Web Scraping

Requests library

Web scraping starts with sending HTTP requests, such as POST or GET, to a website's server, which returns a response containing the needed data

In [31]:
```
#pip install requests
```

In [ ]:
```
pip install BeautifulSoup
```

Requests library provides easy methods for sending HTTP GET and POST requests.
For example, the function to send an HTTP Get request is named as get():

In [2]:
```
import requests
url="https://oxylabs.io/"
response = requests.get(url)
#print(response.text)
```

But this library has a limitation in that it does not parse the extracted HTML data, i.e., it cannot convert the data into a more readable format for analysis.

Also, it cannot be used to scrape websites that are written using purely JavaScript.

Beautiful Soup

Beautiful Soup is a Python library that works with a parser to extract data from HTML and can turn even invalid markup into a parse tree. However, this library is only designed for parsing and cannot request data from web servers in the form of HTML documents/files.
For this reason, it is mostly used alongside the Python Requests Library. Note that Beautiful Soup makes it easy to query and navigate the HTML, but still requires a parser.
The following example demonstrates the use of the html.parser module, which is part of the Python Standard Library.

Part 1 – Get the HTML using Requests

In [ ]:
```
import requests
url='https://oxylabs.io/blog'
response = requests.get(url)
```

Part 2 – Find the element

In [3]:
```
from bs4 import BeautifulSoup
soup = BeautifulSoup(response.text, 'html.parser')
print(soup.title)
```

```
<title>Innovative Proxy Service to Gather Data at Scale | Oxylabs</title>
```

Active Stock Data Using Web Scraping

In [16]:
```
from bs4 import BeautifulSoup
import requests
```

In [17]:
```
url="https://finance.yahoo.com/most-active"
response=requests.get(url)
response.text
```

dware","criteria":{"operator":"EQ","operands":["peer_group","Technology Hardware"]},"dependentFilterLabel":null},{"displayName":"Transportation","criteria":{"operator":"EQ","operands":["peer_group","Transportation"]},"dependentFilterLabel":null},{"displayName":"Consumer Services","criteria":{"operator":"EQ","operands":["peer_group","Consumer Services"]},"dependentFilterLabel":null},{"displayName":"Media","criteria":{"operator":"EQ","operands":["peer_group","Media"]},"dependentFilterLabel":null},{"displayName":"Automobiles","criteria":{"operator":"EQ","operands":["peer_group","Automobiles"]},"dependentFilterLabel":null},{"displayName":"Semiconductors","criteria":{"operator":"EQ","operands":["peer_group","Semiconductors"]},"dependentFilterLabel":null}

In [26]:
```python
for listing in stock_table.find_all('tr'):
    code=listing.find('td',attrs={'aria-label':"Symbol"})
    codes.append(code.text)
```

In [27]:
```python
len(codes)
```

Out[27]: 25

Stock Name

In [28]:
```python
for listing in stock_table.find_all('tr'):
    name=listing.find('td',attrs={'aria-label':"Name"})
    names.append(name.text)
```

In [29]:
```python
len(names)
```

Out[29]: 25

Stock Prices

In [30]:
```python
for listing in stock_table.find_all('tr'):
    price=listing.find('td',attrs={'aria-label':"Price (Intraday)"})
    prices.append(price.text)
```

In [31]:
```python
len(prices)
```

Out[31]: 25

Change

In [32]:
```python
for listing in stock_table.find_all('tr'):
    change=listing.find('td',attrs={'aria-label':"Change"})
    changes.append(change.text)
```

In [33]:
```python
len(changes)
```

Out[33]: 25

total_volume

In [34]:
```python
for listing in stock_table.find_all('tr'):
    total_volume=listing.find('td',attrs={'aria-label':"Volume"})
    total_volumes.append(total_volume.text)
```

In [35]:
```python
len(total_volumes)
```

Out[35]: 25

Percent_changes

In [36]:
```python
for listing in stock_table.find_all('tr'):
    percent_change=listing.find('td',attrs={'aria-label':"% Change"})
    percent_changes.append(percent_change.text)
```

In [37]:
```python
len(percent_changes)
```

Out[37]: 25

Market_caps

In [38]:
```python
for listing in stock_table.find_all('tr'):
    market_cap=listing.find('td',attrs={'aria-label':"Market Cap"})
    market_caps.append(market_cap.text)
```

In [39]:
```python
len(market_caps)
```

Out[39]: 25

In [40]:
```python
import pandas as pd
```

In [41]:
```python
a={
"Symbol":codes,
"Name":names,
"Price":prices,
"Change":changes,
"% Change":percent_changes,
"Volume":total_volumes,
"Market Capital":market_caps
}
```

In [42]:
```python
df=pd.DataFrame(
{
"Symbol":codes,
"Name":names,
"Price":prices,
"Change":changes,
"% Change":percent_changes,
"Volume":total_volumes,
"Market Capital":market_caps
})
```

In [43]:
```python
df
```

Out[43]:

	Symbol	Name	Price	Change	% Change	Volume	Market Capital
0	TWTR	Twitter, Inc.	42.36	-0.77	-1.68%	252.397M	34.422B
1	BAC	Bank of America Corporation	37.57	-1.25	-3.22%	79.032M	302.997B
2	AAPL	Apple Inc.	165.29	-5.11	-3.00%	75.329M	2.697T
3	AMD	Advanced Micro Devices, Inc.	93.06	-4.68	-4.79%	72.482M	151.442B
4	WFC	Wells Fargo & Company	46.35	-2.19	-4.51%	65.583M	176.204B
5	NVDA	NVIDIA Corporation	212.58	-9.45	-4.26%	56.913M	529.749B
6	T	AT&T Inc.	19.54	+0.12	+0.62%	52.707M	139.966B
7	AAL	American Airlines Group Inc.	19.00	+0.05	+0.26%	42.822M	12.334B
8	SOFI	SoFi Technologies, Inc.	7.26	-0.35	-4.60%	42.198M	5.805B
9	PBR	Petróleo Brasileiro S.A. - Petrobras	14.71	-0.03	-0.21%	42.065M	101.164B
10	NIO	NIO Inc.	19.65	-0.77	-3.77%	43.778M	35.945B
11	F	Ford Motor Company	15.48	-0.03	-0.19%	39.313M	62.211B
12	WBD	Warner Bros. Discovery, Inc.	24.88	-1.12	-4.31%	37.866M	12.49B
13	PCG	PG&E Corporation	12.30	-0.55	-4.28%	37.224M	24.445B
14	INTC	Intel Corporation	45.67	-1.34	-2.85%	36.182M	186.731B
15	C	Citigroup Inc.	50.93	+0.78	+1.56%	34.043M	100.458B
16	SWN	Southwestern Energy Company	8.19	-0.07	-0.85%	33.52M	9.141B
17	ITUB	Itaú Unibanco Holding S.A.	5.54	-0.02	-0.36%	32.211M	54.181B
18	TELL	Tellurian Inc.	6.19	+0.40	+6.91%	30.185M	3.209B
19	BBD	Banco Bradesco S.A.	4.5000	-0.0500	-1.10%	30.49M	47.83B
20	DAL	Delta Air Lines, Inc.	42.36	+1.34	+3.27%	29.129M	27.107B
21	MSFT	Microsoft Corporation	279.83	-7.79	-2.71%	28.037M	2.098T
22	VALE	Vale S.A.	19.48	-0.43	-2.16%	25.586M	94.376B
23	AMC	AMC Entertainment Holdings, Inc.	18.02	-0.51	-2.75%	25.334M	9.258
24	CCL	Carnival Corporation & plc	19.56	+0.05	+0.26%	25.317M	22.991B

In [44]:
```python
df.to_csv("active_stock_data.csv")
```

In [45]:
```python
stocks=pd.read_csv("active_stock_data.csv")
```

In [46]:
```python
stocks.tail()
```

Out[46]:

	Unnamed: 0	Symbol	Name	Price	Change	% Change	Volume	Market Capital
20	20	DAL	Delta Air Lines, Inc.	42.36	1.34	+3.27%	29.129M	27.107B
21	21	MSFT	Microsoft Corporation	279.83	-7.79	-2.71%	28.037M	2.098T
22	22	VALE	Vale S.A.	19.48	-0.43	-2.16%	25.586M	94.376B
23	23	AMC	AMC Entertainment Holdings, Inc.	18.02	-0.51	-2.75%	25.334M	9.258
24	24	CCL	Carnival Corporation & plc	19.56	0.05	+0.26%	25.317M	22.991B

In [48]:
```python
df
```

Out[48]:

	Symbol	Name	Price	Change	% Change	Volume	Market Capital
0	TWTR	Twitter, Inc.	42.36	-0.77	-1.68%	252.397M	34.422B
1	BAC	Bank of America Corporation	37.57	-1.25	-3.22%	79.032M	302.997B
2	AAPL	Apple Inc.	165.29	-5.11	-3.00%	75.329M	2.697T
3	AMD	Advanced Micro Devices, Inc.	93.06	-4.68	-4.79%	72.482M	151.442B
4	WFC	Wells Fargo & Company	46.35	-2.19	-4.51%	65.583M	176.204B
5	NVDA	NVIDIA Corporation	212.58	-9.45	-4.26%	56.913M	529.749B
6	T	AT&T Inc.	19.54	+0.12	+0.62%	52.707M	139.966B
7	AAL	American Airlines Group Inc.	19.00	+0.05	+0.26%	42.822M	12.334B
8	SOFI	SoFi Technologies, Inc.	7.26	-0.35	-4.60%	42.198M	5.805B
9	PBR	Petróleo Brasileiro S.A. - Petrobras	14.71	-0.03	-0.21%	42.065M	101.164B
10	NIO	NIO Inc.	19.65	-0.77	-3.77%	43.778M	35.945B
11	F	Ford Motor Company	15.48	-0.03	-0.19%	39.313M	62.211B
12	WBD	Warner Bros. Discovery, Inc.	24.88	-1.12	-4.31%	37.866M	12.49B
13	PCG	PG&E Corporation	12.30	-0.55	-4.28%	37.224M	24.445B
14	INTC	Intel Corporation	45.67	-1.34	-2.85%	36.182M	186.731B
15	C	Citigroup Inc.	50.93	+0.78	+1.56%	34.043M	100.458B
16	SWN	Southwestern Energy Company	8.19	-0.07	-0.85%	33.52M	9.141B
17	ITUB	Itaú Unibanco Holding S.A.	5.54	-0.02	-0.36%	32.211M	54.181B
18	TELL	Tellurian Inc.	6.19	+0.40	+6.91%	30.185M	3.209B
19	BBD	Banco Bradesco S.A.	4.5000	-0.0500	-1.10%	30.49M	47.83B
20	DAL	Delta Air Lines, Inc.	42.36	+1.34	+3.27%	29.129M	27.107B
21	MSFT	Microsoft Corporation	279.83	-7.79	-2.71%	28.037M	2.098T
22	VALE	Vale S.A.	19.48	-0.43	-2.16%	25.586M	94.376B
23	AMC	AMC Entertainment Holdings, Inc.	18.02	-0.51	-2.75%	25.334M	9.258
24	CCL	Carnival Corporation & plc	19.56	+0.05	+0.26%	25.317M	22.991B

In [ ]: