Web Scrapping Assignment-2(Selenium)

In [2]:
```
!pip install selenium
```
```
Requirement already satisfied: selenium in c:\users\sakshi\anaconda3\lib\site-packages (4.24.0)
Requirement already satisfied: urllib3[socks]<3,>=1.26 in c:\users\sakshi\anaconda3\lib\site-packages (from selenium) (1.26.16)
Requirement already satisfied: trio~=0.17 in c:\users\sakshi\anaconda3\lib\site-packages (from selenium) (0.26.2)
Requirement already satisfied: trio-websocket~=0.9 in c:\users\sakshi\anaconda3\lib\site-packages (from selenium) (0.11.1)
Requirement already satisfied: certifi>=2021.10.8 in c:\users\sakshi\anaconda3\lib\site-packages (from selenium) (2024.6.2)
Requirement already satisfied: typing_extensions~=4.9 in c:\users\sakshi\anaconda3\lib\site-packages (from selenium) (4.12.2)
Requirement already satisfied: websocket-client~=1.8 in c:\users\sakshi\anaconda3\lib\site-packages (from selenium) (1.8.0)
Requirement already satisfied: attrs>=23.2.0 in c:\users\sakshi\anaconda3\lib\site-packages (from trio~=0.17->selenium) (24.2.0)
Requirement already satisfied: sortedcontainers in c:\users\sakshi\anaconda3\lib\site-packages (from trio~=0.17->selenium) (2.4.0)
Requirement already satisfied: idna in c:\users\sakshi\anaconda3\lib\site-packages (from trio~=0.17->selenium) (3.4)
Requirement already satisfied: outcome in c:\users\sakshi\anaconda3\lib\site-packages (from trio~=0.17->selenium) (1.3.0.post0)
Requirement already satisfied: sniffio>=1.3.0 in c:\users\sakshi\anaconda3\lib\site-packages (from trio~=0.17->selenium) (1.3.1)
Requirement already satisfied: cffi>=1.14 in c:\users\sakshi\anaconda3\lib\site-packages (from trio~=0.17->selenium) (1.15.1)
Requirement already satisfied: wsproto>=0.14 in c:\users\sakshi\anaconda3\lib\site-packages (from trio-websocket~=0.9->selenium) (1.2.0)
Requirement already satisfied: PySocks!=1.5.7,<2.0,>=1.5.6 in c:\users\sakshi\anaconda3\lib\site-packages (from urllib3[socks]<3,>=1.26->selenium) (1.7.1)
Requirement already satisfied: pycparser in c:\users\sakshi\anaconda3\lib\site-packages (from cffi>=1.14->trio~=0.17->selenium) (2.21)
Requirement already satisfied: h11<1,>=0.9.0 in c:\users\sakshi\anaconda3\lib\site-packages (from wsproto>=0.14->trio-websocket~=0.9->selenium) (0.14.0)
```

In [9]:
```
import selenium
import pandas as pd
from selenium import webdriver
import warnings
warnings.filterwarnings('ignore')
from selenium.webdriver.common.by import By
import time
```

In [11]:
```
driver = webdriver.Chrome()
```
```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
Cell In[11], line 1
----> 1 driver = webdriver.Chrome()

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

Q1: In this question you have to scrape data using the filters available on the webpage 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" lakhs The task will be done as shown in the below steps:

1. first get the web page https://www.naukri.com/
2. Enter "Data Scientist" in "Skill, Designations, and Companies" field.
3. Then click the search button.
4. Then apply the location filter and salary filter by checking the respective boxes
5. Then scrape the data for the first 10 jobs results you get.
6. Finally create a dataframe of the scraped data.

In [12]:
```
driver.get("https://www.naukri.com/")
```
```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
Cell In[12], line 1
----> 1 driver.get("https://www.naukri.com/")

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

In [13]:
```
search_field_designation=driver.find_elements(By.CLASS_NAME,"suggestor-input ")
search_field_designation.send_keys('Data Scientists')
```
```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
Cell In[13], line 1
----> 1 search_field_designation=driver.find_elements(By.CLASS_NAME,"suggestor-input ")
      2 search_field_designation.send_keys('Data Scientists')

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

In [14]:
```
search_field_location=driver.find_elements(By.XPATH,"/html/body/div[1]/div[7]/div/div/div[1]/div/div/div[1]/div[1]/div/input")
search_field_location.send_keys('Bangalore')
```
```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
Cell In[14], line 1
----> 1 search_field_location=driver.find_elements(By.XPATH,"/html/body/div[1]/div[7]/div/div/div[1]/div/div/div[1]/div[1]/div/input")
      2 search_field_location.send_keys('Bangalore')

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

In [15]:
```
search_button=driver.find_elements(By.XPATH,"qsbSubmit")
search_button.click()
```
```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
Cell In[15], line 1
----> 1 search_button=driver.find_elements(By.XPATH,"qsbSubmit")
      2 search_button.click()

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

In [16]:
```
job_title=[]
job_location=[]
company_name=[]
experience_required=[]
```

In [17]:
```
title_tags=driver.find_elements(By.CLASS_NAME,'//a[@class="title "]')
for i in title_tags:
    title=i.text
    job_title.append(title)

location_tags=driver.find_elements(By.CLASS_NAME,'//span[@class="locWdth"]')
for i in location_tags:
    location=i.text
    job_location.append(location)

company_tags=driver.find_elements(By.CLASS_NAME,'//div[@class=" row2"]/span/a[1]')
for i in company_tags:
    company=i.text
    company_name.append(comapny)

experience_tags=driver.find_elements(By.CLASS_NAME,'//span[@class="expwdth"]')
for i in experience_tags:
    experience=i.text
    experience_required.append(exp)
```
```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
Cell In[17], line 1
----> 1 title_tags=driver.find_elements(By.CLASS_NAME,'//a[@class="title "]')
      2 for i in title_tags:
      3     title=i.text

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

In [18]:
```
print(len(job_title),len(job_location),len(company_name),len(experience_required))
```
```
0 0 0 0
```

In [19]:
```
import pandas as pd
df=pd.DataFrame({'Title':job_title,'Location':job_location,'Company_name':company_name,'Experience':experience_required})
df
```

Out[19]:

Title	Location	Company_name	Experience

In [20]:
```
url=driver.find_elements(By.XPATH,'//a[@class="title "]')
url[0:4]
```
```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
Cell In[20], line 1
----> 1 url=driver.find_elements(By.XPATH,'//a[@class="title "]')
      2 url[0:4]

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

In [ ]:
```
for i in url[0:4]:
    print(i.get_attribute('href'))
```

In [ ]:
```
job_titles=[]
```

In [ ]:
```
start=0
end=10
for page in range(start,end):
    titles=driver.find_elements(By.XPATH,'//a[@class="title "]')
    for i in titles:
        job_titles.append(i.text)
    next_button=driver.find_element(By.XPATH,)
    next_button.click()
    time.sleep(3)
```

In [ ]:
```
len(job_titles)
```

In [ ]:
```
job_titles
```

Q2: Write a python program to scrape data for "Data Scientist" Job position in "Bangalore" location. You have to scrape the job-title, job-location, company_name, experience_required. You have to scrape first 10 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 the location" field.
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 [ ]:
```
driver.get("https://www.shine.com/")
```

In [ ]:
```
designation=driver.find_elements(By.CLASS_NAME,"suggestor-input ")
designation.send_keys('Data Analyst')
```

In [ ]:
```
location=driver.find_elements(By.XPATH,"/html/body/div[1]/div[7]/div/div/div[1]/div/div/div[1]/div[1]/div/input")
location.send_keys('Bangalore')
```

In [ ]:
```
search=driver.find_elements(By.CLASS_NAME,"qsbSubmit")
search.click()
```

In [ ]:
```
job_title=[]
job_location=[]
company_name=[]
experience_required=[]
```

In [21]:
```
title_tags=driver.find_elements(By.CLASS_NAME,'//a[@class="title "]')
for i in title_tags:
    title=i.text
    job_title.append(title)

location_tags=driver.find_elements(By.CLASS_NAME,'//span[@class="locWdth"]')
for i in location_tags:
    location=i.text
    job_location.append(location)

