold-chart-indicates-buyers-and-sellers-are-having-a-showdown-how-it-may-be
27	January 17, 2025	FAA grounds SpaceX Starship after midflight explosion, reports property damage	https://www.cnbc.com/2025/01/17/faa-grounds-spacex-starship-reports-property-damage-in-caribbean.html
28	January 17, 2025	Stocks making the biggest moves midday: Novo Nordisk, J.B. Hunt, Rivian, Intel	https://www.cnbc.com/2025/01/17/stocks-making-the-biggest-moves-midday-nvo-grvo-jbht.html
29	January 17, 2025	The future of TikTok is cloudy, but Jim Cramer's advice for investors isn't	https://www.cnbc.com/2025/01/17/the-future-of-tiktok-is-cloudy-but-jim-cramers-advice-for-investors-isnt.html
30	January 17, 2025	Sam Altman posts letter from senators asking about OpenAI efforts to 'cozy up' to Trump	https://www.cnbc.com/2025/01/17/sam-altman-posts-letter-from-senators-concerned-about-openai-donations.html