Web scraping, also known as screen scraping, data mining, web data extracting or web harvesting; is a method of extracting large amounts of data from a website. The extracted data are then analysed bringing out deductions and inferences from the data.

In [269]:

```
import requests
```

In [73]:

```
page = requests.get("http://dataquestio.github.io/web-scraping-pages/simple.html")
```

In [74]:

```
page
```

Out[74]:

```
<Response [200]>
```

In [75]:

```
page.status_code
```

Out[75]:

```
200
```

In [76]:

```
from bs4 import BeautifulSoup
```

In [77]:

```
soup = BeautifulSoup(page.content, "lxml")
```

In [78]:

```
print(soup.prettify())
```

```
<!DOCTYPE html>
<html>
 <head>
  <title>
   A simple example page
  </title>
 </head>
 <body>
  <p>
   Here is some simple content for this page.
  </p>
 </body>
</html>
```

In [79]:

```
list(soup.children)
```

Out[79]:

```
['html', <html>
 <head>
 <title>A simple example page</title>
 </head>
```

</head>
<body>
Here is some simple content for this page.
</body>
</html>]

In [80]:

```
[type(item) for item in list(soup.children)]
```

Out[80]:

```
[bs4.element.Doctype, bs4.element.Tag]
```

In [81]:

```
for item in list(soup.children):
    print(type(item))
```

```
<class 'bs4.element.Doctype'>
<class 'bs4.element.Tag'>
```

In [82]:

```
print(list(soup.children))
```

```
['html', <html>
<head>
<title>A simple example page</title>
</head>
<body>
<p>Here is some simple content for this page.</p>
</body>
</html>]
```

In [83]:

```
html = list(soup.children)[1]
```

In [84]:

```
list(html.children)
```

Out[84]:

```
['\n', <head>
 <title>A simple example page</title>
 </head>, '\n', <body>
 <p>Here is some simple content for this page.</p>
 </body>, '\n']
```

In [85]:

```
list(html.children)
```

Out[85]:

```
['\n', <head>
 <title>A simple example page</title>
 </head>, '\n', <body>
 <p>Here is some simple content for this page.</p>
 </body>, '\n']
```

In [86]:

```
body = list(html.children)[3]
```

In [87]:

```
list(body.children)
```

Out[87]:

```
['\n', <p>Here is some simple content for this page.</p>, '\n']
```

['\n', Here is some simple content for this page. , '\n']

In [88]:

```
p = list(body.children)[1]
```

In [89]:

```
p.get_text()
```

Out[89]:

```
'Here is some simple content for this page.'
```

Finding all instances of a tag at once

In [90]:

```
soup = BeautifulSoup(page.content, "lxml")
```

In [46]:

```
soup
```

Out[46]:

```
<!DOCTYPE html>
<html>
<head>
<title>A simple example page</title>
</head>
<body>
<p>Here is some simple content for this page.</p>
</body>
</html>
```

In [47]:

```
soup.find_all('p')
```

Out[47]:

```
[<p>Here is some simple content for this page.</p>]
```

In [95]:

```
soup.find_all('p')[0].get_text()
```

Out[95]:

```
'Here is some simple content for this page.'
```

In [98]:

```
soup.find('p')
```

Out[98]:

```
<p>Here is some simple content for this page.</p>
```

In [ ]:

```

```

In [ ]:

```

```

In [ ]:

```

```

Searching for tags by class and id

In [49]:

```
page = requests.get('http://dataquestio.github.io/web-scraping-pages/ids_and_classes.html
')
```

In [50]:

```
page
```

Out[50]:

```
<Response [200]>
```

In [51]:

```
page.status_code
```

Out[51]:

```
200
```

In [52]:

```
soup = BeautifulSoup(page.content, "lxml")
```

In [53]:

```
soup
```

Out[53]:

```
<html>
<head>
<title>A simple example page</title>
</head>
<body>
<div>
<p class="inner-text first-item" id="first">
                First paragraph.
            </p>
<p class="inner-text">
                Second paragraph.
            </p>
</div>
<p class="outer-text first-item" id="second">
<b>
                First outer paragraph.
            </b>
</p>
<p class="outer-text">
<b>
                Second outer paragraph.
            </b>
</p>
</body>
</html>
```

