In [ ]:
```
# WEB SCRAPING ASSIGNMENT 2
```

In [4]:
```
# let's first install the selenium library
! pip install selenium
```

```
Requirement already satisfied: selenium in c:\users\gaura\anaconda3\lib\site-packages
(4.17.2)
Requirement already satisfied: urllib3[socks]<3,>=1.26 in c:\users\gaura\anaconda3\li
b\site-packages (from selenium) (1.26.16)
Requirement already satisfied: trio~=0.17 in c:\users\gaura\anaconda3\lib\site-packag
es (from selenium) (0.24.0)
Requirement already satisfied: trio-websocket~=0.9 in c:\users\gaura\anaconda3\lib\si
te-packages (from selenium) (0.11.1)
Requirement already satisfied: certifi>=2021.10.8 in c:\users\gaura\anaconda3\lib\sit
e-packages (from selenium) (2023.7.22)
Requirement already satisfied: typing_extensions>=4.9.0 in c:\users\gaura\anaconda3\l
ib\site-packages (from selenium) (4.9.0)
Requirement already satisfied: attrs>=20.1.0 in c:\users\gaura\anaconda3\lib\site-pac
kages (from trio~=0.17->selenium) (22.1.0)
Requirement already satisfied: sortedcontainers in c:\users\gaura\anaconda3\lib\site-
packages (from trio~=0.17->selenium) (2.4.0)
Requirement already satisfied: idna in c:\users\gaura\anaconda3\lib\site-packages (fr
om trio~=0.17->selenium) (3.4)
Requirement already satisfied: outcome in c:\users\gaura\anaconda3\lib\site-packages
(from trio~=0.17->selenium) (1.3.0.post0)
Requirement already satisfied: sniffio>=1.3.0 in c:\users\gaura\anaconda3\lib\site-pa
ckages (from trio~=0.17->selenium) (1.3.0)
Requirement already satisfied: cffi>=1.14 in c:\users\gaura\anaconda3\lib\site-packag
es (from trio~=0.17->selenium) (1.15.1)
Requirement already satisfied: wsproto>=0.14 in c:\users\gaura\anaconda3\lib\site-pac
kages (from trio-websocket~=0.9->selenium) (1.2.0)
Requirement already satisfied: PySocks!=1.5.7,<2.0,>=1.5.6 in c:\users\gaura\anaconda
3\lib\site-packages (from urllib3[socks]<3,>=1.26->selenium) (1.7.1)
Requirement already satisfied: pycparser in c:\users\gaura\anaconda3\lib\site-package
s (from cffi>=1.14->trio~=0.17->selenium) (2.21)
Requirement already satisfied: h11<1,>=0.9.0 in c:\users\gaura\anaconda3\lib\site-pac
kages (from wsproto>=0.14->trio-websocket~=0.9->selenium) (0.14.0)
```

In [7]:
```
# Let's now import all the required libraries.
import selenium
import pandas as pd
from selenium import webdriver
```

In [ ]:
```
Q1: Write a python program to scrape data for “Data Analyst” Job position in “Bangalor
have to scrape the job-title, job-location, company_name, experience_required. You hav
jobs data.
This task will be done in following steps:
1. First get the webpage https://www.shine.com/
2. Enter “Data Analyst” in “Job title, Skills” field and enter “Bangalore” in “enter t
3. Then click the searchbutton.
4. Then scrape the data for the first 10 jobs results you get.
5. Finally create a dataframe of the scraped data.
```

In [8]:
```
url = "https://www.shine.com"
driver.get(url)
```

```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
Cell In[8], line 2
      1 url = "https://www.shine.com"
----> 2 driver.get(url)

NameError: name 'driver' is not defined
```

In [6]:
```
search_job = driver.find_element_by_xpath("//input[@class='sugInp']")
search_job
```

```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
Cell In[6], line 1
----> 1 search_job = driver.find_element_by_xpath("//input[@class='sugInp']")
      2 search_job

NameError: name 'driver' is not defined
```

