```
Chapter 6 - Data Sourcing via Web

Part 4 - Web Scraping
```

In [2]:

```
from bs4 import BeautifulSoup
import urllib.request
from IPython.display import HTML
import re
```

In [3]:

```
r = urllib.request.urlopen("https://analytics.usa.gov/").read()
soup = BeautifulSoup(r, "lxml")
type(soup)
```

Out[3]:

```
bs4.BeautifulSoup
```

In [4]:

```
print(soup.prettify()[:100])
```

```
<!DOCTYPE html>
<html lang="en">
 <!-- Initalize title and data source variables -->
 <head>
  <!--
```

In [6]:

```
for link in soup.find_all('a'):
    print(link.get('href'))
```

```
/
#explanation
https://analytics.usa.gov/data/ (https://analytics.usa.gov/data/)
https://open.gsa.gov/api/dap/ (https://open.gsa.gov/api/dap/)
data/
#top-pages-realtime
#top-pages-7-days
#top-pages-30-days
https://analytics.usa.gov/data/live/all-pages-realtime.csv (https://analytics.u
sa.gov/data/live/all-pages-realtime.csv)
https://analytics.usa.gov/data/live/all-domains-30-days.csv (https://analytics.
usa.gov/data/live/all-domains-30-days.csv)
https://www.digitalgov.gov/services/dap/ (https://www.digitalgov.gov/services/d
ap/)
https://www.digitalgov.gov/services/dap/common-questions-about-dap-faq/#part-4
 (https://www.digitalgov.gov/services/dap/common-questions-about-dap-faq/#part-
4)
https://support.google.com/analytics/answer/2763052?hl=en (https://support.goog
le.com/analytics/answer/2763052?hl=en)
https://analytics.usa.gov/data/live/second-level-domains.csv (https://analytic
s.usa.gov/data/live/second-level-domains.csv)
https://analytics.usa.gov/data/live/sites.csv (https://analytics.usa.gov/data/l
ive/sites.csv)
mailto:DAP@support.digitalgov.gov
https://analytics.usa.gov/data/ (https://analytics.usa.gov/data/)
https://open.gsa.gov/api/dap/ (https://open.gsa.gov/api/dap/)
mailto:DAP@support.digitalgov.gov
https://github.com/GSA/analytics.usa.gov/issues (https://github.com/GSA/analyti
cs.usa.gov/issues)
https://github.com/GSA/analytics.usa.gov (https://github.com/GSA/analytics.usa.
gov)
https://github.com/18F/analytics-reporter (https://github.com/18F/analytics-rep
orter)
http://www.gsa.gov/ (http://www.gsa.gov/)
https://www.digitalgov.gov/services/dap/ (https://www.digitalgov.gov/services/d
ap/)
https://cloud.gov/ (https://cloud.gov/)
```

In [7]:

```
print(soup.get_text())
```

```


            Much more detailed data is available in downloadable CSV and JSO
N. This includes data on combined browser and OS usage.



Browsers





Internet Explorer





Operating Systems
```

```
In [8]: print(soup.prettify()[0:1000])
```

```
<!DOCTYPE html>
<html lang="en">
 <!-- Initalize title and data source variables -->
 <head>
  <!--

    Hi! Welcome to our source code.

    This dashboard uses data from the Digital Analytics Program, a US
    government team inside the General Services Administration.


    For a detailed tech breakdown of how 18F and friends built this site:

    https://18f.gsa.gov/2015/03/19/how-we-built-analytics-usa-gov/ (https://18
f.gsa.gov/2015/03/19/how-we-built-analytics-usa-gov/)


    This is a fully open source project, and your contributions are welcome.

    Frontend static site: https://github.com/18F/analytics.usa.gov (https://git
hub.com/18F/analytics.usa.gov)
    Backend data reporting: https://github.com/18F/analytics-reporter (https://
github.com/18F/analytics-reporter)

    -->
  <meta charset="utf-8"/>
  <meta content="IE=Edge" http-equiv="X-UA-Compatible"/>
  <meta content="NjbZn6hQe7OwV-nTsa6nLmtrOUcSGPRyFjxm5zkmCcg" name="google-site
-verification"/>
  <link href="/css/vendor/css/uswds.v0.9.6.css" rel="stylesheet"/>
  <link href="/css/public_analytics.css" rel="stylesheet"/>
  <link href="/images/analytics-favicon.ico" rel="ic
```

In [9]:

```
for link in soup.findAll('a', attrs={'href': re.compile("^http")}):
    print(link)
    type(link)
```

```
<a href="https://analytics.usa.gov/data/">Data</a>
<a href="https://open.gsa.gov/api/dap/" target="_blank">API</a>
<a href="https://analytics.usa.gov/data/live/all-pages-realtime.csv">Download t
he full dataset.</a>
<a href="https://analytics.usa.gov/data/live/all-domains-30-days.csv">Download
the full dataset.</a>
<a class="external-link" href="https://www.digitalgov.gov/services/dap/">Digita
l Analytics Program</a>
<a class="external-link" href="https://www.digitalgov.gov/services/dap/common-q
uestions-about-dap-faq/#part-4">does not track individuals</a>
<a class="external-link" href="https://support.google.com/analytics/answer/2763
052?hl=en">anonymizes the IP addresses</a>
<a class="external-link" href="https://analytics.usa.gov/data/live/second-level
-domains.csv">400 executive branch government domains</a>
<a class="external-link" href="https://analytics.usa.gov/data/live/sites.csv">a
bout 5,700 total websites</a>
<a href="https://analytics.usa.gov/data/">download the data here.</a>
<a href="https://open.gsa.gov/api/dap/" target="_blank"> API project</a>
<a class="usa-button usa-button-secondary-inverse" href="https://github.com/GS
A/analytics.usa.gov/issues">
<img alt="Github Icon" class="github-icon" src="/images/github-logo-white.svg"/
>
                  Suggest a feature or report an issue
            </a>
<a href="https://github.com/GSA/analytics.usa.gov">
<img alt="Github Icon" class="github-icon" src="/images/github-logo.svg"/>
              View our code on GitHub</a>
<a href="https://github.com/18F/analytics-reporter">
<img alt="Github Icon" class="github-icon" src="/images/github-logo.svg"/>
              View our code for the data on GitHub</a>
<a href="http://www.gsa.gov/">
<img alt="GSA" src="/images/gsa-logo.svg"/>
</a>
<a href="https://www.digitalgov.gov/services/dap/">Digital Analytics Program</a
>
<a href="https://cloud.gov/">cloud.gov</a>
```

In [10]:

```
file = open("Parsed_data.txt", "w")
for link in soup.findAll('a', attrs={'href': re.compile("^http")}):
    soup_link = str(link)
    print(soup_link)
    file.write(soup_link)
file.flush()
file.close()
```

```
<a href="https://analytics.usa.gov/data/">Data</a>
<a href="https://open.gsa.gov/api/dap/" target="_blank">API</a>
<a href="https://analytics.usa.gov/data/live/all-pages-realtime.csv">Download t
he full dataset.</a>
<a href="https://analytics.usa.gov/data/live/all-domains-30-days.csv">Download
the full dataset.</a>
<a class="external-link" href="https://www.digitalgov.gov/services/dap/">Digita
l Analytics Program</a>
<a class="external-link" href="https://www.digitalgov.gov/services/dap/common-q
uestions-about-dap-faq/#part-4">does not track individuals</a>
<a class="external-link" href="https://support.google.com/analytics/answer/2763
052?hl=en">anonymizes the IP addresses</a>
<a class="external-link" href="https://analytics.usa.gov/data/live/second-level
-domains.csv">400 executive branch government domains</a>
<a class="external-link" href="https://analytics.usa.gov/data/live/sites.csv">a
bout 5,700 total websites</a>
<a href="https://analytics.usa.gov/data/">download the data here.</a>
<a href="https://open.gsa.gov/api/dap/" target="_blank"> API project</a>
<a class="usa-button usa-button-secondary-inverse" href="https://github.com/GS
A/analytics.usa.gov/issues">
<img alt="Github Icon" class="github-icon" src="/images/github-logo-white.svg"/
>
                  Suggest a feature or report an issue
            </a>
<a href="https://github.com/GSA/analytics.usa.gov">
<img alt="Github Icon" class="github-icon" src="/images/github-logo.svg"/>
              View our code on GitHub</a>
<a href="https://github.com/18F/analytics-reporter">
<img alt="Github Icon" class="github-icon" src="/images/github-logo.svg"/>
              View our code for the data on GitHub</a>
<a href="http://www.gsa.gov/">
<img alt="GSA" src="/images/gsa-logo.svg"/>
</a>
<a href="https://www.digitalgov.gov/services/dap/">Digital Analytics Program</a
>
<a href="https://cloud.gov/">cloud.gov</a>
```

In [11]:

```
%pwd
```

Out[11]:

```
'C:\\Users\\danal'
```