Web Scrapping - Case_Internshala.

Prepared by: Sagun Shakya (https://github.com/sagsshakya)

- GITAM Institute of Science.

Imporing the necessary libraries.

In [2]:
```
import requests
from bs4 import BeautifulSoup as BS
```

In [3]:
```
url = 'https://internshala.com/internships/work-from-home-data%20science-jobs'
page = requests.get(url)
soup = BS(page.text, 'html.parser')

print(soup.prettify())
```

```
<!DOCTYPE html>
<html xmlns="http://www.w3.org/1999/xhtml" xmlns:fb="https://www.facebook.com/2008/fbml" xmlns:og="http
 <head>
  <meta content="IE=9" http-equiv="X-UA-Compatible"/>
  <meta charset="utf-8"/>
  <meta content="width=device-width, initial-scale=1.0 user-scalable=0" name="viewport"/>
  <meta content="272234782795210" property="fb:app_id"/>
  <meta content="article" property="og:type"/>
  <meta content="1200" property="og:image:width"/>
  <meta content="630" property="og:image:height"/>
  <meta content="@Internshala" name="twitter:site"/>
  <meta content="summary_large_image" name="twitter:card"/>
  <meta content="@internshala" name="twitter:creator"/>
  <meta content="https://internshala.com/static/images/internships_for_facebook.png" name="twitter:imag
  <meta content="#1295c9" name="theme-color"/>
  <meta content="#1295c9" name="msapplication-navbutton-color"/>
  <script defer="" src="https://internshala.com/static/js/includes/common/jquery-1.11.1.min.js">
  </script>
  <script defer="" src="https://internshala.com/static/cdn/3.3.6/bootstrap.min.js">
  </script>
  <link href="https://internshala.com/static/cdn/3.3.6/bootstrap.css" rel="stylesheet"/>
  <link href="https://internshala.com/static/cdn/fonts/open_sans/open_sans_min.css" rel="stylesheet" ty
  <link href="https://internshala.com/favicon.ico?v=3" rel="icon"/>
  <script type="application/ld+json">
```

In [4]:

```python
company_class = soup.find_all(class_ = 'company')

print(company_class[0])
print('----------------------------------------')
print(company_class[0].h4.a.text)
print('----------------------------------------')
```

```
<div class="company">
<h4 title="Deep Learning">
<a href="/internship/detail/deep-learning-work-from-home-job-internship-at-iit-bombay1585177536">Deep L
<h4>
<a class="link_display_like_text" href="/internships/internship-at-IIT%20Bombay" title="IIT Bombay">
                        IIT Bombay                    </a>
</h4>
</div>
----------------------------------------
Deep Learning
----------------------------------------
```

Getting the Fields list.

In [5]:

```python
fields_list = [company_class[ii].h4.a.text for ii in range(len(company_class))]
print(len(fields_list))
print('---------------------')
print(fields_list)
```

```
40
---------------------
['Deep Learning', 'Data Analytics', 'Teaching (Business Analytics)', 'Data Science', 'Machine Learning'
cs', 'Data Analytics', 'Machine Learning', 'Product Management', 'Data Analytics', 'Machine Learning',
tics', 'Data Science', 'Data Science', 'Business Analytics', 'Business Analytics', 'Web Development', '
ing', 'Machine Learning', 'Flask Application Development', 'Machine Learning', 'Machine Learning', 'Cus
lytics Using R Programming', 'Machine Learning', 'Machine Learning & Computer Vision', 'Artificial Neur
nt & Analytics', 'Stock Market Analysis (Technical & Fundamental)', 'Machine Learning', 'Artificial Int
rocessing)', 'Research Analytics', 'Machine Learning', 'Data Science', 'B2B Alliance Analytics', 'Data 
```

Getting the company list.

In [6]:

```
company = soup.find_all(class_ = 'link_display_like_text')
company_list = [company[ii].text for ii in range(len(company))]
company_list[:5]
```

```
['\n                        IIT Bombay                    ',
 '\n                        Motilal Oswal                    ',
 '\n                        UpGrad                    ',
 '\n                        Shyena Tech Yarns Private Limited                    ',
 '\n                        TechnoYantra                    ']
```

Cleaning process.

