Web Scrapping

In [1]:
```
import urllib
import re
import pandas as pd
```

In [6]:
```
url = 'http://olympus.realpython.org/profiles/poseidon'
```

In [7]:
```
data = urllib.request.urlopen(url)
```

In [8]:
```
data
```

Out[8]:
```
<http.client.HTTPResponse at 0x16d7598a640>
```

In [9]:
```
data1 = data.read()
```

In [10]:
```
data1
```

Out[10]:
```
b'<html>\n<head>\n<title >Profile: Poseidon</title>\n</head>\n<body bgcolor
="yellow">\n<center>\n<br><br>\n<img src="/static/poseidon.jpg" />\n<h2>Name:
Poseidon</h2>\n<br><br>\nFavorite animal: Dolphin\n<br><br>\nFavorite color:
Blue\n<br><br>\nHometown: Sea\n</center>\n</body>\n</html>\n'
```

In [11]:
```
data2  = data1.decode()
```

In [12]:
```
data2
```

Out[12]:
```
'<html>\n<head>\n<title >Profile: Poseidon</title>\n</head>\n<body bgcolor="y
ellow">\n<center>\n<br><br>\n<img src="/static/poseidon.jpg" />\n<h2>Name: Po
seidon</h2>\n<br><br>\nFavorite animal: Dolphin\n<br><br>\nFavorite color: Bl
ue\n<br><br>\nHometown: Sea\n</center>\n</body>\n</html>\n'
```

In [13]:
```
my_pattern = re.sub("<.*?>"," ", data2)
```

In [14]:
```
print(my_pattern)
```

```
 Profile: Poseidon



 Name: Poseidon

Favorite animal: Dolphin

Favorite color: Blue

Hometown: Sea
```

In [26]:
```
my_string = re.findall('[a-zA-z]{1,8}', my_pattern)
```

In [27]:
```
my_string
```

Out[27]:
```
['Profile',
 'Poseidon',
 'Name',
 'Poseidon',
 'Favorite',
 'animal',
 'Dolphin',
 'Favorite',
 'color',
 'Blue',
 'Hometown',
 'Sea']
```

In [29]:
```
df = pd.DataFrame({"Profile":my_string})
```

In [31]:
```
df.to_csv('data.csv', index=False)
```

In [ ]:
```

```

In [36]:
```
df = pd.read_csv('data1.csv')
```

In [37]:
```
df.head()
```

Out[37]:

	Profile
0	Profile
1	Poseidon
2	Name
3	Poseidon
4	Favorite

In [38]:
```
df.tail()
```

Out[38]:

	Profile
20	Favorite
21	color
22	Blue
23	Hometown
24	Sea

In [108]:
```
import re
```

In [109]:
```
my_text = "My name is Shyam Ambilkar"
```

In [110]:
```
my_pattern = re.findall(r"am\B", my_text)
```

In [111]:
```
my_pattern
```

Out[111]:
```
['am']
```

In [112]:
```
if my_pattern:
    print("Yes pattern is available")
else:
    print("No pattern match")
```

```
Yes pattern is available
```

In [115]:
```
my_text = "My name is Shyam Ambilkar"
my_pattern = re.findall("\S", my_text)
```

In [116]:
```
my_pattern
```

Out[116]:
```
['M',
 'y',
 'n',
 'a',
 'm',
 'e',
 'i',
 's',
 'S',
 'h',
 'y',
 'a',
 'm',
 'A',
 'm',
 'b',
 'i',
 'l',
 'k',
 'a',
 'r']
```

In [123]:
```
my_text = "My name is Shyam # Ambilkar 412207"
my_pattern = re.findall("\w", my_text)
```

In [124]:
```
my_pattern
```

Out[124]:
```
['M',
 'y',
 'n',
 'a',
 'm',
 'e',
 'i',
 's',
 'S',
 'h',
 'y',
 'a',
 'm',
 'A',
 'm',
 'b',
 'i',
 'l',
 'k',
 'a',
 'r',
 '4',
 '1',
 '2',
 '2',
 '0',
 '7']
```

In [125]:
```
import re
```

In [128]:
```
my_text = "My name is Python Welcome to DWH"

my_pattern =re.findall("[abc]", my_text)
```

In [129]:
```
my_pattern
```

Out[129]:
```
['a', 'c']
```

In [136]:
```
my_text = "My name is Python Welcome to DWH"

my_pattern =re.findall("[a-eA-D]", my_text)
```

In [137]:
```
my_pattern
```

Out[137]:
```
['a', 'e', 'e', 'c', 'e', 'D']
```

In [138]:
```
my_text = "My name is Python Welcome to DWH"

my_pattern =re.findall("[^arn]", my_text)
```

In [139]:
```
my_pattern
```

Out[139]:
```
['M',
 'y',
 ' ',
 'm',
 'e',
 ' ',
 'i',
 's',
 ' ',
 'P',
 'y',
 't',
 'h',
 'o',
 ' ',
 'W',
 'e',
 'l',
 'c',
 'o',
 'm',
 'e',
 ' ',
 't',
 'o',
 ' ',
 'D',
 'W',
 'H']
```

In [142]:
```
my_text = "My name is Python Welcome to DWH 9923090436"

my_pattern =re.findall("[0123]", my_text)
```

In [143]:
```
print(my_pattern)
```

```
['2', '3', '0', '0', '3']
```

In [ ]:
```

```