Web Scraping

What is Web Scraping?

web scraping is the practice of gathering data through any means other than a program interacting with an API (or, obviously, through a human using a web browser)

Why

web scrapers are excellent at gathering and processing large amounts of data quickly

WARNING

There are legal issues to consider when scraping a website (robot.txt)

In [ ]:
```
from urllib.request import urlopen
from bs4 import BeautifulSoup


html = urlopen('http://www.pythonscraping.com/pages/page1.html')
bs = BeautifulSoup(html.read(), 'html.parser') #/lxml/html5lib
print(bs.h1)
```

In [1]:
```
html_doc = """
       <html lang="en">
        <head>
            <meta charset="UTF-8">
            <meta name="viewport" content="width=device-width, initial-scal
            <meta http-equiv="X-UA-Compatible" content="ie=edge">
            <title>My Story</title>
        </head>
        <body>
        <h1>The story of my life</h1>
        <p class="page-1">
            Lorem ipsum dolor sit amet consectetur adipisicing elit. Repudi
        </p>
        <p class="page-1">
            Color sit amet consectetur adipisicing elit. Repudiandae, earum
        </p>
        <a class="home" href="">my home</a>
        <a class="girl" href="">girlfriend</a>
        <a href="">food</a>
        <a clas="food" href="">pet</a>
        <p class="page-2">Lorem ipsum dolor Dolor, accusantium odio.</p>
        </body>
        </html>
"""
```

In [3]:
```
from bs4 import BeautifulSoup

soup = BeautifulSoup(html_doc, 'lxml')
```

In [37]:
```
soup.find('p')
```

Out[37]:
```
<p class="page-1">
            Lorem ipsum dolor sit amet consectetur adipisicing elit. Repu
diandae, earum?
        </p>
```

In [13]:
```
soup.p
```

Out[13]:
```
<p class="page-1">
            Lorem ipsum dolor sit amet consectetur adipisicing elit. Repu
diandae, earum?
        </p>
```

In [38]:
```
soup.findAll('a') # find all items
```

Out[38]:
```
[<a class="home" href="">my home</a>,
 <a class="girl" href="">girlfriend</a>,
 <a href="">food</a>,
 <a clas="food" href="">pet</a>]
```

In [4]:
```
# finding a attributes
para = soup.find_all('p', class_ = 'page-1')
print(para)
```

```
[<p class="page-1">
            Lorem ipsum dolor sit amet consectetur adipisicing elit. Repu
diandae, earum?
        </p>, <p class="page-1">
            Color sit amet consectetur adipisicing elit. Repudiandae, ear
um?
        </p>]
```

In [46]:
```
a_tag = soup.find('a', class_='girl')
print(a_tag)
```

```
<a class="girl" href="">girlfriend</a>
```

In [ ]:

In [50]:
```
# Searching for all tags

from bs4 import BeautifulSoup

html_doc = """
<html>
    <head>
        <title>The Dormouse's story</title>
    </head>
<body>
    <p class="title">
        <b>The Dormouse's story</b>
    </p>
    <p class="story">Once upon a time there were three little sisters; and
    <a href="http://example.com/elsie" class="sister" id="link1">Elsie</a>,
    <a href="http://example.com/lacie" class="sister" id="link2">Lacie</a>
    <a href="http://example.com/tillie" class="sister" id="link3">Tillie</a
    and they lived at the bottom of a well.</p>
    <p class="story">...</p>
</body>
</html>
"""

soup = BeautifulSoup(html_doc, 'lxml')
```

In [51]:
```
a_tags = soup.find_all('a')

print(a_tags)

len(a_tags)
```

```
[<a class="sister" href="http://example.com/elsie" id="link1">Elsie</a>,
<a class="sister" href="http://example.com/lacie" id="link2">Lacie</a>, <
a class="sister" href="http://example.com/tillie" id="link3">Tillie</a>]
```

Out[51]: 3

In [ ]:

In [52]:
```
# Search with tag names and other attributes

from bs4 import BeautifulSoup

html_doc = """
<html>
    <head>
        <title>The Dormouse's story</title>
    </head>
<body>
    <p class="title">
        <b>The Dormouse's story</b>
    </p>
    <p class="story">Once upon a time there were three little sisters; and
    <a href="http://example.com/elsie" class="sister" id="link1">Elsie</a>,
    <a href="http://example.com/lacie" class="sister" id="link2">Lacie</a>
    <a href="http://example.com/tillie" class="sister" id="link3">Tillie</a
    and they lived at the bottom of a well.</p>
    <p class="story">...</p>
</body>
</html>
"""
```

In [54]:
```
soup = BeautifulSoup(html_doc, 'lxml')

a = soup.find_all('a', {'id':'link1'})

print(a)
```

```
[<a class="sister" href="http://example.com/elsie" id="link1">Elsie</a>]
```

In [ ]:
```

```

In [56]:
```
# Search with tag name and strings

from bs4 import BeautifulSoup

html_doc = """
<html>
    <head>
        <title>The Dormouse's story</title>
    </head>
<body>
    <p class="title">
        <b>The Dormouse's story</b>
    </p>
    <p class="story">Once upon a time there were three little sisters; and
    <a href="http://example.com/elsie" class="sister" id="link1">Elsie</a>,
    <a href="http://example.com/lacie" class="sister" id="link2">Lacie</a>
    <a href="http://example.com/tillie" class="sister" id="link3">Tillie</a
    and they lived at the bottom of a well.</p>
    <p class="story">...</p>
</body>
</html>
"""
```