In [54]:

```
soup.prettify()
```

Out[54]:

```
'<html>\n <head>\n  <title>\n   A simple example page\n  </title>\n </head>\n <body>\n  <
div>\n   <p class="inner-text first-item" id="first">\n    First paragraph.\n   </p>\n
<p class="inner-text">\n    Second paragraph.\n   </p>\n  </div>\n  <p class="outer-text
first-item" id="second">\n   <b>\n    First outer paragraph.\n   </b>\n  </p>\n  <p class
="outer-text">\n   <b>\n    Second outer paragraph.\n   </b>\n  </p>\n </body>\n</html>'
```

In [60]:

```
soup.find_all('p', class_ = 'outer-text')
```

Out[60]:

```
[<p class="outer-text first-item" id="second">
 <b>
                 First outer paragraph.
             </b>
 </p>, <p class="outer-text">
 <b>
                 Second outer paragraph.
             </b>
 </p>]
```

In [62]:

```
soup.find_all(id = "first")
```

Out[62]:

```
[<p class="inner-text first-item" id="first">
                 First paragraph.
             </p>]
```

Using CSS Selectors

In [63]:

```
soup
```

Out[63]:

```
<html>
<head>
<title>A simple example page</title>
</head>
<body>
<div>
<p class="inner-text first-item" id="first">
                 First paragraph.
             </p>
<p class="inner-text">
                 Second paragraph.
             </p>
</div>
<p class="outer-text first-item" id="second">
<b>
                 First outer paragraph.
             </b>
</p>
<p class="outer-text">
<b>
                 Second outer paragraph.
             </b>
</p>
</body>
</html>
```

In [64]:

```
soup.select("div p")
```

Out[64]:

```
[<p class="inner-text first-item" id="first">
                 First paragraph.
             </p>, <p class="inner-text">
                 Second paragraph.
             </p>]
```

In [65]:

```
soup.select("html body")
```

Out[65]:

```
[<body>
 <div>
 <p class="inner-text first-item" id="first">
                First paragraph.
            </p>
 <p class="inner-text">
                Second paragraph.
            </p>
 </div>
 <p class="outer-text first-item" id="second">
 <b>
                First outer paragraph.
            </b>
 </p>
 <p class="outer-text">
 <b>
                Second outer paragraph.
            </b>
 </p>
 </body>]
```

In [66]:

```
soup.select("p#first")
```

Out[66]:

```
[<p class="inner-text first-item" id="first">
                First paragraph.
            </p>]
```

Downloading weather data

In [137]:

```
page = requests.get("https://forecast.weather.gov/MapClick.php?lat=37.7772&lon=-122.4168#
.XJ4dLJhKhEY")
```

In [138]:

```
page.status_code
```

Out[138]:

```
200
```

In [143]:

```
soup = BeautifulSoup(page.content, "lxml")
```

In [153]:

```
seven_day = soup.find_all(id="seven-day-forecast")[0]
forecast_items = seven_day.find_all(class_="tombstone-container")
```

In [154]:

```
tonight = forecast_items[0]
```

In [155]:

```
print(tonight.prettify())
```

```
<div class="tombstone-container">
```

```
 <p class="period-name">
  Today
  <br/>
  <br/>
 </p>
 <p>
  <img alt="Today: Mostly sunny, with a high near 60. West wind 5 to 10 mph increasing to
12 to 17 mph in the afternoon. Winds could gust as high as 23 mph. " class="forecast-icon
" src="newimages/medium/sct.png" title="Today: Mostly sunny, with a high near 60. West wi
nd 5 to 10 mph increasing to 12 to 17 mph in the afternoon. Winds could gust as high as 2
3 mph. "/>
 </p>
 <p class="short-desc">
  Mostly Sunny
 </p>
 <p class="temp temp-high">
  High: 60 °F
 </p>
</div>
```

