Web Scrapping - From Craig's List (Udemy).

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

- GITAM Institute of Science.

Importing necessary libraries.

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

Finding the webpage document.

In [2]:
```

url = "https://boston.craigslist.org/search/sof"

```

In [3]:
```
page = requests.get(url)
```

In [6]:
```
soup = BS(page.text, 'html.parser')
```

Finding all the href (URL EXTRACTION).

In [13]:
```
links = soup.find_all('a')
```

In [17]:
```
for ii in links:
    print(ii.get('href'))
    print('--------------')
```

```
--------------
https://boston.craigslist.org/gbs/sof/d/boston-senior-cloud-devops-engineer/7
097886190.html (https://boston.craigslist.org/gbs/sof/d/boston-senior-cloud-d
evops-engineer/7097886190.html)
--------------
https://boston.craigslist.org/gbs/sof/d/boston-senior-cloud-devops-engineer/7
097886190.html (https://boston.craigslist.org/gbs/sof/d/boston-senior-cloud-d
evops-engineer/7097886190.html)
--------------
#
--------------
https://boston.craigslist.org/gbs/sof/d/boston-data-science-fellowship/709778
3859.html (https://boston.craigslist.org/gbs/sof/d/boston-data-science-fellow
ship/7097783859.html)
--------------
https://boston.craigslist.org/gbs/sof/d/boston-data-science-fellowship/709778
3859.html (https://boston.craigslist.org/gbs/sof/d/boston-data-science-fellow
ship/7097783859.html)
--------------
#
```

Finding the title names.

In [9]:
```
title = soup.find_all('a', {'class':'result-title hdrlnk'})
```

In [10]:
```
for ii in title:
    print(ii.text)
```

```
Data Science Fellowship
Software Developer for new startup (React, Python)
Sr. Software Engineer - Join a THRIVING Company actively recruiting!!
Data Engineer Manager- Join a THRIVING Company actively recruiting!!
Data Science Fellowship
Data Science Fellowship
Co-founder wanted for tech business startup
Cloud or mobile engineer, side gig, major equity, proven team
Looking for a result-oriented IT sales person
Senior Cloud/DevOps Engineer
Data Science Fellowship
MS Dynamics CRM Developer
Software Engineer
Looking For Cobol Programmer
Java PL/SQL Developer
```

Finding the addresses.

In [11]:
```
address = soup.find_all(class_ = 'result-hood')
```

In [12]:

```
for ii in address:
    print(ii.text)
```

```
(Beverly)
(Jupiter)
(Jupiter)
(South Shore / Cape)
(Remote)
(Northborough)
(boston: boston/cambridge/brookline)
(boston: boston/cambridge/brookline)
(Natick)
```

Extracting the title, location, link and date in groups.

In [53]:

```
from numpy import nan
```

In [18]:

```
jobs = soup.find_all('p', class_ = 'result-info')
```

In [63]:

```python
for ii in range(len(jobs)):
    titles = jobs[ii].find('a', {'class':'result-title hdrlnk'})

    locationFind = jobs[ii].find('span', class_ = 'result-hood')
    if locationFind:
        locations = str(locationFind.text)[2:-1]
    else:
        locations = nan

    links = jobs[ii].find('a')
    dates = jobs[ii].find('time', class_ = 'result-date')


    job_page = requests.get( str(links.get('href')) )
    job_soup = BS(job_page.text, 'html.parser')

    description = job_soup.find('section', id = 'postingbody')

    print('Title : ', titles.text,
          '\nLocation : ', locations,
          '\nLink : ', links.get('href'),
          '\nDate : ', str(dates.get('title'))[:11],
          '\nDescription : \n', description.text,
         )

    print('---------------')
```

```
Title :  Data Science Fellowship
Location :  nan
Link :  https://boston.craigslist.org/gbs/sof/d/boston-data-science-fellowshi
p/7108163205.html (https://boston.craigslist.org/gbs/sof/d/boston-data-scienc
e-fellowship/7108163205.html)
Date :  Tue 14 Apr
Description :


QR Code Link to This Post


What is Pathrise

Pathrise (YC W18) invests in undervalued university students or young profess
ionals by coaching them to get a competitive job. The program is completely f
ree upfront. In exchange, Pathrise fellows agree to pay back a share of their
first year's salary if and only if they get hired.
```

Exporting the above info to a .csv file.

First, we read the content into a dataframe using pandas.

In [64]:

```
import pandas as pd
```

In [65]:

```
Titles = []
Location = []
Link = []
Date = []
Description = []
```

In [66]:

```
for ii in range(len(jobs)):
    titles = jobs[ii].find('a', {'class':'result-title hdrlnk'})

    locationFind = jobs[ii].find('span', class_ = 'result-hood')
    if locationFind:
        locations = str(locationFind.text)[2:-1]
    else:
        locations = nan

    links = jobs[ii].find('a')
    dates = jobs[ii].find('time', class_ = 'result-date')


    job_page = requests.get( str(links.get('href')) )
    job_soup = BS(job_page.text, 'html.parser')

    description = job_soup.find('section', id = 'postingbody')

    Titles.append(titles.text)
    Location.append(locations)
    Link.append(links.get('href'))
    Date.append(str(dates.get('title'))[:11])
    Description.append(description.text)
```

In [72]:

```
df = pd.DataFrame({
    'Titles' : Titles,
    'Location' : Location,
    'Link' : Link,
    'Date' : Date,
    'Description' : Description
})

df.head()
```

Out[72]:

	Titles	Location	Link	Date	Description
0	Data Science Fellowship	NaN	https://boston.craigslist.org/gbs/sof/d/boston...	Tue 14 Apr	\n\nQR Code Link to This Post\n\n\nWhat is Pat...
1	Software Developer for new startup (React, Pyt...	Beverly	https://boston.craigslist.org/nos/sof/d/beverl...	Thu 09 Apr	\n\nQR Code Link to This Post\n\n\nSoftware De...
2	Sr. Software Engineer - Join a THRIVING Compan...	Jupiter	https://boston.craigslist.org/gbs/sof/d/jupite...	Thu 09 Apr	\n\nQR Code Link to This Post\n\n\nAre you a S...
3	Data Engineer Manager- Join a THRIVING Company...	Jupiter	https://boston.craigslist.org/gbs/sof/d/jupite...	Thu 09 Apr	\n\nQR Code Link to This Post\n\n\nAre you a D...
4	Data Science Fellowship	NaN	https://boston.craigslist.org/gbs/sof/d/boston...	Mon 06 Apr	\n\nQR Code Link to This Post\n\n\nWhat is Pat...

In [70]:

```
import os
os.chdir(r'C:\Users\Habeeb\Documents\Sagun\Python\csv files')
```

In [73]:

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