In [57]:
```
soup = BeautifulSoup(html_doc, 'lxml')

a_elsie = soup.find_all('a', string = 'Elsie')

print(a_elsie)
```

```
[<a class="sister" href="http://example.com/elsie" id="link1">Elsie</a>]
```

In [ ]:

In [ ]:
```
# DSearch Parent, Child and Siblings

from bs4 import BeautifulSoup

html_doc = """
<html>
    <head>
        <title>The Dormouse's story</title>
    </head>
<body>

    <p class="title">
        <b>The Dormouse's story</b>
    </p>

    <p class="story">Once upon a time there were three little sisters; and
        <a href="http://example.com/elsie" class="sister" id="link1">Elsie<
        <a href="http://example.com/lacie" class="sister" id="link2">Lacie<
        <a href="http://example.com/tillie" class="sister" id="link3">Tilli
    and they lived at the bottom of a well.
    </p>

    <p class="story">...</p>

</body>
</html>
"""
```

In [59]:
```
soup = BeautifulSoup(html_doc, 'lxml')


# search for all child
p = soup.find('p', class_ ='story')
all_p_children = p.findChildren()
print(all_p_children)
```

```
[<a class="sister" href="http://example.com/elsie" id="link1">Elsie</a>,
<a class="sister" href="http://example.com/lacie" id="link2">Lacie</a>, <
a class="sister" href="http://example.com/tillie" id="link3">Tillie</a>]
```

In [ ]:
```

```

In [60]:
```
# Search scope in BeatifulSoup object
from bs4 import BeautifulSoup

html_doc = """
<html>
    <head>
        <title>The Dormouse's story</title>
    </head>
<body>

    <p class="title">
        <b>The Dormouse's story</b>
    </p>

    <p class="story">Once upon a time there were three little sisters; and
        <a href="http://example.com/elsie" class="sister" id="link1">Elsie<
        <a href="http://example.com/lacie" class="sister" id="link2">Lacie<
        <a href="http://example.com/tillie" class="sister" id="link3">Tilli
    and they lived at the bottom of a well.
    </p>

    <p class="story">...</p>

</body>
</html>
"""
```

In [62]:
```
soup = BeautifulSoup(html_doc, 'lxml')

first_p = soup.find('p')

print(first_p.find('a'))

print(first_p.find('b'))
```

```
None
<b>The Dormouse's story</b>
```

In [ ]:
```

```

In [63]:
```
# Scraping text content

from bs4 import BeautifulSoup

html_doc = """
<html>
    <head>
        <title>The Dormouse's story</title>
    </head>
<body>

    <p class="title">
        <b>The Dormouse's story</b>
    </p>

    <p class="story">Once upon a time there were three little sisters; and
        <a href="http://example.com/elsie" class="sister" id="link1">Elsie<
        <a href="http://example.com/lacie" class="sister" id="link2">Lacie<
        <a href="http://example.com/tillie" class="sister" id="link3">Tilli
    and they lived at the bottom of a well.
    </p>

    <p class="story">...</p>

</body>
</html>
"""

soup = BeautifulSoup(html_doc, 'lxml')

p = soup.find('p')

print(p.text)

a =(soup.find('a'))

print(a.text)
```

```
The Dormouse's story

Elsie
```

In [ ]:
```

```

In [64]:
```
# Scrape for links
from bs4 import BeautifulSoup

html_doc = """
<html>
    <head>
        <title>The Dormouse's story</title>
    </head>
<body>

    <p class="title">
        <b>The Dormouse's story</b>
    </p>

    <p class="story">Once upon a time there were three little sisters; and
        <a href="http://example.com/elsie" class="sister" id="link1">Elsie<
        <a href="http://example.com/lacie" class="sister" id="link2">Lacie<
        <a href="http://example.com/tillie" class="sister" id="link3">Tilli
    and they lived at the bottom of a well.
    </p>

    <p class="story">...</p>

</body>
</html>
"""

soup = BeautifulSoup(html_doc, 'lxml')

a_tags = soup.find_all('a')

for a in a_tags:
    print(a['href'])
```

```
http://example.com/elsie (http://example.com/elsie)
http://example.com/lacie (http://example.com/lacie)
http://example.com/tillie (http://example.com/tillie)
```

In [ ]:

In [67]:
```
# Scrape Data inside Tables

from bs4 import BeautifulSoup

soup = BeautifulSoup(open('sample.html'), 'lxml')

# print soup.prettify()

for tr in soup.find_all('tr'):
    for td in tr.find_all('td'):
        print(td.text)
```

```
Chicken noodle soup
120
2
Caesar salad
400
26
```

In [ ]:
```

```

Project 1 Assigment

Scrape nba.com (https://nba.com)

- **Get the list of all players**
- **Get the list of all couches**

In [ ]:
```
from urllib.request import urlopen
from bs4 import BeautifulSoup


html = urlopen('http://www.nba.com/')
soup = BeautifulSoup(html.read(), 'lxml') #/lxml/html5lib
print(soup.h1)
```

In [ ]:
```
# All players
```

In [ ]:
```
# all couches
section = soup.find('section',{'id':'nbaArticleContent'})

for p in section.find_all('p'):
    for a in p.find_all('a'):
        # print a.text + "is coach of " + a.find_previous_sibling().text.re
        print(a.text + " ---> " + a.find_previous_sibling().text.replace(':
```