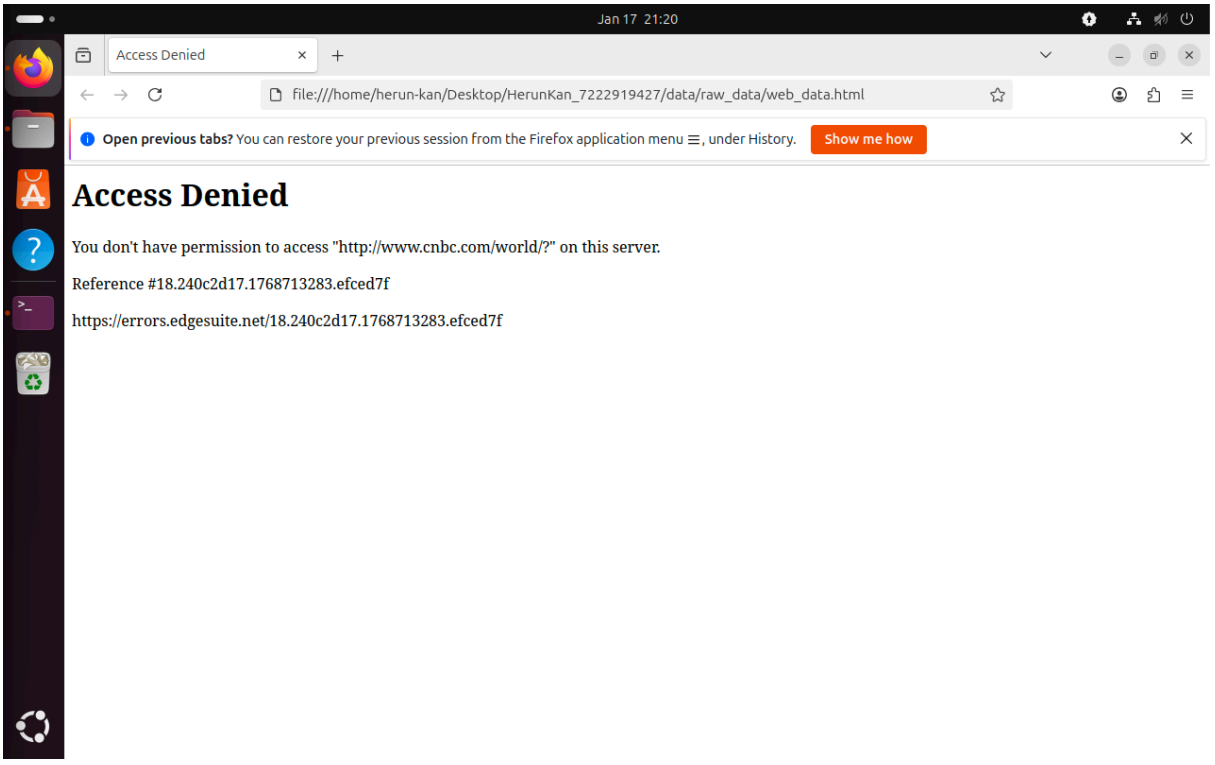
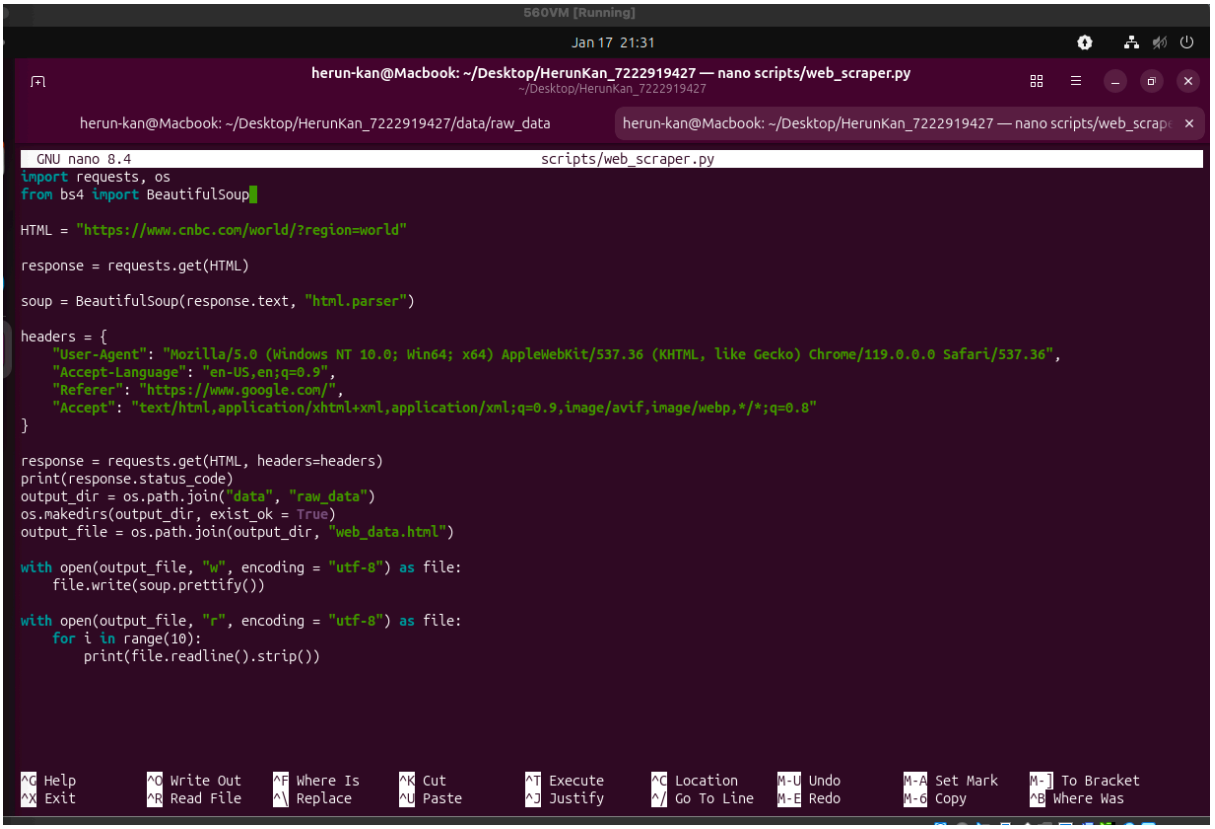
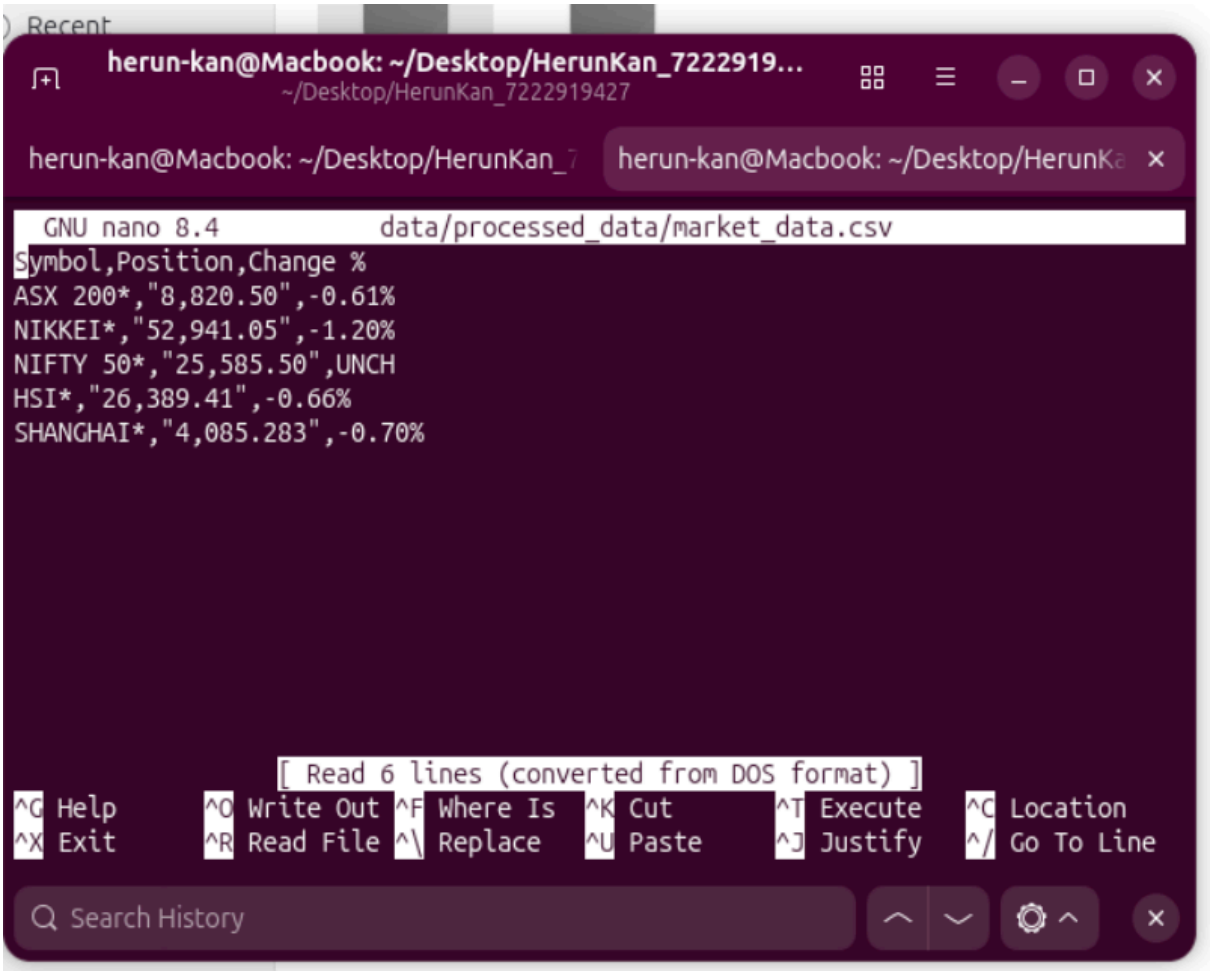
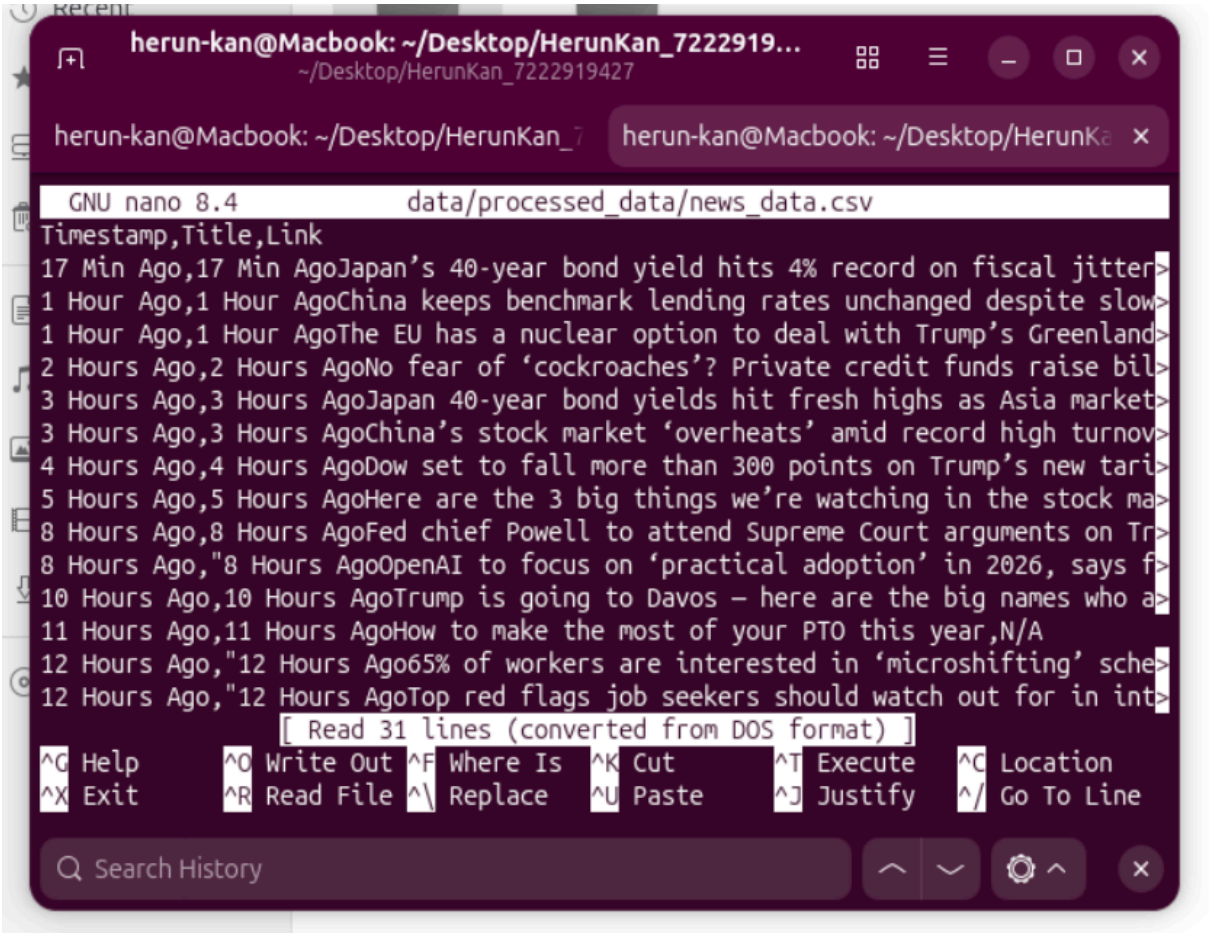
Lab1 Herun Kan 7222919427

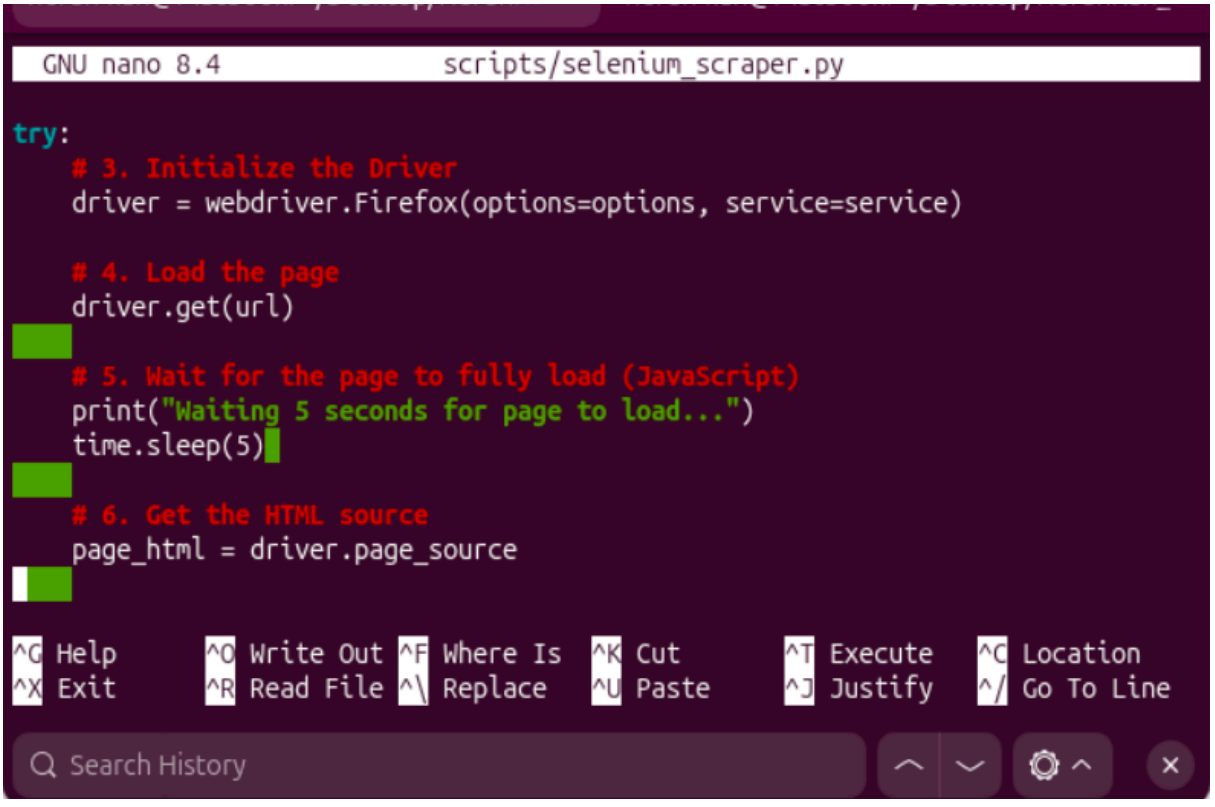
I ran into some problems with Part 2.3 and 2.4, first try I wrote the webscraper without adding the headers and got access denied. I then did some research and learned that adding a headers dictionary can bypass this, however it still did not work. And Tanya suggested to use