company_tags=driver.find_elements(By.ClASS_NAME,'//div[@class=" row2"]/span/a[1]')
for i in company_tags:
    company=i.text
    company_name.append(comapny)

experience_tags=driver.find_elements(By.CLASS_NAME,'//span[@class="expwdth"]')
for i in experience_tags:
    experience=i.text
    experience_required.append(exp)
```
```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
Cell In[21], line 1
----> 1 title_tags=driver.find_elements(By.CLASS_NAME,'//a[@class="title "]')
      2 for i in title_tags:
      3     title=i.text

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

In [22]:
```
print(len(job_title),len(job_location),len(company_name),len(experience_required))
```
```
0 0 0 0
```

In [23]:
```
import pandas as pd
df=pd.DataFrame({'Title':job_title,'Location':job_location,'Company_name':company_name,'Experience':experience_required})
df
```

Out[23]:

Title	Location	Company_name	Experience

In [24]:
```
for i in url[0:4]:
    print(i.get_attribute('href'))
```
```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
Cell In[24], line 1
----> 1 for i in url[0:4]:
      2     print(i.get_attribute('href'))

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

In [25]:
```
job_titles=[]
```

In [26]:
```
start=0
end=10
for page in range(start,end):
    titles=driver.find_elements(By.XPATH,'//a[@class="title "]')
    for i in titles:
        job_titles.append(i.text)
    next_button=driver.find_element(By.XPATH,)
    next_button.click()
    time.sleep(3)
```
```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
Cell In[26], line 4
      2 end=10
      3 for page in range(start,end):
----> 4     titles=driver.find_elements(By.XPATH,'//a[@class="title "]')
      5     for i in titles:
      6         job_titles.append(i.text)

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

In [27]:
```
len(job_titles)
```

Out[27]:
```
0
```

In [28]:
```
job_titles
```

Out[28]:
```
[]
```

Q3: Scrape 100 reviews data from flipkart.com for iphone11 phone. You have to go the link: https://www.flipkart.com/apple-iphone-11-black-64-gb/product￼reviews/itm4e5041ba101fd?pid=MOBFWQ6BXGJCEYNY&lid=LSTMOBFWQ6BXGJCEYNYZXSHRJ&marketplace=F LIPKART As shown in the above page you have to scrape the tick marked attributes. These are:

1. Rating
2. Review summary
3. Full review
4. You have to scrape this data for first 100reviews.

In [29]:
```
driver.get("https://www.flipkart.com/apple-iphone-11-black-64-gb/productreviews/itm4e5041ba101fd?pid=MOBFWQ6BXGJCEYNY&lid=LSTMOBFWQ6BXGJCEYNYZXSHRJ&marketplace
```
```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
Cell In[29], line 1
----> 1 driver.get("https://www.flipkart.com/apple-iphone-11-black-64-gb/product￼reviews/itm4e5041ba101fd?pid=MOBFWQ6BXGJCEYNY&lid=LSTMOBFWQ6BXGJCEYNYZXSHRJ&m
arketplace=F")

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

In [30]:
```
Rating=driver.find_elements(By.CLASS_NAME,"suggestor-input ")
Rating.send_keys('iphone11')
```
```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
Cell In[30], line 1
----> 1 Rating=driver.find_elements(By.CLASS_NAME,"suggestor-input ")
      2 Rating.send_keys('iphone11')

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

In [31]:
```
Rating=[]
Review_summary=[]
Full_review=[]
```

In [32]:
```
Rating_tags=driver.find_elements(By.CLASS_NAME,'//a[@class="title "]')
for i in title_tags:
    Rating=i.text
    Rating.append(title)

summary_tags=driver.find_elements(By.CLASS_NAME,'//span[@class="locWdth"]')
for i in location_tags:
    summary=i.text
    Review_summary.append(location)

Full_tags=driver.find_elements(By.ClASS_NAME,'//div[@class=" row2"]/span/a[1]')
for i in company_tags:
    Full=i.text
    Full_review.append(comapny)
```
```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
Cell In[32], line 1
----> 1 Rating_tags=driver.find_elements(By.CLASS_NAME,'//a[@class="title "]')
      2 for i in title_tags:
      3     Rating=i.text

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

In [33]:
```
Rating=[]
```

In [34]:
```
start=0
end=100
for page in range(start,end):
    Rating=driver.find_elements(By.XPATH,'//a[@class="title "]')
    for i in titles:
        Rating.append(i.text)
    next_button=driver.find_element(By.XPATH,)
    next_button.click()
    time.sleep(3)
```
```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
Cell In[34], line 4
      2 end=100
      3 for page in range(start,end):
----> 4     Rating=driver.find_elements(By.XPATH,'//a[@class="title "]')
      5     for i in titles:
      6         Rating.append(i.text)

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

In [ ]:

Q4: Scrape data forfirst 100 sneakers you find whenyouvisitflipkart.com and search for "sneakers" inthe 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 [35]:
```
driver.get("https://www.shine.com/")
```
```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
Cell In[35], line 1
----> 1 driver.get("https://www.shine.com/")

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

In [36]:
```
Brand=driver.find_elements(By.CLASS_NAME,"")
Brand.send_keys('sneakers')
```
```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
Cell In[36], line 1
----> 1 Brand=driver.find_elements(By.CLASS_NAME,"")
      2 Brand.send_keys('sneakers')

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

In [37]:
```
Brand=[]
ProductDescription=[]
Price=[]
```

In [38]:
```
Brand_tags=driver.find_elements(By.CLASS_NAME,'//a[@class="title "]')
for i in title_tags:
    Brand=i.text
    Brand.append(title)

ProductDescription_tags=driver.find_elements(By.CLASS_NAME,'//span[@class="locWdth"]')
for i in location_tags:
    ProductDescription=i.text
    ProductDescription.append(location)

Price_tags=driver.find_elements(By.ClASS_NAME,'//div[@class=" row2"]/span/a[1]')
for i in company_tags:
    Price=i.text
    Price.append(comapny)
```
```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
Cell In[38], line 1
----> 1 Brand_tags=driver.find_elements(By.CLASS_NAME,'//a[@class="title "]')
      2 for i in title_tags:
      3     Brand=i.text

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

Q5: Go to webpage https://www.amazon.in/ Enter "Laptop" in the search field and then click the search icon. Then set CPU Type filter to "Intel Core i7" as shown in the below image: Aftersetting the filters scrape first 10 laptops data. You have to scrape 3 attributes for each laptop:

1. Title
2. Ratings
3. Price

In [39]:
```
driver.get("https://www.amazon.in/")
```
```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
Cell In[39], line 1
----> 1 driver.get("https://www.amazon.in/")

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

In [40]:
```
Title=[]
Ratings=[]
Price=[]
```

In [41]:
```
Title_tags=driver.find_elements(By.CLASS_NAME,'//a[@class="title "]')
for i in title_tags:
    Title=i.text
    Title.append(title)

Ratings_tags=driver.find_elements(By.CLASS_NAME,'//span[@class="locWdth"]')
for i in location_tags:
    Ratings=i.text
    Ratings.append(Ratings)

Price_tags=driver.find_elements(By.ClASS_NAME,'//div[@class=" row2"]/span/a[1]')
for i in company_tags:
    Price=i.text
    Price.append(price)
```
```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
Cell In[41], line 1
----> 1 Title_tags=driver.find_elements(By.CLASS_NAME,'//a[@class="title "]')
      2 for i in title_tags:
      3     Title=i.text

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

In [ ]:

Q6: Write a python program to scrape data for Top 1000 Quotes of All Time. The above task will be done in following steps:

1. First get the webpagehttps://www.azquotes.com/
2. Click on TopQuote
3. Than scrap a)Quote b) Author c) Type Of Quotes

In [42]:
```
driver.get("//www.azquotes.com/")
```
```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
Cell In[42], line 1
----> 1 driver.get("//www.azquotes.com/")

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

In [43]:
```
Quote=[]
Author=[]
Types_of_Quotes=[]
```

In [44]:
```
Quote_tags=driver.find_elements(By.CLASS_NAME,'//a[@class="title "]')
for i in title_tags:
    Quote=i.text
    Quote.append(Quote)

Author_tags=driver.find_elements(By.CLASS_NAME,'//span[@class="locWdth"]')
for i in location_tags:
    Author=i.text
    Author.append(Author)

Types_of_tags=driver.find_elements(By.ClASS_NAME,'//div[@class=" row2"]/span/a[1]')
for i in company_tags:
    Types=i.text
    Types.append(Types)
```
```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
Cell In[44], line 1
----> 1 Quote_tags=driver.find_elements(By.CLASS_NAME,'//a[@class="title "]')
      2 for i in title_tags:
      3     Quote=i.text

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

In [ ]:

Q7: Write a python program to display list of respected former Prime Ministers of India (i.e. Name, Born-Dead, Term of office, Remarks) from https://www.jagranjosh.com/general-knowledge/list-of￼all-prime-ministers-of-india-1473165149-1 scrap the mentioned data and make the DataFrame

In [45]:
```
driver.get("https://www.jagranjosh.com/general-knowledge/list-ofall-prime-ministers-of-india-1473165149-1")
```
```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
Cell In[45], line 1
----> 1 driver.get("https://www.jagranjosh.com/general-knowledge/list-of￼all-prime-ministers-of-india-1473165149-1")

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

In [46]:
```
Name=[]
Born-Dead=[]
Term_of_office=[]
```
```
  Cell In[46], line 2
    Born-Dead=[]
    ^
SyntaxError: cannot assign to expression here. Maybe you meant '==' instead of '='?
```

In [47]:
```
Name_tags=driver.find_elements(By.CLASS_NAME,'//a[@class="title "]')
for i in title_tags:
    Name=i.text
    Name.append(Name)

Born-Dead_tags=driver.find_elements(By.CLASS_NAME,'//span[@class="locWdth"]')
for i in location_tags:
    Born-Dead=i.text
    Born-Dead.append(Born-Dead)

officeTerm_tags=driver.find_elements(By.ClASS_NAME,'//div[@class=" row2"]/span/a[1]')
for i in company_tags:
    officeTerm=i.text
    officeTerm.append(officeTerm)
```
```
  Cell In[47], line 6
    Born-Dead_tags=driver.find_elements(By.CLASS_NAME,'//span[@class="locWdth"]')
    ^
SyntaxError: cannot assign to expression here. Maybe you meant '==' instead of '='?
```

Q8: Write a python program to display list of 50 Most expensive cars in the world (i.e. Car name and Price) from https://www.motor1.com/ This task will be done in following steps:

1. First get the webpage https://www.motor1.com/
2. Then You have to type in the search bar '50 most expensive cars'
3. Then click on 50 most expensive carsin the world..
4. Then scrap thementioned data and make the dataframe

In [48]:
```
driver.get("https://www.motor1.com/")
```
```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
Cell In[48], line 1
----> 1 driver.get("https://www.motor1.com/")

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

In [49]:
```
Car_name=driver.find_elements(By.CLASS_NAME," ")
Car_name.send_keys('cars')
```
```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
Cell In[49], line 1
----> 1 Car_name=driver.find_elements(By.CLASS_NAME," ")
      2 Car_name.send_keys('cars')

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

In [ ]:
```
price=driver.find_elements(By.CLASS_NAME,"suggestor-input ")
price.send_keys('cars')
```

In [ ]:
```
search=driver.find_elements(By.CLASS_NAME,"qsbSubmit")
search.click()
```

In [50]:
```
car_name=[]
price=[]
```

In [51]:
```
car_tags=driver.find_elements(By.CLASS_NAME,'//a[@class="title "]')
for i in title_tags:
    car=i.text
    car.append(car)

price_tags=driver.find_elements(By.CLASS_NAME,'//span[@class="locWdth"]')
for i in location_tags:
    price=i.text
    price.append(price)
```
```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
Cell In[51], line 1
----> 1 car_tags=driver.find_elements(By.CLASS_NAME,'//a[@class="title "]')
      2 for i in title_tags:
      3     car=i.text

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