In [8]:

```
for ii in range(len(company_list)):
    company_list[ii] = company_list[ii].replace('\n','')
company_list[:5]
```

```
['                        IIT Bombay                    ',
 '                        Motilal Oswal                    ',
 '                        UpGrad                    ',
 '                        Shyena Tech Yarns Private Limited                    ',
 '                        TechnoYantra                    ']
```

Removing the extraspaces at the front and the back of the string.

In [9]:

```
for ii in range(len(company_list)):
    company_list[ii] = company_list[ii].replace('                        ','')
for ii in range(len(company_list)):
    company_list[ii] = company_list[ii].replace('                    ','')
company_list[:5]
```

```
['IIT Bombay',
 'Motilal Oswal',
 'UpGrad',
 'Shyena Tech Yarns Private Limited',
 'TechnoYantra']
```

Testing the above results in a dataframe.

In [10]:

```
import pandas as pd
pd.DataFrame({'Companies':company_list,
             'Fields': fields_list})
```

	Companies	Fields
0	IIT Bombay	Deep Learning
1	Motilal Oswal	Data Analytics
2	UpGrad	Teaching (Business Analytics)
3	Shyena Tech Yarns Private Limited	Data Science
4	TechnoYantra	Machine Learning
5	Uttam Blastech Pvt. Ltd.	Machine Learning
6	Predicon.io	Data Analytics
7	NULL Innovation Private Limited	Data Analytics
8	Quesite Services Private Limited	Machine Learning
9	Bhigusa Health Care	Product Management
10	Analyticscosm	Data Analytics
11	Challenge Katta	Machine Learning
12	Kangaroo Rooms	Machine Learning
13	MedTourEasy	Business Analytics
14	Challenge Katta	Data Science
15	360DigiTMG	Data Science
16	Prutha Technologies Private Limited	Business Analytics
17	HR Ocuos	Business Analytics
18	Skill Hives	Web Development
19	Stirring Minds	Business Analytics
20	LineupX	Machine Learning
21	Intel Index	Machine Learning
22	AutomizeApps	Flask Application Development
23	Neolen	Machine Learning
24	CogniAble	Machine Learning
25	ShootMe.in	Customer Service Analytics
26	UFC Food LLP	Data Analytics Using R Programming
27	SkillBit	Machine Learning
28	Nion Technologies	Machine Learning & Computer Vision
29	Xovex IT International	Artificial Neural Networks
30	Frontlobe Insights	Environment Management & Analytics
31	FinThink Academy	Stock Market Analysis (Technical & Fundamental)
32	TransWeb Global Incorporation	Machine Learning
33	InventGrid India Private Limited	Artificial Intelligence/Machine Learning (Imag...
34	GoOffer Hyperlocal Private Limited	Research Analytics

	Companies	Fields
35	SJTech Solutions	Machine Learning
36	SkillBit	Data Science
37	ShootMe.in	B2B Alliance Analytics
38	SkillAngels	Data Analytics
39	Plan My Health	Financial Analytics

Getting the internship description.

In [11]:

```
x = soup.find_all(class_ = 'table-responsive')
print(x[0])
print('------------------')
print(x[1])
```

```
<div class="table-responsive">
<table class="table">
<thead>
<tr>
<th>Start Date</th>
<th>Duration</th>
<th>Stipend</th>
<th>Posted On</th>
<th>Apply By</th>
</tr>
</thead>
<tbody>
<tr>
<td>
<div id="start-date-first">Immediately</div>
</td>
<td>
                                6 Months                            </td>
<td class="stipend_container_table_cell">
<i class="fa fa-inr"></i>2000-4000 /month
                            </td>
<td>26 Mar'20</td>
<td>23 Apr'20</td>
</tr>
</tbody>
</table>
</div>
------------------
<div class="table-responsive">
<table class="table">
<thead>
<tr>
<th>Start Date</th>
<th>Duration</th>
<th>Stipend</th>
<th>Posted On</th>
<th>Apply By</th>
</tr>
</thead>
<tbody>
<tr>
<td>
<div id="start-date-first">20 Apr - 30 Apr'20</div>
</td>
<td>
                                4 Months                            </td>
<td class="stipend_container_table_cell">
<i class="fa fa-inr"></i>1000 /month
                            </td>
```

```
<td>20 Mar'20</td>
<td>17 Apr'20</td>
</tr>
</tbody>
</table>
</div>
```

Putting all the data into lists.

In [13]:

```
start_date_list = []
duration_list = []
stipend_list = []
posted_on_list = []
apply_by_list = []

x = soup.find_all(class_ = 'table-responsive')
for ii in range(len(x)):
    details = x[ii].find_all('td')

    start_date_list.append(details[0].text)
    duration_list.append(details[1].text)
    stipend_list.append(details[2].text)
    posted_on_list.append(details[3].text)
    apply_by_list.append(details[4].text)
```

Cleaning the data one by one.

Cleaning start_date_list.

In [15]:

```
start_date_list[:5]
```

```
['Immediately',
 "20 Apr - 30 Apr'20",
 'Immediately',
 'Immediately',
 'Immediately']
```

In [16]:

```
for ii in range(len(start_date_list)):
    start_date_list[ii] = start_date_list[ii].replace('\n','')
start_date_list[:10]
```

```
['Immediately',
 "20 Apr - 30 Apr'20",
 'Immediately',
 'Immediately',
 'Immediately',
 "11 May - 18 May'20",
 'Immediately',
 'Immediately',
 'Immediately',
 'Immediately']
```

Cleaning duration_list.

In [17]:

```
duration_list[:5]
```

```
['6 Months', '4 Months', '6 Months', '4 Months', '3 Months']
```

In [18]:

```
for ii in range(len(duration_list)):
    duration_list[ii] = duration_list[ii].replace('\n','')
    duration_list[ii] = duration_list[ii].replace('
    duration_list[ii] = duration_list[ii].replace('                    ',

duration_list[:5]
```

```
['6 Months', '4 Months', '6 Months', '4 Months', '3 Months']
```

Cleaning stipend_list.

In [19]:

```
stipend_list[:5]
```

```
['2000-4000 /month',
 '1000 /month',
 '15000 /month',
 '1000 /month',
 '5000-10000 /month']
```

In [20]:

```
for ii in range(len(stipend_list)):
    stipend_list[ii] = stipend_list[ii].replace('                              ',''
    stipend_list[ii] = stipend_list[ii].replace('\n','')
    #duration_list[ii] = duration_list[ii].replace('

stipend_list[:5]
```

```
['2000-4000 /month',
 '1000 /month',
 '15000 /month',
 '1000 /month',
 '5000-10000 /month']
```

Converting our data into a dataframe and exporti .csv file.

In [21]:

```
import pandas as pd
import os
os.chdir(r'C:\Users\acer\Desktop\PythonProgramming')
```

In [22]:

```
df = pd.DataFrame()

df['Companies'] = company_list
df['Fields'] =  fields_list
df['Start Date'] = start_date_list
df['Duration'] = duration_list
df['Stipend'] = stipend_list
df['Posted On'] = posted_on_list
df['Apply By'] = apply_by_list
df.head(10)
```

	Companies	Fields	Start Date	Duration	
0	IIT Bombay	Deep Learning	Immediately	6 Months	2000-4000 /m
1	Motilal Oswal	Data Analytics	20 Apr - 30 Apr'20	4 Months	1000 /month
2	UpGrad	Teaching (Business Analytics)	Immediately	6 Months	15000 /month
3	Shyena Tech Yarns Private Limited	Data Science	Immediately	4 Months	1000 /month
4	TechnoYantra	Machine Learning	Immediately	3 Months	5000-10000 /r
5	Uttam Blastech Pvt. Ltd.	Machine Learning	11 May - 18 May'20	2 Months	10000 lump-S
6	Predicon.io	Data Analytics	Immediately	6 Months	10000-20000
7	NULL Innovation Private Limited	Data Analytics	Immediately	4 Months	3000 /month
8	Quesite Services Private Limited	Machine Learning	Immediately	4 Weeks	2000 /month + Incentives
9	Bhigusa Health Care	Product Management	Immediately	2 Months	10000 /month

Exporting into .csv file.

In [24]:

```
df.to_csv('internshala.csv', index = False)
```

The End.