selenium and it worked. managed to get the html and filtered out latest news and market data




560VM [Running]
Jan 17 21:31
herun-kan@Macbook: ~/Desktop/HerunKan_7222919427 — nano scripts/web_scraper.py
~/Desktop/HerunKan_7222919427
herun-kan@Macbook: ~/Desktop/HerunKan_7222919427/data/raw_data
herun-kan@Macbook: ~/Desktop/HerunKan_7222919427 — nano scripts/web_scrape
GNU nano 8.4 scripts/web_scraper.py
import requests, os
from bs4 import BeautifulSoup
HTML = "https://www.cnbc.com/world/?region=world"
response = requests.get(HTML)
soup = BeautifulSoup(response.text, "html.parser")
headers = {
"User-Agent": "Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/119.0.0.0 Safari/537.36",
"Accept-Language": "en-US,en;q=0.9",
"Referer": "https://www.google.com/",
"Accept": "text/html,application/xhtml+xml,application/xml;q=0.9,image/avif,image/webp,*/*;q=0.8"
}
response = requests.get(HTML, headers=headers)
print(response.status_code)
output_dir = os.path.join("data", "raw_data")
os.makedirs(output_dir, exist_ok = True)
output_file = os.path.join(output_dir, "web_data.html")
with open(output_file, "w", encoding = "utf-8") as file:
file.write(soup.prettify())
with open(output_file, "r", encoding = "utf-8") as file:
for i in range(10):
print(file.readline().strip())
^G Help ^O Write Out ^F Where Is ^K Cut ^T Execute ^C Location M-U Undo M-A Set Mark M-] To Bracket
^X Exit ^R Read File ^\ Replace ^U Paste ^J Justify ^/ Go To Line M-E Redo M-6 Copy ^B Where Was


herun-kan@Macbook: ~/Desktop/HerunKan_7222919...
