Web Scraping Project by Kate Rogatina

Objectives

In this lab we 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

We will extract the data from the below web site:

```
#this url contains the data you need to scrape
url = "https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DA0321EN-SkillsNetwork/labs/datasets/Programming_Languages.html"
```

The data we 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 browser and study the contents of the web page before you start to scrape.
Import the required libraries

```
# Your code here
from bs4 import BeautifulSoup # this module helps in web scrapping.
import requests  # this module helps us to download a web page
```

Download the webpage at the url

```
#your code goes here
data  = requests.get(url).text
```

Create a soup object

```
#your code goes here
%pip install html5lib
soup = BeautifulSoup(data,"html5lib")  # create a soup
object using the variable 'data'
```

```
#your code goes here
%pip install html5lib
soup = BeautifulSoup(data,"html5lib")  # create a soup
object using the variable 'data'
```

```
Requirement already satisfied: html5lib in /opt/conda/
envs/Python-RT23.1/lib/python3.10/site-packages (1.1)
Requirement already satisfied: six>=1.9 in /opt/conda/
envs/Python-RT23.1/lib/python3.10/site-packages (from
html5lib) (1.16.0)
Requirement already satisfied: webencodings in /opt/
conda/envs/Python-RT23.1/lib/python3.10/site-packages
(from html5lib) (0.5.1)
Note: you may need to restart the kernel to use updated
packages.
```

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

```
#find a html table in the web page
table = soup.find('table') # in html table is
represented by the tag <table>
# your code goes here
#Get all rows from the table
for row in table.find_all('Language'): # in html table
row is represented by the tag <tr>
cols = row.find_all('td') # in html a column is
represented by the tag <td>
color_name = cols[2].getText() # store the value in
column 3 as color_name
color_code = cols[3].getText() # store the value in
column 4 as color_code
```

```
print("{}--->{}".format(color_name,color_code))

table = soup.find("table")

# Initialize lists to store language names and salaries
languages = []
salaries = []

# Loop through rows in the table (skipping the header
row)
for row in table.find_all("tr")[1:]:
    # Extract language name and salary from each row
    language = row.find_all("td")[1].text
    salary = row.find_all("td")[3].text

    # Append data to lists
    languages.append(language)
    salaries.append(salary)

# Print the scraped data
for language, salary in zip(languages, salaries):
    print(f"Language: {language}, Average Annual
Salary: {salary}")
```

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

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

```
import csv
from bs4 import BeautifulSoup
```

```
data = list(zip(languages, salaries))

# Save data to a CSV file
csv_file_path = "popular-languages.csv"
with open(csv_file_path, mode="w", newline="", 
encoding="utf-8") as csv_file:
    csv_writer = csv.writer(csv_file)

    # Write header
    csv_writer.writerow(["Language", "Average Annual 
Salary"])

    # Write data rows
    csv_writer.writerows(data)

print(f"Data has been successfully saved to 
{csv_file_path}")
```

```
Data has been successfully saved to popular-
languages.csv
```