Web Scrapping - Case_Internshala.

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

- GITAM Institute of Science.

Imporing the necessary libraries.

In [1]:
```
import requests
from bs4 import BeautifulSoup as BS
```

Fetching the next page of internashala search re

In [2]:
```
url = 'https://internshala.com/internships/work-from-home-data%20science-jobs/pag
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 [3]:
```
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="Machine Learning">
<a href="/internship/detail/machine-learning-work-from-home-job-internship-at-uniconverge-technologies-
Learning</a> </h4>
<h4>
<a class="link_display_like_text" href="/internships/internship-at-UniConverge%20Technologies%20Private
nologies Private Limited">
                        UniConverge Technologies Private Limited                    </a>
</h4>
</div>
----------------------------------------
Machine Learning
----------------------------------------
```

Getting the Fields list.

In [4]:
```
fields_list = [company_class[ii].h4.a.text for ii in range(len(company_class))]
print(len(fields_list))
print('----------------------')
print(fields_list)
```

```
14
----------------------
['Machine Learning', 'Natural Language Processing', 'Blockchain Analytics', 'Data Science - Content Dev
lution Neural Network', 'Actuary Or Actuarial Intern', 'Deep Learning', 'Data Science', 'Machine Learni
Analytics', 'Deep Learning', 'Financial Analysts (Intraday Trading)', 'Fund Analytics']
```

Getting the company list.

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

```
['\n                        UniConverge Technologies Private Limited                    ',
 '\n                        Arihant Patawari                    ',
 '\n                        BuyUcoin                    ',
 '\n                        DPhi                    ',
 '\n                        The Shaadi Times                    ']
```

Cleaning process.

In [6]:
```
for ii in range(len(company_list)):
    company_list[ii] = company_list[ii].replace('\n','')
company_list[:5]
```

```
['                        UniConverge Technologies Private Limited                    ',
 '                        Arihant Patawari                    ',
 '                        BuyUcoin                    ',
 '                        DPhi                    ',
 '                        The Shaadi Times                    ']
```

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

In [7]:
```
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]
```

```
['UniConverge Technologies Private Limited',
 'Arihant Patawari',
 'BuyUcoin',
 'DPhi',
 'The Shaadi Times']
```

Testing the above results in a dataframe.

In [8]:
```
import pandas as pd
pd.DataFrame({'Companies':company_list,
              'Fields': fields_list})
```

	Companies	Fields
0	UniConverge Technologies Private Limited	Machine Learning
1	Arihant Patawari	Natural Language Processing
2	BuyUcoin	Blockchain Analytics
3	DPhi	Data Science - Content Development
4	The Shaadi Times	Data Analytics
5	Softsensor.ai	Convolution Neural Network
6	Get RIA	Actuary Or Actuarial Intern
7	Medi-Caps University	Deep Learning
8	Kapwise Technologies	Data Science
9	Mavenai Technologies LTD	Machine Learning (AWS)
10	HeyCloudy	Web Scrapping And Data Analytics
11	Pucho Technology Information Private Limited	Deep Learning
12	Trader For Tomorrow	Financial Analysts (Intraday Trading)
13	Sapne	Fund Analytics

Getting the internship description.

In [9]:
```
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>
                                1 Month                                </td>
<td class="stipend_container_table_cell">
<i class="fa fa-inr"></i>2000 /month
                            </td>
<td>24 Mar'20</td>
<td>21 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">Immediately</div>
</td>
<td>
                                2 Months                                </td>
<td class="stipend_container_table_cell">
<i class="fa fa-inr"></i>5000 /month
                            </td>
```

```
<td>23 Mar'20</td>
<td>21 Apr'20</td>
</tr>
</tbody>
</table>
</div>
```

Putting all the data into lists.

In [12]:
```
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 [13]:
```
start_date_list[:5]
```

```
['\nImmediately\n',
 '\nImmediately\n',
 '\nImmediately\n',
 '\nImmediately\n',
 '\nImmediately\n']
```

In [14]:
```
for ii in range(len(start_date_list)):
    start_date_list[ii] = start_date_list[ii].replace('\n','')
start_date_list[:10]
```

```
['Immediately',
 'Immediately',
 'Immediately',
 'Immediately',
 'Immediately',
 'Immediately',
 'Immediately',
 'Immediately',
 'Immediately',
 'Immediately']
```

Cleaning duration_list.

In [15]:
```
duration_list[:5]
```

```
['\n                                1 Month                              ',
 '\n                                2 Months                              ',
 '\n                                3 Months                              ',
 '\n                                3 Months                              ',
 '\n                                1 Month                             ']
```

In [16]:
```
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]
```

```
['1 Month', '2 Months', '3 Months', '3 Months', '1 Month']
```

Cleaning stipend_list.

In [17]:
```
stipend_list[:5]
```

```
['\n2000 /month\n                          ',
 '\n5000 /month\n                          ',
 '\n6000-10000 /month\n                          ',
 '\n6000-9000 /month\n                          ',
 '\n1000 /month\n                          ']
```

In [18]:
```
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 /month',
 '5000 /month',
 '6000-10000 /month',
 '6000-9000 /month',
 '1000 /month']
```

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

In [19]:
```
import pandas as pd
import os
os.chdir(r'C:\Users\acer\Desktop\PythonProgramming')
```

In [20]:
```
df1 = pd.DataFrame()

df1['Companies'] = company_list
df1['Fields'] =  fields_list
df1['Start Date'] = start_date_list
df1['Duration'] = duration_list
df1['Stipend'] = stipend_list
df1['Posted On'] = posted_on_list
df1['Apply By'] = apply_by_list
df1.head(10)
```

	Companies	Fields	Start Date	Duration	
0	UniConverge Technologies Private Limited	Machine Learning	Immediately	1 Month	2000 /m
1	Arihant Patawari	Natural Language Processing	Immediately	2 Months	5000 /m
2	BuyUcoin	Blockchain Analytics	Immediately	3 Months	6000-10 /month
3	DPhi	Data Science - Content Development	Immediately	3 Months	6000-90
4	The Shaadi Times	Data Analytics	Immediately	1 Month	1000 /m
5	Softsensor.ai	Convolution Neural Network	Immediately	6 Months	Unpaid
6	Get RIA	Actuary Or Actuarial Intern	Immediately	4 Months	5000-70
7	Medi-Caps University	Deep Learning	Immediately	1 Month	10000-2 /month
8	Kapwise Technologies	Data Science	Immediately	1 Month	1000-50
9	Mavenai Technologies LTD	Machine Learning (AWS)	Immediately	2 Months	8000 /m

Exporting into .csv file.

In [22]:
```
df1.to_csv('internshala1.csv', index = False)
```

The End.