~/Desktop/HerunKan_7222919427
herun-kan@Macbook: ~/Desktop/HerunKan_7
herun-kan@Macbook: ~/Desktop/HerunKa
GNU nano 8.4 data/processed_data/market_data.csv
Symbol,Position,Change %
ASX 200*,"8,820.50",-0.61%
NIKKEI*,"52,941.05",-1.20%
NIFTY 50*,"25,585.50",UNCH
HSI*,"26,389.41",-0.66%
SHANGHAI*,"4,085.283",-0.70%
[Read 6 lines (converted from DOS format)]
^G Help ^O Write Out ^F Where Is ^K Cut ^T Execute ^C Location
^X Exit ^R Read File ^\ Replace ^U Paste ^J Justify ^/ Go To Line
Search History


herun-kan@Macbook: ~/Desktop/HerunKan_7222919...
~/Desktop/HerunKan_7222919427
herun-kan@Macbook: ~/Desktop/HerunKan_7
herun-kan@Macbook: ~/Desktop/HerunKa
GNU nano 8.4 data/processed_data/news_data.csv
Timestamp,Title,Link
17 Min Ago,17 Min AgoJapan's 40-year bond yield hits 4% record on fiscal jitter>
1 Hour Ago,1 Hour AgoChina keeps benchmark lending rates unchanged despite slow>
1 Hour Ago,1 Hour AgoThe EU has a nuclear option to deal with Trump's Greenland>
2 Hours Ago,2 Hours AgoNo fear of 'cockroaches'? Private credit funds raise bil>
3 Hours Ago,3 Hours AgoJapan 40-year bond yields hit fresh highs as Asia market>
3 Hours Ago,3 Hours AgoChina's stock market 'overheats' amid record high turnov>
4 Hours Ago,4 Hours AgoDow set to fall more than 300 points on Trump's new tari>
5 Hours Ago,5 Hours AgoHere are the 3 big things we're watching in the stock ma>
8 Hours Ago,8 Hours AgoFed chief Powell to attend Supreme Court arguments on Tr>
8 Hours Ago,"8 Hours AgoOpenAI to focus on 'practical adoption' in 2026, says f>
10 Hours Ago,10 Hours AgoTrump is going to Davos — here are the big names who a>
11 Hours Ago,11 Hours AgoHow to make the most of your PTO this year,N/A
12 Hours Ago,"12 Hours Ago65% of workers are interested in 'microshifting' sche>
12 Hours Ago,"12 Hours AgoTop red flags job seekers should watch out for in int>
[Read 31 lines (converted from DOS format)]
^G Help ^O Write Out ^F Where Is ^K Cut ^T Execute ^C Location
^X Exit ^R Read File ^\ Replace ^U Paste ^J Justify ^/ Go To Line
Search History


GNU nano 8.4 scripts/selenium_scraper.py
try:
# 3. Initialize the Driver
driver = webdriver.Firefox(options=options, service=service)
# 4. Load the page
driver.get(url)
# 5. Wait for the page to fully load (JavaScript)
print("Waiting 5 seconds for page to load...")
time.sleep(5)
# 6. Get the HTML source
page_html = driver.page_source
^G Help ^O Write Out ^F Where Is ^K Cut ^T Execute ^C Location
^X Exit ^R Read File ^\ Replace ^U Paste ^J Justify ^/ Go To Line
Search History