In [7]:
```
search_job.send_keys('Data Analyst')
```

```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
Cell In[7], line 1
----> 1 search_job.send_keys('Data Analyst')

NameError: name 'search_job' is not defined
```

In [8]:
```
search_loc=driver.find_element_by_id('qsb-location-sugg')
search_loc.send_keys("Bangalore")
```

```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
Cell In[8], line 1
----> 1 search_loc=driver.find_element_by_id('qsb-location-sugg')
      2 search_loc.send_keys("Bangalore")

NameError: name 'driver' is not defined
```

In [ ]:
```
search_btn= driver .find_element_by_xpath("//button[@class='btn']")
search_btn
```

In [ ]:
```
search_btn=driver.find_element_by_xpath("//button[@class='btn']")
search_btn.click()
```

In [ ]:
```
title_tags=driver.find_elements_by_xpath("//a[@class='title fw500 ellipsis']")
title_tags
```

In [ ]:
```
# extract the text of the job title  from the tags
job_titles=[]
for i in title_tags:
    if i.text is None:
        job_titles.append('Not')
    else:
        job_titles.append(i.text)
job_titles[:10]
```

In [ ]:
```
# so lets extract  all the tags having the experience required data
skill_tags=driver.find_elements_by_xpath("//li[@class='fleft grey-text br2 placeHolder
skill_tags
```

In [ ]:
```
# no we will extract the text  from these tags only by one by looping over these tags
skill_list=[]
for i in skill_tags:
    skill_list.append(i.text)
skill_list[:10]
```

In [ ]:
```
locations_list=[]
for i in locations_tags:
    locations_list.append(i.text)
locations_list[:10]
```

In [ ]:
```
locations_tags=driver.find_elements_by_xpath("//li[@class='fleft grey-text br2 placeHc
locations_tags
```

In [ ]:
```
#So lets check th length of ech element.
print(len(job_titles[:10])),print(len(print(len(skills_list[:10])),print(len(locations
```

In [10]:
```
import selenium
import pandas as pd
from selenium import webdriver

jobs=pd.DataFrame({})
jobs['title']=job_titles[:10]
jobs['skill_required']=skill_list[:10]
jobs['location']=locations_list[:10]
```

```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
Cell In[10], line 6
      3 from selenium import webdriver
      5 jobs=pd.DataFrame({})
----> 6 jobs['title']=job_titles[:10]
      7 jobs['skill_required']=skill_list[:10]
      8 jobs['location']=locations_list[:10]

NameError: name 'job_titles' is not defined
```

In [ ]:
```
jobs
```

In [ ]:
```
Q2:Write a python program to scrape data for “Data Scientist” Job position in“Bangalor
have to scrape the job-title, job-location, company_name. You have to scrape first 10
This task will be done in following steps:
1. First get the webpage https://www.shine.com/
2. Enter “Data Scientist” in “Job title, Skills” field and enter “Bangalore” in “enter
3. Then click the search button.
4. Then scrape the data for the first 10 jobs results you get.
5. Finally create a dataframe of the scraped data.
```

In [ ]:
```
# let's first connect to the web driver
driver = webdriver.Chrome(r"C:\Users\Neha\Downloads\chromedriver_win32\chromedriver.ex
```

In [ ]:
```
url = "https://www.shine.com"
driver.get(url)
```

In [ ]:
```
# finding element for job search bar
search_job = driver.find_element_by_xpath("//input[@class='sugInp']")
search_job
```

In [ ]:
```
# write on search bar
search_job.send_keys('Data Scientist')
```

In [ ]:
```
# finding element for job location bar
search_loc=driver.find_element_by_id('qsb-location-sugg')
search_loc.send_keys("Bangalore")
```

In [ ]:
```
search_btn= driver .find_element_by_xpath("//button[@class='btn']")
search_btn
```

In [ ]:
```
# clicking  using xpath function
search_btn=driver.find_element_by_class_name('btn')
search_btn.click()
```

In [ ]:
```
#so let's extract all the tags having the job titles
title_tag=driver.find_elements_by_xpath("//a[@class='title fw500 ellipsis']")
title_tag
```