Extracting information from the page

In [167]:

```
period = tonight.find(class_= "period-name").get_text()
```

In [166]:

```
period.get_text()
```

Out[166]:

```
'Today'
```

In [175]:

```
short_desc = tonight.find(class_ = "short-desc")
```

In [187]:

```
short_desc.get_text()
```

Out[187]:

```
'Mostly Sunny'
```

In [202]:

```
image = tonight.select("img")[0]
```

In [218]:

```
image["title"]
```

Out[218]:

```
'Today: Mostly sunny, with a high near 60. West wind 5 to 10 mph increasing to 12 to 17 m
ph in the afternoon. Winds could gust as high as 23 mph. '
```

In [178]:

```
img = tonight.find("img")
```

In [183]:

```
desc = img["title"]
```

In [181]:

```
print(desc)
```

```
Today: Mostly sunny, with a high near 60. West wind 5 to 10 mph increasing to 12 to 17 mp
h in the afternoon. Winds could gust as high as 23 mph.
```

Extracting all the information from the page

In [185]:

```
seven_day.prettify
```

Out[185]:

```
<bound method Tag.prettify of <div class="panel panel-default" id="seven-day-forecast">
<div class="panel-heading">
<b>Extended Forecast for</b>
<h2 class="panel-title">
          San Francisco CA </h2>
</div>
<div class="panel-body" id="seven-day-forecast-body">
<div id="seven-day-forecast-container"><ul class="list-unstyled" id="seven-day-forecast-l
ist"><li class="forecast-tombstone">
<div class="tombstone-container">
<p class="period-name">Today<br/><br/></p>
<p><img alt="Today: Mostly sunny, with a high near 60. West wind 5 to 10 mph increasing t
o 12 to 17 mph in the afternoon. Winds could gust as high as 23 mph. " class="forecast-ic
on" src="newimages/medium/sct.png" title="Today: Mostly sunny, with a high near 60. West
wind 5 to 10 mph increasing to 12 to 17 mph in the afternoon. Winds could gust as high as
23 mph. "/></p><p class="short-desc">Mostly Sunny</p><p class="temp temp-high">High: 60 °
F</p></div></li><li class="forecast-tombstone">
<div class="tombstone-container">
<p class="period-name">Tonight<br/><br/></p>
<p><img alt="Tonight: Mostly clear, with a low around 48. West northwest wind 13 to 18 mp
h decreasing to 6 to 11 mph in the evening. Winds could gust as high as 23 mph. " class="
forecast-icon" src="newimages/medium/nfew.png" title="Tonight: Mostly clear, with a low a
round 48. West northwest wind 13 to 18 mph decreasing to 6 to 11 mph in the evening. Wind
s could gust as high as 23 mph. "/></p><p class="short-desc">Mostly Clear</p><p class="te
mp temp-low">Low: 48 °F</p></div></li><li class="forecast-tombstone">
<div class="tombstone-container">
<p class="period-name">Saturday<br/><br/></p>
<p><img alt="Saturday: Mostly sunny, with a high near 64. North wind 5 to 10 mph becoming
west 12 to 17 mph in the afternoon. Winds could gust as high as 23 mph. " class="forecast
-icon" src="newimages/medium/sct.png" title="Saturday: Mostly sunny, with a high near 64.
North wind 5 to 10 mph becoming west 12 to 17 mph in the afternoon. Winds could gust as h
igh as 23 mph. "/></p><p class="short-desc">Mostly Sunny</p><p class="temp temp-high">Hig
h: 64 °F</p></div></li><li class="forecast-tombstone">
<div class="tombstone-container">
