In [1]:
```
!pip install bs4
```

```
Collecting bs4
  Downloading https://files.pythonhosted.org/packages/10/ed/7e8b97591f6
f456174139ec089c769f89a94a1a4025fe967691de971f314/bs4-0.0.1.tar.gz
Collecting beautifulsoup4 (from bs4)
  Downloading https://files.pythonhosted.org/packages/d1/41/e6495bd7d37
81cee623ce23ea6ac73282a373088fcd0ddc809a047b18eae/beautifulsoup4-4.9.3-
py3-none-any.whl (115kB)
     |████████████████████████████████| 122kB 987kB/s eta 0:00:01
Collecting soupsieve>1.2; python_version >= "3.0" (from beautifulsoup4-
>bs4)
  Downloading https://files.pythonhosted.org/packages/36/69/d82d04022f0
2733bf9a72bc3b96332d360c0c5307096d76f6bb7489f7e57/soupsieve-2.2.1-py3-n
one-any.whl
Building wheels for collected packages: bs4
  Building wheel for bs4 (setup.py) ... done
  Stored in directory: /home/jupyterlab/.cache/pip/wheels/a0/b0/b2/4f80
b9456b87abedbc0bf2d52235414c3467d8889be38dd472
Successfully built bs4
Installing collected packages: soupsieve, beautifulsoup4, bs4
Successfully installed beautifulsoup4-4.9.3 bs4-0.0.1 soupsieve-2.2.1
```

In [2]:
```
from bs4 import BeautifulSoup # this module helps in web scrapping.
import requests
```

In [3]:
```
%%html
<!DOCTYPE html>
<html>
<head>
<title>Page Title</title>
</head>
<body>
<h3><b id='boldest'>Lebron James</b></h3>
<p> Salary: $ 92,000,000 </p>
```

```
<h3> Stephen Curry</h3>
<p> Salary: $85,000, 000 </p>
<h3> Kevin Durant </h3>
<p> Salary: $73,200, 000</p>
</body>
</html>
```

Lebron James

Salary: $ 92,000,000

Stephen Curry

Salary: $85,000, 000

Kevin Durant

Salary: $73,200, 000

In [4]:
```
html="<!DOCTYPE html><html><head><title>Page Title</title></head><body>
<h3><b id='boldest'>Lebron James</b></h3><p> Salary: $ 92,000,000 </p><
h3> Stephen Curry</h3><p> Salary: $85,000, 000 </p><h3> Kevin Durant </
h3><p> Salary: $73,200, 000</p></body></html>"
```

In [5]:
```
soup = BeautifulSoup(html, 'html5lib')
```

In [6]:
```
print(soup.prettify())
```

```
<!DOCTYPE html>
<html>
 <head>
  <title>
   Page Title
  </title>
 </head>
 <body>
```

```
  <h3>
   <b id="boldest">
    Lebron James
   </b>
  </h3>
  <p>
   Salary: $ 92,000,000
  </p>
  <h3>
   Stephen Curry
  </h3>
  <p>
   Salary: $85,000, 000
  </p>
  <h3>
   Kevin Durant
  </h3>
  <p>
   Salary: $73,200, 000
  </p>
 </body>
</html>
```

In [7]:
```
tag_object=soup.title
print("tag object:",tag_object)
```

```
tag object: <title>Page Title</title>
```

In [8]:
```
print("tag object type:",type(tag_object))
```

```
tag object type: <class 'bs4.element.Tag'>
```

In [9]:
```
tag_object=soup.h3
tag_object
```

Out[9]:
```
<h3><b id="boldest">Lebron James</b></h3>
```

In [10]:
```
tag_child =tag_object.b
```

```
tag_child
```

Out[10]:
```
<b id="boldest">Lebron James</b>
```

In [11]:
```
parent_tag=tag_child.parent
parent_tag
```

Out[11]:
```
<h3><b id="boldest">Lebron James</b></h3>
```

In [12]:
```
tag_object
```

Out[12]:
```
<h3><b id="boldest">Lebron James</b></h3>
```

In [13]:
```
tag_object.parent
```

Out[13]:
```
<body><h3><b id="boldest">Lebron James</b></h3><p> Salary: $ 92,000,000
</p><h3> Stephen Curry</h3><p> Salary: $85,000, 000 </p><h3> Kevin Dura
nt </h3><p> Salary: $73,200, 000</p></body>
```

In [14]:
```
sibling_1=tag_object.next_sibling
sibling_1
```

Out[14]:
```
<p> Salary: $ 92,000,000 </p>
```

In [15]:
```
sibling_2=sibling_1.next_sibling
sibling_2
```

Out[15]:
```
<h3> Stephen Curry</h3>
```

In [16]:
```
sibling_3=sibling_2.next_sibling
sibling_3
```

Out[16]:
```
<p> Salary: $85,000, 000 </p>
```

In [17]:
```
tag_child['id']
```

Out[17]:
```
'boldest'
```

In [18]:
```
tag_child.get('id')
```

Out[18]:
```
'boldest'
```

In [19]:
```
tag_string=tag_child.string
tag_string
```

Out[19]:
```
'Lebron James'
```

In [20]:
```
type(tag_string)
```

Out[20]:
```
bs4.element.NavigableString
```

In [21]:
```
unicode_string = str(tag_string)
unicode_string
```

Out[21]:
```
'Lebron James'
```

In [22]:
```
%%html
<table>
  <tr>
    <td id='flight' >Flight No</td>
    <td>Launch site</td>
    <td>Payload mass</td>
   </tr>
  <tr>
    <td>1</td>
    <td><a href='https://en.wikipedia.org/wiki/Florida'>Florida</a></td
>
    <td>300 kg</td>
  </tr>
  <tr>
    <td>2</td>
    <td><a href='https://en.wikipedia.org/wiki/Texas'>Texas</a></td>
    <td>94 kg</td>
  </tr>
  <tr>
    <td>3</td>
```

```
    <td><a href='https://en.wikipedia.org/wiki/Florida'>Florida<a> </td
>
    <td>80 kg</td>
  </tr>
</table>
```

Flight No	Launch site	Payload mass
1	Florida	300 kg
2	Texas	94 kg
3	Florida	80 kg

In [26]:
```
table="<table><tr><td id='flight'>Flight No</td><td>Launch site</td> <t
d>Payload mass</td></tr><tr> <td>1</td><td><a href='https://en.wikipedi
a.org/wiki/Florida'>Florida<a></td><td>300 kg</td></tr><tr><td>2</td><t
d><a href='https://en.wikipedia.org/wiki/Texas'>Texas</a></td><td>94 kg
</td></tr><tr><td>3</td><td><a href='https://en.wikipedia.org/wiki/Flor
ida'>Florida<a> </td><td>80 kg</td></tr></table>"
```

In [27]:
```
table_rows=table_bs.find_all('tr')
table_rows
```

```
---------------------------------------------------------------------------
----
NameError                                 Traceback (most recent call l
ast)
<ipython-input-27-608ac036c0fc> in <module>
----> 1 table_rows=table_bs.find_all('tr')
      2 table_rows

NameError: name 'table_bs' is not defined
```

In [28]:
```
table_bs.find_all(id="flight")
```

```
---------------------------------------------------------------------------
----
NameError                                 Traceback (most recent call l
```

```
ast)
<ipython-input-28-c3bb6531bac0> in <module>
----> 1 table_bs.find_all(id="flight")

NameError: name 'table_bs' is not defined
```

In [29]:
```
url = "http://www.ibm.com"
```

In [30]:
```
data  = requests.get(url).text
```

In [31]:
```
soup = BeautifulSoup(data,"html5lib")  # create a soup object using the variable 'data'
```

In [32]:
```
for link in soup.find_all('a',href=True):

    print(link.get('href'))
```

```
#main-content
http://www.ibm.com/
https://www.ibm.com/cloud/automation/mayflower-autonomous-ship?lnk=ushpv18l1
https://www.ibm.com/cloud/hybrid/value-calculator/?lnk=ushpv18f1
https://www.ibm.com/cloud/websphere-hybrid-edition?lnk=ushpv18f2
https://www.ibm.com/blogs/journey-to-ai/2021/04/extended-planning-and-analysis-xpa/?lnk=ushpv18f3
https://www.ibm.com/watson/trustworthy-ai?lnk=ushpv18f4
https://www.ibm.com/products/offers-and-discounts?link=ushpv18t5&lnk2=trial_mktpl_MPDISC
https://www.ibm.com/cloud/free?lnk=ushpv18t1&lnk2=trial_Cloud&psrc=none&pexp=def
https://www.ibm.com/products/cognos-analytics?lnk=ushpv18t2&lnk2=trial_CogAnalytics&psrc=none&pexp=def
https://www.ibm.com/cloud/watson-assistant?lnk=ushpv18t3&lnk2=trial_WatAssist&psrc=none&pexp=def
https://www.ibm.com/products/digital-learning-subscription/pricing?lnk=ushpv18t4&lnk2=trial_DigLearning&psrc=none&pexp=def
https://www.ibm.com/search?lnk=ushpv18srch&locale=en-us&q=
https://www.ibm.com/products?lnk=ushpv18p1&lnk2=trial_mktpl&psrc=none&p
```

```
exp=def
https://developer.ibm.com/depmodels/cloud/?lnk=ushpv18ct16
https://developer.ibm.com/technologies/artificial-intelligence?lnk=ushp
v18ct19
https://www.ibm.com/demos/?lnk=ushpv18ct12
https://developer.ibm.com/?lnk=ushpv18ct9
https://www.ibm.com/docs/en?lnk=ushpv18ct14
https://www.redbooks.ibm.com/?lnk=ushpv18ct10
https://www.ibm.com/support/home/?lnk=ushpv18ct11
https://www.ibm.com/training/?lnk=ushpv18ct15
https://www.ibm.com/cloud/hybrid?lnk=ushpv18ct20
https://www.ibm.com/cloud/learn/public-cloud?lnk=ushpv18ct17
https://www.ibm.com/cloud/redhat?lnk=ushpv18ct13
https://www.ibm.com/artificial-intelligence?lnk=ushpv18ct3
https://www.ibm.com/quantum-computing?lnk=ushpv18ct18
https://www.ibm.com/cloud/learn/kubernetes?lnk=ushpv18ct8
https://www.ibm.com/products/spss-statistics?lnk=ushpv18ct7
https://www.ibm.com/blockchain?lnk=ushpv18ct1
https://www-03.ibm.com/employment/technicaltalent/developer/?lnk=ushpv1
8ct2
https://www.ibm.com/search?lnk=ushpv18srch&locale=en-us&q=
https://www.ibm.com/products?lnk=ushpv18p1&lnk2=trial_mktpl&psrc=none&p
exp=def
https://www.ibm.com/cloud/hybrid?lnk=ushpv18pt14&bv=true
https://www.ibm.com/watson?lnk=ushpv18pt17&bv=true
https://www.ibm.com/us-en/products/categories?technologyTopics[0][0]=ca
t.topic:Blockchain&isIBMOffering[0]=true&lnk=ushpv18pt4&bv=true
https://www.ibm.com/us-en/products/category/technology/analytics?lnk=us
hpv18pt1&bv=true
https://www.ibm.com/financing?lnk=ushpv18pt3&bv=true
https://www.ibm.com/cloud/public?lnk=ushpv18pt15&bv=true
https://www.ibm.com/garage?lnk=ushpv18pt13&bv=true
https://www.ibm.com/cloud/automation?lnk=ushpv18ct21
https://www.ibm.com/us-en/products/category/technology/security?lnk=ush
pv18pt9&bv=true
https://www.ibm.com/quantum-computing?lnk=ushpv18pt16&bv=true
https://www.ibm.com/cloud/hybrid?lnk=ushpv18ct20
https://www.ibm.com/cloud/public?lnk=ushpv18ct17
https://www.ibm.com/cloud/redhat?lnk=ushpv18ct13
https://www.ibm.com/artificial-intelligence?lnk=ushpv18ct3
```

```
https://www.ibm.com/quantum-computing?lnk=ushpv18ct18
https://www.ibm.com/cloud/learn/kubernetes?lnk=ushpv18ct8
https://www.ibm.com/products/spss-statistics?lnk=ushpv18ct7
https://www.ibm.com/blockchain?lnk=ushpv18ct1
https://www-03.ibm.com/employment/technicaltalent/developer/?lnk=ushpv1
8ct2
https://www.ibm.com/
```

In [33]:

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

```
<img alt="" aria-hidden="true" role="presentation" src="
Ly93d3cudzMub3JnLzIwMDAvc3ZnIiB2ZXJzaW9uPSIxLjEiLz4=" style="max-width:
100%;display:block;margin:0;border:none;padding:0"/>

G1sbnM9Imh0dHA6Ly93d3cudzMub3JnLzIwMDAvc3ZnIiB2ZXJzaW9uPSIxLjEiLz4=
<img alt="leadspace mobile image" class="ibm-resize" decoding="async" s
rc="https://1.dam.s81c.com/public/content/dam/worldwide-content/homepag
e/ul/g/6a/68/20210531-Mayflower-AI-25917-mobile-720x360.jpg" style="pos
ition:absolute;top:0;left:0;bottom:0;right:0;box-sizing:border-box;padd
ing:0;border:none;margin:auto;display:block;width:0;height:0;min-width:
100%;max-width:100%;min-height:100%;max-height:100%"/>
https://1.dam.s81c.com/public/content/dam/worldwide-content/homepage/u
l/g/6a/68/20210531-Mayflower-AI-25917-mobile-720x360.jpg
<img alt="" aria-hidden="true" role="presentation" src="
d3cudzMub3JnLzIwMDAvc3ZnIiB2ZXJzaW9uPSIxLjEiLz4=" style="max-width:10
0%;display:block;margin:0;border:none;padding:0"/>

nM9Imh0dHA6Ly93d3cudzMub3JnLzIwMDAvc3ZnIiB2ZXJzaW9uPSIxLjEiLz4=
<img alt="Hybrid cloud, by the numbers
" class="ibm-resize ibm-ab-image featured-image" decoding="async" src
="https://1.dam.s81c.com/public/content/dam/worldwide-content/homepage/
ul/g/18/52/20210426-f-hybrid-cloud-value-calculator.jpg" style="positio
n:absolute;top:0;left:0;bottom:0;right:0;box-sizing:border-box;padding:
0;border:none;margin:auto;display:block;width:0;height:0;min-width:10
```

0%;max-width:100%;min-height:100%;max-height:100%"/>
https://1.dam.s81c.com/public/content/dam/worldwide-content/homepage/ul/g/18/52/20210426-f-hybrid-cloud-value-calculator.jpg

https://1.dam.s81c.com/public/content/dam/worldwide-content/homepage/ul/g/45/c0/20210531-websphere-hybrid-edition-444x320.jpg

https://1.dam.s81c.com/public/content/dam/worldwide-content/homepage/ul/g/7e/f0/20210531-Extended-Planning-Analysis-25919-444x320.jpg


```
<img alt="Your business needs trustworthy&amp;nbsp;AI" class="ibm-resiz
e ibm-ab-image featured-image" decoding="async" src="https://1.dam.s81
c.com/public/content/dam/worldwide-content/homepage/ul/g/e3/26/20210531
-trust-ai-watson-b-25922-444x320.jpg" style="position:absolute;top:0;le
ft:0;bottom:0;right:0;box-sizing:border-box;padding:0;border:none;margi
n:auto;display:block;width:0;height:0;min-width:100%;max-width:100%;min
-height:100%;max-height:100%"/>
https://1.dam.s81c.com/public/content/dam/worldwide-content/homepage/u
l/g/e3/26/20210531-trust-ai-watson-b-25922-444x320.jpg
<img alt="" aria-hidden="true" role="presentation" src="
d3cudzMub3JnLzIwMDAvc3ZnIiB2ZXJzaW9uPSIxLjEiLz4=" style="max-width:10
0%;display:block;margin:0;border:none;padding:0"/>

nM9Imh0dHA6Ly93d3cudzMub3JnLzIwMDAvc3ZnIiB2ZXJzaW9uPSIxLjEiLz4=
<img alt="IBM Cloud" class="ibm-resize ibm-ab-image trials-image" decod
ing="async" src="
AAAALAAAAAABAAEAAAIBRAA7" style="position:absolute;top:0;left:0;bottom:
0;right:0;box-sizing:border-box;padding:0;border:none;margin:auto;displ
ay:block;width:0;height:0;min-width:100%;max-width:100%;min-height:10
0%;max-height:100%"/>

AIBRAA7
<img alt="" aria-hidden="true" role="presentation" src="
d3cudzMub3JnLzIwMDAvc3ZnIiB2ZXJzaW9uPSIxLjEiLz4=" style="max-width:10
0%;display:block;margin:0;border:none;padding:0"/>

nM9Imh0dHA6Ly93d3cudzMub3JnLzIwMDAvc3ZnIiB2ZXJzaW9uPSIxLjEiLz4=
<img alt="IBM Cognos Analytics" class="ibm-resize ibm-ab-image trials-i
mage" decoding="async" src="
P///yH5BAEAAAAALAAAAAABAAEAAAIBRAA7" style="position:absolute;top:0;lef
t:0;bottom:0;right:0;box-sizing:border-box;padding:0;border:none;margi
n:auto;display:block;width:0;height:0;min-width:100%;max-width:100%;min
-height:100%;max-height:100%"/>

AIBRAA7
<img alt="" aria-hidden="true" role="presentation" src="
d3cudzMub3JnLzIwMDAvc3ZnIiB2ZXJzaW9uPSIxLjEiLz4=" style="max-width:10
```

```
0%;display:block;margin:0;border:none;padding:0"/>

nM9Imh0dHA6Ly93d3cudzMub3JnLzIwMDAvc3ZnIiB2ZXJzaW9uPSIxLjEiLz4=
<img alt="IBM Watson Assistant" class="ibm-resize ibm-ab-image trials-i
mage" decoding="async" src="
P///yH5BAEAAAAALAAAAAABAAEAAAIBRAA7" style="position:absolute;top:0;lef
t:0;bottom:0;right:0;box-sizing:border-box;padding:0;border:none;margi
n:auto;display:block;width:0;height:0;min-width:100%;max-width:100%;min
-height:100%;max-height:100%"/>

AIBRAA7
<img alt="" aria-hidden="true" role="presentation" src="
d3cudzMub3JnLzIwMDAvc3ZnIiB2ZXJzaW9uPSIxLjEiLz4=" style="max-width:10
0%;display:block;margin:0;border:none;padding:0"/>

nM9Imh0dHA6Ly93d3cudzMub3JnLzIwMDAvc3ZnIiB2ZXJzaW9uPSIxLjEiLz4=
<img alt="IBM Digital Learning Subscription" class="ibm-resize ibm-ab-i
mage trials-image" decoding="async" src="
AQABAIAAAAAAAP///yH5BAEAAAAALAAAAAABAAEAAAIBRAA7" style="position:absol
ute;top:0;left:0;bottom:0;right:0;box-sizing:border-box;padding:0;borde
r:none;margin:auto;display:block;width:0;height:0;min-width:100%;max-wi
dth:100%;min-height:100%;max-height:100%"/>

AIBRAA7
```

In [34]:
```
url = "https://cf-courses-data.s3.us.cloud-object-storage.appdomain.clo
ud/IBM-DA0321EN-
```

```
  File "<ipython-input-34-c3d53f6c9e73>", line 1
    url = "https://cf-courses-data.s3.us.cloud-object-storage.appdomai
n.cloud/IBM-DA0321EN-

^
SyntaxError: EOL while scanning string literal
```

In [ ]:
```

```