In [ ]:
```
# extract the text of the job title  from the tags
job1_titles=[]
for i in title_tag:
    if i.text is None:
        job1_titles.append('Not')
    else:
        job1_titles.append(i.text)
job1_titles[:10]
```

In [ ]:
```
# lets extract all the tags having company names
company_tag=driver.find_elements_by_xpath("//a[@class='subTitle ellipsis fleft']")
company_tag
```

In [ ]:
```
# Now we will extract the text from the tags by looping over these tags
companies1_names=[]

for i in company_tag:
    companies1_names.append(i.text)
companies1_names[:10]
```

In [ ]:
```
# lets extract all the tags having locations
locations_tag=driver.find_elements_by_xpath("//li[@class='fleft grey-text br2 placeHol
locations_tag
```

In [ ]:
```
locations1_list=[]
for i in locations_tag:
    locations1_list.append(i.text)
locations1_list[:10]
```

In [1]:
```
print(len(job1_titles[:10])),print(len(companies1_names[:10])),print(len(locations1_li
```

```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
Cell In[1], line 1
----> 1 print(len(job1_titles[:10])),print(len(companies1_names[:10])),print(len(loca
tions1_list[:10]))

NameError: name 'job1_titles' is not defined
```

In [ ]:
```
driver=webdriver.Chrome(r"C:\Users\Neha\Downloads\chromedriver_win32\chromedriver.exe"
driver.get('https://www.shine.com/data-scientist-jobs-in-banglore-bagaluru')
```

In [ ]:
```
urls=[]
```

In [ ]:
```
job_description=[]
```

In [ ]:
```
for i in driver.find_elements_by_xpath("//a[@class='title fw500 ellipsis']"):
    urls.append(i.get_attribute("href"))
```

In [ ]:
```
for url in urls[:10]:

    try:
        driver.get(url)
        description=driver.find_element_by_xpath("//section[@class='job-desc']").text
        job_description.append(description)

    except NoSuchElementException:
        job_description.append("Not Available")
```

In [ ]:
```
job_description
```

In [ ]:
```
print(len(job_description))
```

In [ ]:
```
#Creating a dataframe for the Data Analyst Jobs
```

In [ ]:
```
job1=pd.DataFrame({})
job1['title']=job1_titles[:10]
job1['company_name']=companies1_names[:10]
job1['location']=locations1_list[:10]
job1['job_desc']=job_description
```

In [ ]:
```
job1
```

In [ ]:
```
#So here we can see that we have created the dataset for Data scientist jobs named job
```

In [ ]:
```
driver.close()
```

In [ ]:
```
Q3: In this question you have to scrape data using the filters available on the webpag
 You have to use the location and salary filter.
You have to scrape data for “Data Scientist” designation for first 10 job results.
You have to scrape the job-title, job-location, company name, experience required.
The location filter to be used is “Delhi/NCR”. The salary filter to be used is “3-6” l
```

In [ ]:
```
# let's first connect to the web driver
driver = webdriver.Chrome(r"C:\Users\Neha\Downloads\chromedriver_win32\chromedriver.ex
```

In [ ]:
```
url = "https://www.shine.com"
driver.get(url)
```

In [ ]:
```
# finding element for job search bar
search_job = driver.find_element_by_xpath("//input[@class='sugInp']")
search_job
```

In [ ]:
```
# write on search bar
search_job.send_keys('Data Scientist')
```

In [ ]:
```
search_btn= driver .find_element_by_xpath("//button[@class='btn']")
search_btn
```

In [ ]:
```
# clicking  using xpath function
search_btn=driver.find_element_by_class_name('btn')
search_btn.click()
```

In [ ]:
```
#so let's extract all the tags having the job titles
title_t1=driver.find_elements_by_xpath("//a[@class='title fw500 ellipsis']")
title_t1
```

In [ ]:
```
# extract the text of the job title  from the tags
job_titles=[]
for i in title_t1:
    if i.text is None:
        job_titles.append('Not')
    else:
        job_titles.append(i.text)
job_titles[:10]
```

