Web Scraping using BeautifulSoup

Objectives:

- Using `requests` to download server-side rendered HTML code
- Using `BeautifulSoup` to parse HTML code

Scrape for Latest COE Price

We will extract latest COE price from following website:

- https://www.onemotoring.com.sg/content/onemotoring/home/buying/coe-open-bidding.html (https://www.onemotoring.com.sg/content/onemotoring/home/buying/coe-open-bidding.html)

Confirm that the desired data in webpage is **server-side rendered**.

- Copy a string of the desired data on webpage
- Right click on webpage and select `View Page Source`
- The string should be found in the HTML code

In [2]:

```
!pip install beautifulsoup4
```

```
Requirement already satisfied: beautifulsoup4 in c:\users\isszq\anaconda3\l
ib\site-packages (4.9.1)
Requirement already satisfied: soupsieve>1.2 in c:\users\isszq\anaconda3\li
b\site-packages (from beautifulsoup4) (2.0.1)
```

In [6]:

```
import bs4
bs4.__version__
```

Out[6]:

```
'4.9.1'
```

Make Soup

Import libraries.

In [1]:

```
from bs4 import BeautifulSoup
import requests
```

Use `requests` to send GET request to server and download HTML.

- Use status code to make sure request is successful.

In [12]:

```python
URL = 'https://www.onemotoring.com.sg/content/onemotoring/home/buying/coe

resp = requests.get(URL)
print(resp.status_code)
```

```
200
```

Make a soup from HTML code, which is in `resp.text` .

In [5]:

```python
soup = BeautifulSoup(resp.text)
print(soup.title)
print(soup.title.text)
```

```
<title>COE Open Bidding | Buying | One Motoring</title>
COE Open Bidding | Buying | One Motoring
```

Inspect HTML Elements

Open URL in web browser; Right click on targeted element in webpage and select `Inspect` from context menu.

- It will open the `Element` pane in **Chrome DevTools**
- Examine the HTML code. The data are contained in 2 `<table>` element with attribute `style="width: 100%;"` .

Find the 2 tables using `find_all()` method.

In [11]:

```python
tables = soup.find_all('table', {'style':"width: 100%;"})
print(len(tables))
```

```
2
```

Extract 1st Table - COE Price

Extract all `<tr>` which each contains a row.

In [47]:

```python
tr_list = tables[0].find_all('tr')
print(len(tr_list))
```

```
6
```

Header

Extract table header from each `<tr>` .

In [48]:

```python
th_list = tr_list[0].find_all('th')
header = [ th.text for th in th_list ]
print(header)
header.insert(1, 'Description')
print(header)
```

```
['Category', 'Quota', 'QP($)']
['Category', 'Description', 'Quota', 'QP($)']
```

Table Data

Extract table data from each `<tr>` .

In [49]:

```python
data = []
for tr in tr_list:
    td_list = tr.find_all('td')
    row = [ td.text for td in td_list ]
    if row:
        data.append(row)

print(data)
```

```
[['A', 'CAR UP TO 1600CC & 97KW', '1035', '37766'], ['B', 'CAR ABOVE 1600CC
OR 97KW', '904', '41510'], ['C', 'GOODS VEHICLE & BUS', '354', '26644'],
['D', 'MOTORCYCLE', '496', '7399'], ['E', 'OPEN-ALL EXCEPT MOTORCYCLE', '47
0', '40790']]
```

Write to csv file `coe_price.csv` .

In [50]:

```python
import csv

with open('coe_price.csv', 'w', newline='') as f:
    writer = csv.writer(f)
    writer.writerow(header)
    writer.writerows(data)
```

Examine data in file `coe_price.csv` .

In [51]:

```python
!notepad coe_price.csv
```

Exercise

Extract 2nd Table - COE Bids

Extract all `<tr>` which each contains a row.

In [52]:

```python
tr_list = tables[1].find_all('tr')
print(len(tr_list))
```

```
6
```

Header

Extract table header from each `<tr>`.

In [53]:

```python
th_list = tr_list[0].find_all('th')
header = [ th.text for th in th_list ]
print(header)
header.insert(1, 'Description')
print(header)
```

```
['Category', 'Received', 'Successful', 'Unsuccessful', 'Unused']
['Category', 'Description', 'Received', 'Successful', 'Unsuccessful', 'Unus
ed']
```

Table Data

Extract table data from each `<tr>`.

In [55]:

```python
data = []
for tr in tr_list:
    td_list = tr.find_all('td')
    row = [ td.text for td in td_list ]
    print(row)
    if row:
        data.append(row)
```

```
[]
['A', 'CAR UP TO 1600CC & 97KW', '1737', '1035', '702', '0']
['B', 'CAR ABOVE 1600CC OR 97KW', '1715', '892', '823', '12']
['C', 'GOODS VEHICLE & BUS', '525', '350', '175', '4']
['D', 'MOTORCYCLE', '691', '488', '203', '8']
['E', 'OPEN-ALL EXCEPT MOTORCYCLE', '672', '470', '202', '0']
```

In [56]:

```python
print(data)
```

```
[['A', 'CAR UP TO 1600CC & 97KW', '1737', '1035', '702', '0'], ['B', 'CAR A
BOVE 1600CC OR 97KW', '1715', '892', '823', '12'], ['C', 'GOODS VEHICLE & B
US', '525', '350', '175', '4'], ['D', 'MOTORCYCLE', '691', '488', '203',
'8'], ['E', 'OPEN-ALL EXCEPT MOTORCYCLE', '672', '470', '202', '0']]
```

Write to csv file `coe_bids.csv`.

In [57]:

```python
import csv

with open('coe_bids.csv', 'w', newline='') as f:
    writer = csv.writer(f)
    writer.writerow(header)
    writer.writerows(data)
```

Examine data in file `coe_bids.csv` .

In [58]:

```python
!notepad coe_bids.csv
```