Chapter 6 - Data Sourcing via Web

Segment 4 - Web scraping

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

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

```
bs4.BeautifulSoup
```

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

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

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

```
/
#explanation
https://analytics.usa.gov/data/
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.usa.gov/data/live/all-domains-30-days.csv
https://www.digitalgov.gov/services/dap/
```

```
    https://www.digitalgov.gov/services/dap/common-questions-about-dap-faq/#part-4
    https://support.google.com/analytics/answer/2763052?hl=en
    https://analytics.usa.gov/data/live/second-level-domains.csv
    https://analytics.usa.gov/data/live/sites.csv
    dap@gsa.gov
    https://analytics.usa.gov/data/
    https://open.gsa.gov/api/dap/
    dap@gsa.gov
    https://github.com/GSA/analytics.usa.gov/issues
    https://github.com/GSA/analytics.usa.gov
    https://github.com/18F/analytics-reporter
    http://www.gsa.gov/
    https://www.digital.gov/guides/dap/
    https://cloud.gov/
```

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

```
    analytics.usa.gov | The US government's web traffic.




    (function(w,d,s,l,i){w[l]=w[l]||[];w[l].push({'gtm.start':
        new Date().getTime(),event:'gtm.js'});var f=d.getElementsByTagName(s)[0],
        j=d.createElement(s),dl=l!='dataLayer'?'&l='+l:'';j.async=true;j.src=
        'https://www.googletagmanager.com/gtm.js?id='+i+dl;f.parentNode.insertBefore(j,f);
        })(window,document,'script','dataLayer','GTM-MQSGZS');
```

```
  (function(i,s,o,g,r,a,m){i['GoogleAnalyticsObject']=r;i[r]=i[r]||function(){
(i[r].q=i[r].q||[]).push(arguments)},i[r].l=1*new Date();a=s.createElement(o),
m=s.getElementsByTagName(o)[0];a.async=1;a.src=g;m.parentNode.insertBefore(a,m)
})(window,document,'script','https://www.google-analytics.com/analytics.js','ga');

  ga('create', 'UA-48605964-36', 'auto');
  ga('set', 'anonymizeIp', true);
  ga('set', 'forceSSL', true);
  ga('send', 'pageview');









analytics.usa.gov
```

```
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/


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

    Frontend static site: https://github.com/18F/analytics.usa.gov
    Backend data reporting: 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
```

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

```
https://analytics.usa.gov/data/
https://open.gsa.gov/api/dap/
https://analytics.usa.gov/data/live/all-pages-realtime.csv
https://analytics.usa.gov/data/live/all-domains-30-days.csv
https://www.digitalgov.gov/services/dap/
https://www.digitalgov.gov/services/dap/common-questions-about-dap-faq/#part-4
https://support.google.com/analytics/answer/2763052?hl=en
https://analytics.usa.gov/data/live/second-level-domains.csv
https://analytics.usa.gov/data/live/sites.csv
https://analytics.usa.gov/data/
https://open.gsa.gov/api/dap/
https://github.com/GSA/analytics.usa.gov/issues
https://github.com/GSA/analytics.usa.gov
https://github.com/18F/analytics-reporter
http://www.gsa.gov/
https://www.digital.gov/guides/dap/
https://cloud.gov/
```

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

```
https://analytics.usa.gov/data/
https://open.gsa.gov/api/dap/
https://analytics.usa.gov/data/live/all-pages-realtime.csv
https://analytics.usa.gov/data/live/all-domains-30-days.csv
https://www.digitalgov.gov/services/dap/
https://www.digitalgov.gov/services/dap/common-questions-about-dap-faq/#part-4
https://support.google.com/analytics/answer/2763052?hl=en
https://analytics.usa.gov/data/live/second-level-domains.csv
https://analytics.usa.gov/data/live/sites.csv
https://analytics.usa.gov/data/
https://open.gsa.gov/api/dap/
https://github.com/GSA/analytics.usa.gov/issues
https://github.com/GSA/analytics.usa.gov
https://github.com/18F/analytics-reporter
http://www.gsa.gov/
https://www.digital.gov/guides/dap/
https://cloud.gov/
```

```
%pwd
```

```
'/content'
```