In [ ]:
```
# lets extract all the tags having company names
company_t1=driver.find_elements_by_xpath("//a[@class='subTitle ellipsis fleft']")
company_t1
```

In [ ]:
```
# Now we will extract the text from the tags by looping over these tags
companies_names=[]

for i in company_t1:
    companies_names.append(i.text)
companies_names[:10]
```

In [ ]:
```
# so lets extract  all the tags having the experience required data
experience_t1=driver.find_elements_by_xpath("//li[@class='fleft grey-text br2 placeHol
experience_t1
```

In [ ]:
```
# no we will extract the text  from these tags only by one by looping over these tags
experience_list=[]
for i in experience_t1:
    experience_list.append(i.text)
experience_list[:10]
```

In [ ]:
```
# So lets extract all the tags having locations
locations_t1=driver.find_elements_by_xpath("//li[@class='fleft grey-text br2 placeHold
locations_t1
```

In [ ]:
```
#Now we wil extract the text from these tags only by one by looping over these tags
locations_list=[]
for i in locations_t1:
    locations_list.append(i.text)
locations_list[:10]
```

In [ ]:
```
#So lets check th length of ech element.
print(len(job_titles[:10])),print(len(companies_names[:10])),print(len(experience_list
```

In [ ]:
```
#Creating a DataFarme for the Data Analyst jobs
```

In [ ]:
```
jobs2=pd.DataFrame({})
jobs2['title']=job_titles[:10]
jobs2['company']=companies_names[:10]
jobs2['experience_required']=experience_list[:10]
jobs2['location']=locations_list[:10]
```

In [ ]:
```
jobs2
```

In [ ]:
```
driver.close()
```

In [ ]:
```
Q4: Scrape data of first 100 sunglasses listings on flipkart.com. You have to scrape f
6. Brand
7. ProductDescription
8. Price

To scrape the data you have to go through following steps:
1. Go to Flipkart webpage by url :https://www.flipkart.com/
2. Enter “sunglasses” in the search fieldwhere “search for products, brands and more”
click the search icon
3. After that you will reach to the page having a lot of sunglasses. From this page yo
required data as usual.
4. After scraping data from the first page, go to the “Next” Button at the bottom othe
click on it.
5. Now scrape data from this page as usual
6. Repeat this until you get data for 100sunglasses.
Note: That all of the above steps have to be done by coding only and not manually.
```

In [ ]:
```
# let's first connect to the web driver
driver = webdriver.Chrome(r"C:\Users\Neha\Downloads\chromedriver_win32\chromedriver.ex
```

In [ ]:
```
url="https://www.flipkart.com/"
driver.get(url)
```

In [ ]:
```
# finding element for job search bar
search_g= driver.find_element_by_xpath("//input[@type='text']")
search_g
```

In [ ]:
```
# write on search bar
search_g.send_keys('sunglasses')
```

In [ ]:
```
search_btn=driver.find_element_by_xpath("//button[@class='L0Z3Pu']")
search_btn
```

In [ ]:
```
search_btn=driver.find_element_by_class_name('L0Z3Pu')
search_btn.click()
```

In [ ]:
```
B_name=[]
Price=[]
P_desc=[]
Discount=[]
```

In [ ]:
```
for i in range(3):
    b_name=driver.find_elements_by_xpath("//div[@class='_2WkVRV']")
    p_desc=driver.find_elements_by_xpath("//a[@class='IRpwTa']")
    price =driver.find_elements_by_xpath("//div[@class='_25b18c']")
    discount=driver.find_elements_by_xpath("//div[@class='_3Ay6Sb']")

    for j  in b_name:
        B_name.append(j.text)
    B_name[:100]



    for k in p_desc:
        P_desc.append(k.text)
    P_desc[:100]


    for l in price:
        Price.append(l.text)
    Price[:100]


    for t in discount:
        Discount.append(t.text)
    Discount[:100]
```

In [ ]:
```
B_name[:100]
```

In [ ]:
```
print(len(B_name[:100])),print(len(Price[:100])),print(len(P_desc[:100])),print(len(Di
```

