Hands-on Lab : Web Scraping

Estimated time needed: **30 to 45** minutes

Objectives

In this lab you will perform the following:

- Extract information from a given web site
- Write the scraped data into a csv file.

Extract information from the given web site

You will extract the data from the below web site:

In [1]:
```
#this url contains the data you need to scrape
url = "https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DA0321EN
```

The data you need to scrape is the **name of the programming language** and **average annual salary**.
It is a good idea to open the url in your web broswer and study the contents of the web page before you start to scrape.

Import the required libraries

In [2]:
```
# Your code here
from bs4 import BeautifulSoup
import requests
import pandas as pd
```

```
C:\Users\Administrator\anaconda3\Lib\site-packages\pandas\core\arrays\masked.py:60: U
serWarning: Pandas requires version '1.3.6' or newer of 'bottleneck' (version '1.3.5'
currently installed).
  from pandas.core import (
```

Download the webpage at the url

In [3]:
```
#your code goes here
data = requests.get(url).text
```

Create a soup object

In [4]:
```
#your code goes here
soup = BeautifulSoup(data)
```

Scrape the `Language name` and `annual average salary`.

In [10]:
```
#your code goes here
table = soup.find('table')
Lang = []
Avg_salary = []
for row in table.find_all('tr'):
    col = row.find_all('td')
    Language = col[1].getText()
    Annual_Salary = col[3].getText()
    Lang.append(Language)
    Avg_salary.append(Annual_Salary)
    print(f'{Language}-----> {Annual_Salary}')
```

```
Language-----> Average Annual Salary
Python-----> $114,383
Java-----> $101,013
R-----> $92,037
Javascript-----> $110,981
Swift-----> $130,801
C++-----> $113,865
C#-----> $88,726
PHP-----> $84,727
SQL-----> $84,793
Go-----> $94,082
```

In [ ]:
```
```

In [11]:
```
Lang
```

Out[11]:
```
['Language',
 'Python',
 'Java',
 'R',
 'Javascript',
 'Swift',
 'C++',
 'C#',
 'PHP',
 'SQL',
 'Go']
```

In [53]:
```
df = pd.DataFrame(Lang, Avg_salary)
df
```

Out[53]:

	0
Average Annual Salary	Language
$114,383	Python
$101,013	Java
$92,037	R
$110,981	Javascript
$130,801	Swift
$113,865	C++
$88,726	C#
$84,727	PHP
$84,793	SQL
$94,082	Go

In [54]:
```
df.columns = df.iloc[0]
df
```

Out[54]:

Average Annual Salary	Language
Average Annual Salary	Language
$114,383	Python
$101,013	Java
$92,037	R
$110,981	Javascript
$130,801	Swift
$113,865	C++
$88,726	C#
$84,727	PHP
$84,793	SQL
$94,082	Go

In [56]:
```
row_1 = df.iloc[0]
row_1
df2 = df.drop(row_1)
```

```
---------------------------------------------------------------------------
KeyError                                  Traceback (most recent call last)
Cell In[56], line 3
      1 row_1 = df.iloc[0]
      2 row_1
----> 3 df2 = df.drop(row_1)

File ~\anaconda3\Lib\site-packages\pandas\core\frame.py:5568, in DataFrame.drop(self,
labels, axis, index, columns, level, inplace, errors)
   5420 def drop(
   5421     self,
   5422     labels: IndexLabel | None = None,
   (...)
   5429     errors: IgnoreRaise = "raise",
   5430 ) -> DataFrame | None:
   5431     """
   5432     Drop specified labels from rows or columns.
   5433
   (...)
   5566             weight  1.0     0.8
   5567     """
-> 5568     return super().drop(
   5569         labels=labels,
   5570         axis=axis,
   5571         index=index,
   5572         columns=columns,
   5573         level=level,
   5574         inplace=inplace,
   5575         errors=errors,
   5576     )

File ~\anaconda3\Lib\site-packages\pandas\core\generic.py:4782, in NDFrame.drop(self,
labels, axis, index, columns, level, inplace, errors)
   4780 for axis, labels in axes.items():
   4781     if labels is not None:
-> 4782         obj = obj._drop_axis(labels, axis, level=level, errors=errors)
   4784 if inplace:
   4785     self._update_inplace(obj)

File ~\anaconda3\Lib\site-packages\pandas\core\generic.py:4824, in NDFrame._drop_axis
(self, labels, axis, level, errors, only_slice)
   4822         new_axis = axis.drop(labels, level=level, errors=errors)
   4823     else:
-> 4824         new_axis = axis.drop(labels, errors=errors)
   4825     indexer = axis.get_indexer(new_axis)
   4827 # Case for non-unique axis
   4828 else:

File ~\anaconda3\Lib\site-packages\pandas\core\indexes\base.py:7069, in Index.drop(se
lf, labels, errors)
   7067 if mask.any():
   7068     if errors != "ignore":
-> 7069         raise KeyError(f"{labels[mask].tolist()} not found in axis")
   7070     indexer = indexer[~mask]
   7071 return self.delete(indexer)

KeyError: "['Language'] not found in axis"
```

In [ ]:

In [ ]:

In [ ]:

Save the scrapped data into a file named *popular-languages.csv*

In [ ]:
```
# your code goes here
```

Authors

Ramesh Sannareddy

Other Contributors

Rav Ahuja

Change Log

Date (YYYY-MM-DD)	Version	Changed By	Change Description
2020-10-17	0.1	Ramesh Sannareddy	Created initial version of the lab

Copyright © 2020 IBM Corporation. This notebook and its source code are released under the terms of the MIT License.