<p class="period-name">Saturday<br/>Night</p>
<p><img alt="Saturday Night: Partly cloudy, with a low around 49. West wind 11 to 16 mph
decreasing to 5 to 10 mph in the evening. Winds could gust as high as 21 mph. " class="fo
recast-icon" src="newimages/medium/nsct.png" title="Saturday Night: Partly cloudy, with a
low around 49. West wind 11 to 16 mph decreasing to 5 to 10 mph in the evening. Winds cou
ld gust as high as 21 mph. "/></p><p class="short-desc">Partly Cloudy</p><p class="temp t
emp-low">Low: 49 °F</p></div></li><li class="forecast-tombstone">
<div class="tombstone-container">
<p class="period-name">Sunday<br/><br/></p>
<p><img alt="Sunday: Mostly sunny, with a high near 68. Light and variable wind becoming
west 11 to 16 mph in the afternoon. Winds could gust as high as 21 mph. " class="forecast
-icon" src="newimages/medium/sct.png" title="Sunday: Mostly sunny, with a high near 68. L
ight and variable wind becoming west 11 to 16 mph in the afternoon. Winds could gust as h
igh as 21 mph. "/></p><p class="short-desc">Mostly Sunny</p><p class="temp temp-high">Hig
h: 68 °F</p></div></li><li class="forecast-tombstone">
<div class="tombstone-container">
<p class="period-name">Sunday<br/>Night</p>
<p><img alt="Sunday Night: Mostly cloudy, with a low around 52." class="forecast-icon" sr
c="newimages/medium/nbkn.png" title="Sunday Night: Mostly cloudy, with a low around 52."/
></p><p class="short-desc">Mostly Cloudy</p><p class="temp temp-low">Low: 52 °F</p></div>
</li><li class="forecast-tombstone">
<div class="tombstone-container">
<p class="period-name">Monday<br/><br/></p>
<p><img alt="Monday: A 40 percent chance of rain.  Mostly cloudy, with a high near 65." c
lass="forecast-icon" src="newimages/medium/ra40.png" title="Monday: A 40 percent chance o
```

```
t rain.  Mostly cloudy, with a high near 65."/></p><p class="short-desc">Chance Rain</p><
p class="temp temp-high">High: 65 °F</p></div></li><li class="forecast-tombstone">
<div class="tombstone-container">
<p class="period-name">Monday<br/>Night</p>
<p><img alt="Monday Night: A chance of showers.  Mostly cloudy, with a low around 52." cl
ass="forecast-icon" src="newimages/medium/nshra.png" title="Monday Night: A chance of sho
wers.  Mostly cloudy, with a low around 52."/></p><p class="short-desc">Chance<br/>Shower
s</p><p class="temp temp-low">Low: 52 °F</p></div></li><li class="forecast-tombstone">
<div class="tombstone-container">
<p class="period-name">Tuesday<br/><br/></p>
<p><img alt="Tuesday: A chance of showers.  Mostly cloudy, with a high near 62." class="f
orecast-icon" src="newimages/medium/shra.png" title="Tuesday: A chance of showers.  Mostl
y cloudy, with a high near 62."/></p><p class="short-desc">Chance<br/>Showers</p><p class
="temp temp-high">High: 62 °F</p></div></li></ul></div>
<script type="text/javascript">
// equalize forecast heights
$(function () {
 var maxh = 0;
 $(".forecast-tombstone .short-desc").each(function () {
  var h = $(this).height();
  if (h > maxh) { maxh = h; }
 });
 $(".forecast-tombstone .short-desc").height(maxh);
});
</script> </div>
</div>>
```