In [ ]:
```
#Creating a dataframe of the above data
```

In [ ]:
```
sun_gl=pd.DataFrame({})
sun_gl['Brand_name']=B_name[:100]
sun_gl['P_price']=Price[:100]
sun_gl['Pr_desc']=P_desc[:100]
sun_gl['P_discount']=Discount[:100]
```

In [ ]:
```
sun_gl
```

In [ ]:
```
driver.close()
```

In [ ]:
```

```

In [ ]:
```
Q5: Scrape data forfirst 100 sneakers you find whenyou visit flipkart.com and search f
search field.
You have to scrape 3 attributes of each sneaker:
1. Brand
2. ProductDescription
3. Price
As shown in the below image, you have to scrape the above attributes.
```

In [ ]:
```
# let's first connect to the web driver
driver = webdriver.Chrome(r"C:\Users\Neha\Downloads\chromedriver_win32\chromedriver.ex
```

In [ ]:
```
url="https://www.flipkart.com/"
driver.get(url)
```

In [ ]:
```
# finding element for job search bar
search_g= driver.find_element_by_xpath("//input[@type='text']")
search_g
```

In [ ]:
```
# write on search bar
search_g.send_keys('sunglasses')
```

In [ ]:
```
search_btn=driver.find_element_by_xpath("//button[@class='L0Z3Pu']")
search_btn
```

In [ ]:
```
search_btn=driver.find_element_by_class_name('L0Z3Pu')
search_btn.click()
```

In [ ]:
```
B_name=[]
Price=[]
P_desc=[]
Discount=[]
```

In [ ]:
```
for i in range(3):
    b_name=driver.find_elements_by_xpath("//div[@class='_2WkVRV']")
    p_desc=driver.find_elements_by_xpath("//a[@class='IRpwTa']")
    price =driver.find_elements_by_xpath("//div[@class='_25b18c']")
    discount=driver.find_elements_by_xpath("//div[@class='_3Ay6Sb']")

    for j  in b_name:
        B_name.append(j.text)
    B_name[:100]



    for k in p_desc:
        P_desc.append(k.text)
    P_desc[:100]


    for l in price:
        Price.append(l.text)
    Price[:100]


    for t in discount:
```

```
        Discount.append(t.text)
    Discount[:100]
```

In [1]:
```
B_name[:100]
```

```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
Cell In[1], line 1
----> 1 B_name[:100]

NameError: name 'B_name' is not defined
```

In [2]:
```
print(len(B_name[:100])),print(len(Price[:100])),print(len(P_desc[:100])),print(len(Di
```

```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
Cell In[2], line 1
----> 1 print(len(B_name[:100])),print(len(Price[:100])),print(len(P_desc[:100])),pri
nt(len(Discount[:100]))

NameError: name 'B_name' is not defined
```

In [ ]:
```
Creating a dataframe of the above data
```

In [ ]:
```
sun_gl=pd.DataFrame({})
sun_gl['Brand_name']=B_name[:100]
sun_gl['P_price']=Price[:100]
sun_gl['Pr_desc']=P_desc[:100]
sun_gl['P_discount']=Discount[:100]
```

In [3]:
```
sun_gl
```

```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
Cell In[3], line 1
----> 1 sun_gl

NameError: name 'sun_gl' is not defined
```

In [ ]:
```
Q6: Go to webpage https://www.amazon.in/ Enter “Laptop” in the search field and then c
set CPU Type filter to “Intel Core i7” as shown in the below image:

After setting the filters scrape first 10 laptops data. You have to scrape 3 attribute
1. Title
2. Ratings
3. Price
```

In [ ]:
```
# let's first connect to the web driver
driver = webdriver.Chrome(r"C:\Users\Neha\Downloads\chromedriver_win32\chromedriver.ex
```

In [ ]:
```
url=" https://www.amazon.in "
driver.get(url)
```

In [ ]:
```
# finding element for job search bar
search_g= driver.find_element_by_xpath("//input[@type='text']")
search_g
```