Skills
Network

Web Scraping Lab

Estimated time needed: **30** minutes

Objectives

After completing this lab you will be able to:

- Download a webpage using requests module
- Scrape all links from a web page
- Scrape all image urls from a web page
- Scrape data from html tables

Scrape www.ibm.com

Import the required modules and functions

In [1]:
```
from bs4 import BeautifulSoup # this module helps in web scrapping.
import requests  # this module helps us to download a web page
```

Download the contents of the web page

In [2]:
```
url = "http://www.ibm.com"
```

In [3]:
```
# get the contents of the webpage in text format and store in a variable called data
data  = requests.get(url).text
```

Create a soup object using the class BeautifulSoup

In [4]:
```
soup = BeautifulSoup(data,"html.parser")  # create a soup object using the variable 'data'
```

Scrape all links

In [5]:
```
for link in soup.find_all('a'):  # in html anchor/link is represented by the tag <a>
    print(link.get('href'))
```

```
https://www.ibm.com/cloud?lnk=hpUSbt1
```

Scrape all images

In [6]:
```
for link in soup.find_all('img'):# in html image is represented by the tag <img>
    print(link.get('src'))
```

Scrape data from html tables

In [7]:
```
#The below url contains a html table with data about colors and color codes.
```

In [8]:
```
url = "https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DA0321EN-SkillsNetwork/labs/datasets/HTMLColorC
```

Before proceeding to scrape a web site, you need to examine the contents, and the way data is organized on the website. Open the above url in your browser and check how many rows and columns are there in the color table.

In [9]:
```
# get the contents of the webpage in text format and store in a variable called data
data  = requests.get(url).text
```

In [10]:
```
soup = BeautifulSoup(data,"html.parser")
```

In [11]:
```
#find a html table in the web page
table = soup.find('table') # in html table is represented by the tag <table>
```

In [12]:
```
#Get all rows from the table
for row in table.find_all('tr'): # in html table row is represented by the tag <tr>
    # Get all columns in each row.
    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))
```

```
Color Name--->Hex Code#RRGGBB
lightsalmon--->#FFA07A
salmon--->#FA8072
darksalmon--->#E9967A
lightcoral--->#F08080
coral--->#FF7F50
tomato--->#FF6347
orangered--->#FF4500
gold--->#FFD700
orange--->#FFA500
darkorange--->#FF8C00
lightyellow--->#FFFFE0
lemonchiffon--->#FFFACD
papayawhip--->#FFEFD5
moccasin--->#FFE4B5
peachpuff--->#FFDAB9
palegoldenrod--->#EEE8AA
khaki--->#F0E68C
darkkhaki--->#BDB76B
yellow--->#FFFF00
lawngreen--->#7CFC00
chartreuse--->#7FFF00
limegreen--->#32CD32
lime--->#00FF00
forestgreen--->#228B22
green--->#008000
powderblue--->#B0E0E6
lightblue--->#ADD8E6
lightskyblue--->#87CEFA
skyblue--->#87CEEB
deepskyblue--->#00BFFF
lightsteelblue--->#B0C4DE
dodgerblue--->#1E90FF
```

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