In [216]:

```
period_tags = seven_day.select(".tombstone-container .period-name")
```

In [217]:

```
print(period_tags)
```

```
[<p class="period-name">Today<br/><br/></p>, <p class="period-name">Tonight<br/><br/></p>
, <p class="period-name">Saturday<br/><br/></p>, <p class="period-name">Saturday<br/>Nigh
t</p>, <p class="period-name">Sunday<br/><br/></p>, <p class="period-name">Sunday<br/>Nig
ht</p>, <p class="period-name">Monday<br/><br/></p>, <p class="period-name">Monday<br/>Ni
ght</p>, <p class="period-name">Tuesday<br/><br/></p>]
```

In [211]:

```
periods =[pa.get_text() for pa in period_tags]
```

In [207]:

```
periods
```

Out[207]:

```
['Today',
 'Tonight',
 'Saturday',
 'SaturdayNight',
 'Sunday',
 'SundayNight',
 'Monday',
 'MondayNight',
 'Tuesday']
```

In [236]:

```
short_desc_tags = seven_day.select(".tombstone-container .short-desc")
```

In [237]:

```
print(short_desc_tags)
```

```
[<p class="short-desc">Mostly Sunny</p>, <p class="short-desc">Mostly Clear</p>, <p class
="short-desc">Mostly Sunny</p>, <p class="short-desc">Partly Cloudy</p>, <p class="short-
desc">Mostly Sunny</p>, <p class="short-desc">Mostly Cloudy</p>, <p class="short-desc">Ch
```

```
ance Rain</p>, <p class="short-desc">Chance<br/>Showers</p>, <p class="short-desc">Chance
<br/>Showers</p>]
```

In [243]:

```
short_description = [pa.get_text() for pa in short_desc_tags]
```

In [244]:

```
short_description
```

Out[244]:

```
['Mostly Sunny',
 'Mostly Clear',
 'Mostly Sunny',
 'Partly Cloudy',
 'Mostly Sunny',
 'Mostly Cloudy',
 'Chance Rain',
 'ChanceShowers',
 'ChanceShowers']
```

In [245]:

```
temp_tags = seven_day.select(".tombstone-container .temp")
```

In [246]:

```
print(temp_tags)
```

```
[<p class="temp temp-high">High: 60 °F</p>, <p class="temp temp-low">Low: 48 °F</p>, <p c
lass="temp temp-high">High: 64 °F</p>, <p class="temp temp-low">Low: 49 °F</p>, <p class=
"temp temp-high">High: 68 °F</p>, <p class="temp temp-low">Low: 52 °F</p>, <p class="temp
temp-high">High: 65 °F</p>, <p class="temp temp-low">Low: 52 °F</p>, <p class="temp temp-
high">High: 62 °F</p>]
```

In [247]:

```
temp_description = [pa.get_text() for pa in temp_tags]
```

In [248]:

```
print(temp_description)
```

```
['High: 60 °F', 'Low: 48 °F', 'High: 64 °F', 'Low: 49 °F', 'High: 68 °F', 'Low: 52 °F', '
High: 65 °F', 'Low: 52 °F', 'High: 62 °F']
```

In [251]:

```
temp_description
```

Out[251]:

```
['High: 60 °F',
 'Low: 48 °F',
 'High: 64 °F',
 'Low: 49 °F',
 'High: 68 °F',
 'Low: 52 °F',
 'High: 65 °F',
 'Low: 52 °F',
 'High: 62 °F']
```

In [258]:

```
image_tags = seven_day.select(".tombstone-container img")
```

In [259]:

```
image_tags
```

Out[259]:

```
[<img alt="Today: Mostly sunny, with a high near 60. West wind 5 to 10 mph increasing to
12 to 17 mph in the afternoon. Winds could gust as high as 23 mph. " class="forecast-icon
" src="newimages/medium/sct.png" title="Today: Mostly sunny, with a high near 60. West wi
nd 5 to 10 mph increasing to 12 to 17 mph in the afternoon. Winds could gust as high as 2
3 mph. "/>,
 <img alt="Tonight: Mostly clear, with a low around 48. West northwest wind 13 to 18 mph
decreasing to 6 to 11 mph in the evening. Winds could gust as high as 23 mph. " class="fo
recast-icon" src="newimages/medium/nfew.png" title="Tonight: Mostly clear, with a low aro
und 48. West northwest wind 13 to 18 mph decreasing to 6 to 11 mph in the evening. Winds
could gust as high as 23 mph. "/>,
 <img alt="Saturday: Mostly sunny, with a high near 64. North wind 5 to 10 mph becoming w
est 12 to 17 mph in the afternoon. Winds could gust as high as 23 mph. " class="forecast-
icon" src="newimages/medium/sct.png" title="Saturday: Mostly sunny, with a high near 64.
North wind 5 to 10 mph becoming west 12 to 17 mph in the afternoon. Winds could gust as h
igh as 23 mph. "/>,
 <img alt="Saturday Night: Partly cloudy, with a low around 49. West wind 11 to 16 mph de
creasing to 5 to 10 mph in the evening. Winds could gust as high as 21 mph. " class="fore
cast-icon" src="newimages/medium/nsct.png" title="Saturday Night: Partly cloudy, with a l
ow around 49. West wind 11 to 16 mph decreasing to 5 to 10 mph in the evening. Winds coul
d gust as high as 21 mph. "/>,
 <img alt="Sunday: Mostly sunny, with a high near 68. Light and variable wind becoming we
st 11 to 16 mph in the afternoon. Winds could gust as high as 21 mph. " class="forecast-i
con" src="newimages/medium/sct.png" title="Sunday: Mostly sunny, with a high near 68. Lig
ht and variable wind becoming west 11 to 16 mph in the afternoon. Winds could gust as hig
h as 21 mph. "/>,
 <img alt="Sunday Night: Mostly cloudy, with a low around 52." class="forecast-icon" src=
"newimages/medium/nbkn.png" title="Sunday Night: Mostly cloudy, with a low around 52."/>,
 <img alt="Monday: A 40 percent chance of rain.  Mostly cloudy, with a high near 65." cla
ss="forecast-icon" src="newimages/medium/ra40.png" title="Monday: A 40 percent chance of
rain.  Mostly cloudy, with a high near 65."/>,
 <img alt="Monday Night: A chance of showers.  Mostly cloudy, with a low around 52." clas
s="forecast-icon" src="newimages/medium/nshra.png" title="Monday Night: A chance of showe
rs.  Mostly cloudy, with a low around 52."/>,
 <img alt="Tuesday: A chance of showers.  Mostly cloudy, with a high near 62." class="for
ecast-icon" src="newimages/medium/shra.png" title="Tuesday: A chance of showers.  Mostly
cloudy, with a high near 62."/>]
```

In [264]:

```
image_description = [d["title"] for d in image_tags]
```

In [265]:

```
image_description
```

Out[265]:

```
['Today: Mostly sunny, with a high near 60. West wind 5 to 10 mph increasing to 12 to 17
mph in the afternoon. Winds could gust as high as 23 mph. ',
 'Tonight: Mostly clear, with a low around 48. West northwest wind 13 to 18 mph decreasin
g to 6 to 11 mph in the evening. Winds could gust as high as 23 mph. ',
 'Saturday: Mostly sunny, with a high near 64. North wind 5 to 10 mph becoming west 12 to
17 mph in the afternoon. Winds could gust as high as 23 mph. ',
 'Saturday Night: Partly cloudy, with a low around 49. West wind 11 to 16 mph decreasing
to 5 to 10 mph in the evening. Winds could gust as high as 21 mph. ',
 'Sunday: Mostly sunny, with a high near 68. Light and variable wind becoming west 11 to
16 mph in the afternoon. Winds could gust as high as 21 mph. ',
 'Sunday Night: Mostly cloudy, with a low around 52.',
 'Monday: A 40 percent chance of rain.  Mostly cloudy, with a high near 65.',
 'Monday Night: A chance of showers.  Mostly cloudy, with a low around 52.',
 'Tuesday: A chance of showers.  Mostly cloudy, with a high near 62.']
```

Combining our data into a Pandas Dataframe

In [250]:

```
import pandas as pd
```

In [266]:

```
weather = pd.DataFrame(
   { "periods" : periods,
    "short_description" : short_description,
    "temp_description" : temp_description,
    "image_description" : image_description
}
)
```

In [267]:

```
weather
```

Out[267]:

	periods	short_description	temp_description	image_description
0	Today	Mostly Sunny	High: 60 °F	Today: Mostly sunny, with a high near 60. West...
1	Tonight	Mostly Clear	Low: 48 °F	Tonight: Mostly clear, with a low around 48. W...
2	Saturday	Mostly Sunny	High: 64 °F	Saturday: Mostly sunny, with a high near 64. N...
3	SaturdayNight	Partly Cloudy	Low: 49 °F	Saturday Night: Partly cloudy, with a low arou...
4	Sunday	Mostly Sunny	High: 68 °F	Sunday: Mostly sunny, with a high near 68. Lig...
5	SundayNight	Mostly Cloudy	Low: 52 °F	Sunday Night: Mostly cloudy, with a low around...
6	Monday	Chance Rain	High: 65 °F	Monday: A 40 percent chance of rain. Mostly c...
7	MondayNight	ChanceShowers	Low: 52 °F	Monday Night: A chance of showers. Mostly clo...
8	Tuesday	ChanceShowers	High: 62 °F	Tuesday: A chance of showers. Mostly cloudy, ...

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]: