AAM1 — AAM1 TASK 1: WEB SCRAPING

Nanako Ohashi

Western Governors University

A.

Note: There is no text that matches “Current Estimates” on the link provided (https://www.census.gov/programs-surveys/popest.html), so I am assuming that the data scraping needed to be done on the link provided.

1. Open a Python program.

2. Using bs4, import the BeautifulSoup package. Also import the urllib, re, and csv libraries.

3. Create a variable and call the urllib.request.urlopen function. Insert the United States Census Bureau link provided (United States Census Bureau, n.d.) in the function and then use the .read() function.

4. Create a Beautiful Soup object and call the BeautifulSoup constructor, and pass in r using the ‘xlml’ parser.

5. Use the prettify function to add some structure and make it easy to read.

6. Create an empty list.

7. Create a for loop for links in soup, call the find all method and then pass in a string that reads “a”. For each of the “a” tags we want to append the links that have an attribute of ‘href’ to the empty list (Pierson, 2017).

Fig 1. Import libraries and use urllib to read link.

```
In [131]: from bs4 import BeautifulSoup
          import urllib
          import re
          import csv

In [132]: r = urllib.request.urlopen('https://www.census.gov/programs-surveys/popest.html').read()
          soup = BeautifulSoup(r, "lxml")
          type(soup)

Out[132]: bs4.BeautifulSoup
```

Fig 2. Scraping a webpage and saving your results.

Scraping a webpage and saving your results

In [133]:
```
print(soup.prettify()[:100])
```

```
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Strict//EN" "http://www.w3.org/TR/xhtml1/DTD/xhtml1-str
```

In [134]:
```
# Gather all the hrefs for the a tags
links = []
for link in soup.find_all('a'):
    links.append(link.get('href'))
print(links)
```

```
[None, '#content', 'https://www.census.gov/en.html', None, None, None, None, None, None, None, 'https://www.census.gov/topic
s/population/age-and-sex.html', 'https://www.census.gov/businessandeconomy', 'https://www.census.gov/topics/education.html',
'https://www.census.gov/topics/preparedness.html', 'https://www.census.gov/topics/employment.html', 'https://www.census.gov/t
opics/families.html', 'https://www.census.gov/topics/population/migration.html', 'https://www.census.gov/programs-surveys/geo
graphy.html', 'https://www.census.gov/topics/health.html', 'https://www.census.gov/topics/population/hispanic-origin.html',
'https://www.census.gov/topics/housing.html', 'https://www.census.gov/topics/income-poverty.html', 'https://www.census.gov/to
pics/international-trade.html', 'https://www.census.gov/topics/population.html', 'https://www.census.gov/topics/population/po
pulation-estimates.html', 'https://www.census.gov/topics/public-sector.html', 'https://www.census.gov/topics/population/race.
html', 'https://www.census.gov/topics/research.html', 'https://www.census.gov/topics/public-sector/voting.html', 'https://ww
w.census.gov/about/index.html', 'https://www.census.gov/data.html', 'https://www.census.gov/academy', 'https://www.census.go
v/about/what/admin-data.html', 'https://www.census.gov/data/data-tools.html', 'https://www.census.gov/developers/', 'https://
www.census.gov/data/related-sites.html', 'https://www.census.gov/data/software.html', 'https://www.census.gov/data/tables.htm
l', 'https://www.census.gov/data/training-workshops.html', 'https://www.census.gov/library/visualizations.html', 'https://ww
w.census.gov/library.html', 'https://www.census.gov/AmericaCounts', 'https://www.census.gov/library/audio.html', 'https://ww
w.census.gov/library/visualizations.html', 'https://www.census.gov/library/photos.html', 'https://www.census.gov/library/publ
ications.html', 'https://www.census.gov/library/video.html', 'https://www.census.gov/library/working-papers.html', 'https://w
ww.census.gov/programs-surveys.html', 'https://www.census.gov/programs-surveys/are-you-in-a-survey.html', 'https://www.censu
s.gov/2020census', 'https://www.census.gov/programs-surveys/decennial-census/decade/2010.html', 'https://www.census.gov/progr
ams-surveys/acs', 'https://www.census.gov/programs-surveys/ahs.html', 'https://www.census.gov/programs-surveys/abs.html', 'ht
```

B.

I used a for loop that checked to see if each link started with a backslash.

Fig 3. For loop finding relative links.

```
In [147]: # Turn relative links into absolute links
          for i in range(len(links)):
              l = links[i]
              if l[:1] == "/":
                  links[i] = "https://www.census.gov"+l
                  print(l, "-->", links[i])
```

C.

I used a for loop to find the relative links that started with a backslash and appended "https://www.census.gov" in front of the link.

Fig 4. For loop saving relative links as absolute URIs.

```
In [147]: # Turn relative links into absolute links
          for i in range(len(links)):
              l = links[i]
              if l[:1] == "/":
                  links[i] = "https://www.census.gov"+l
                  print(l, "-->", links[i])
```

```
/en.html --> https://www.census.gov/en.html
/programs-surveys.html --> https://www.census.gov/programs-surveys.html
/programs-surveys/popest/about.html --> https://www.census.gov/programs-surveys/popest/about.html
/programs-surveys/popest/data.html --> https://www.census.gov/programs-surveys/popest/data.html
/programs-surveys/popest/geographies.html --> https://www.census.gov/programs-surveys/popest/geographies.html
/programs-surveys/popest/guidance.html --> https://www.census.gov/programs-surveys/popest/guidance.html
/programs-surveys/popest/guidance-geographies.html --> https://www.census.gov/programs-surveys/popest/guidance-geographies.html
/programs-surveys/popest/library.html --> https://www.census.gov/programs-surveys/popest/library.html
/programs-surveys/popest/news.html --> https://www.census.gov/programs-surveys/popest/news.html
/programs-surveys/popest/technical-documentation.html --> https://www.census.gov/programs-surveys/popest/technical-documentation.html
/programs-surveys.html --> https://www.census.gov/programs-surveys.html
/programs-surveys/popest/about.html --> https://www.census.gov/programs-surveys/popest/about.html
/programs-surveys/popest/data.html --> https://www.census.gov/programs-surveys/popest/data.html
/programs-surveys/popest/geographies.html --> https://www.census.gov/programs-surveys/popest/geographies.html
/programs-surveys/popest/guidance.html --> https://www.census.gov/programs-surveys/popest/guidance.html
/programs-surveys/popest/guidance-geographies.html --> https://www.census.gov/programs-surveys/popest/guidance-geographies.html
/programs-surveys/popest/library.html --> https://www.census.gov/programs-surveys/popest/library.html
/programs-surveys/popest/news.html --> https://www.census.gov/programs-surveys/popest/news.html
/programs-surveys/popest/technical-documentation.html --> https://www.census.gov/programs-surveys/popest/technical-documentation.html
/programs-surveys.html --> https://www.census.gov/programs-surveys.html
/programs-surveys/popest/about.html --> https://www.census.gov/programs-surveys/popest/about.html
/programs-surveys/popest/data/tables.html --> https://www.census.gov/programs-surveys/popest/data/tables.html
/programs-surveys/popest/technical-documentation.html --> https://www.census.gov/programs-surveys/popest/technical-documentation.html
/programs-surveys/popest/about/schedule.html --> https://www.census.gov/programs-surveys/popest/about/schedule.html
/newsroom/press-releases/2019/estimates-characteristics.html --> https://www.census.gov/newsroom/press-releases/2019/estimates-characteristics.html
/newsroom/press-releases/2019/estimates-characteristics/estimates-characteristics-sp.html --> https://www.census.gov/newsroom/press-releases/2019/estimates-characteristics/estimates-characteristics-sp.html
/newsroom/press-releases/2019/characteristics-embargo-advisory.html --> https://www.census.gov/newsroom/press-releases/2019/characteristics-embargo-advisory.html
/programs-surveys/popest/news.html --> https://www.census.gov/programs-surveys/popest/news.html
/data/tables/time-series/demo/popest/pre-1980-national.html --> https://www.census.gov/data/tables/time-series/demo/popest/pre-1980-national.html
/data/tables/time-series/demo/popest/pre-1980-state.html --> https://www.census.gov/data/tables/time-series/demo/popest/pre-1980-state.html
/data/tables/time-series/demo/popest/pre-1980-county.html --> https://www.census.gov/data/tables/time-series/demo/popest/pre-1980-county.html
```

D. & E.

I created an empty list, then I added all unique entries to that list to ensure that there were no duplicated links. The code I used to execute this can be found below (fig 5 and 6).

Fig 5. For loop removing unnecessary characters to be able to compare all links for uniqueness.

```
In [148]: # Strip https://, http://, and trailing backslashes from list
          for i in range(len(http_links)):
              l = http_links[i]
              if l[:8] == "https://":
                  http_links[i] = l[8:]

              l = http_links[i]
              if l[:7] == "http://":
                  http_links[i] = l[7:]

              l = http_links[i]
              if l[-1:] == "/":
                  http_links[i] = l[:len(l)-1]
          print(http_links)
```

```
['www.census.gov/en.html', 'www.census.gov/topics/population/age-and-sex.html', 'www.census.gov/businessandeconomy', 'www.census.gov/topics/education.html', 'www.census.gov/topics/preparedness.html', 'www.census.gov/topics/employment.html', 'www.census.gov/topics/families.html', 'www.census.gov/topics/population/migration.html', 'www.census.gov/programs-surveys/geography.html', 'www.census.gov/topics/health.html', 'www.census.gov/topics/population/hispanic-origin.html', 'www.census.gov/topics/housing.html', 'www.census.gov/topics/income-poverty.html', 'www.census.gov/topics/international-trade.html', 'www.census.gov/topics/population.html', 'www.census.gov/topics/population/population-estimates.html', 'www.census.gov/topics/public-sector.html', 'www.census.gov/topics/population/race.html', 'www.census.gov/topics/research.html', 'www.census.gov/topics/public-sector/voting.html', 'www.census.gov/about/index.html', 'www.census.gov/data.html', 'www.census.gov/academy', 'www.census.gov/about/what/admin-data.html', 'www.census.gov/data/data-tools.html', 'www.census.gov/developers', 'www.census.gov/data/related-sites.html', 'www.census.gov/data/software.html', 'www.census.gov/data/tables.html', 'www.census.gov/data/training-workshops.html', 'www.census.gov/library/visualizations.html', 'www.census.gov/library.html', 'www.census.gov/AmericaCounts', 'www.census.gov/library/audio.html', 'www.census.gov/library/visualizations.html', 'www.census.gov/library/photos.html', 'www.census.gov/library/publications.html', 'www.census.gov/library/video.html', 'www.census.gov/library/working-papers.html', 'www.census.gov/programs-surveys.html', 'www.census.gov/programs-surveys/are-you-in-a-survey.html', 'www.census.gov/2020census', 'www.census.gov/programs-surveys/decennial-census/decade/2010.html', 'www.census.gov/programs-surveys/acs', 'www.census.gov/programs-surveys/ahs.html', 'www.census.gov/programs-surveys/abs.html', 'www.census.gov/programs-surveys/asm.html', 'www.census.gov/programs-surveys/cog.html', 'www.census.gov/programs-surveys/cbp.html', 'www.census.gov/programs-surveys/cps.html', 'www.census.gov/EconomicCensus', 'www.census.gov/internationalprograms', 'www.census.gov/programs-surveys/metro-micro.html', 'www.census.gov/programs-surveys/popproj.html', 'www.census.gov/programs-surveys/saipe.html', 'www.census.gov/programs-surveys/susb.html',
```

Fig 6. For loop inserting unique links into empty list.

```
In [149]: # remove duplicates
          unique_links = []
          for l in http_links:
              if not l in unique_links:
                  unique_links.append(l)
          print(len(http_links))
          print(len(unique_links))
          unique_links.sort()
```

```
176
90
```

F.

Provided on a separate document: www.census.gov-programs-surveys-popest.html

G.

Provided in a separate document: unique_links.csv

I used the following code to create a .csv file (Martelli, 2010):

“with open('unique_links.csv', 'w', newline='') as myfile:

    wr = csv.writer(myfile, quoting=csv.QUOTE_ALL)

    wr.writerow(unique_links)”

H.

In [151]:
```
for l in unique_links:
    print(l)
```

```
twitter.com/uscensusbureau
www.census.gov/2020census
www.census.gov/AmericaCounts
www.census.gov/EconomicCensus
www.census.gov/about-us
www.census.gov/about/business-opportunities.html
www.census.gov/about/contact-us.html
www.census.gov/about/contact-us/staff-finder.html
www.census.gov/about/faqs.html
www.census.gov/about/history.html
www.census.gov/about/index.html
www.census.gov/about/policies.html
www.census.gov/about/policies/privacy/privacy-policy.html#accessibility
www.census.gov/about/what.html
www.census.gov/about/what/admin-data.html
www.census.gov/about/who.html
www.census.gov/academy
www.census.gov/businessandeconomy
www.census.gov/careers
www.census.gov/data.html
www.census.gov/data/data-tools.html
www.census.gov/data/developers/data-sets/Geocoding-services.html
www.census.gov/data/related-sites.html
www.census.gov/data/software.html
www.census.gov/data/tables.html
www.census.gov/data/training-workshops.html
www.census.gov/datalinkage
www.census.gov/developers
www.census.gov/en.html
www.census.gov/fieldjobs
www.census.gov/foia
www.census.gov/internationalprograms
www.census.gov/library.html
www.census.gov/library/audio.html
www.census.gov/library/photos.html
www.census.gov/library/publications.html
www.census.gov/library/reference/code-lists/naics.html
www.census.gov/library/reference/code-lists/schedule/b.html
www.census.gov/library/video.html
```

I.

References

Martelli, A. (2010, January 18). Create a .csv file with values from a Python list [Online Forum Comment]. Retrieved from https://stackoverflow.com/questions/2084069/create-a-csv-file-with-values-from-a-python-list

Pierson, L. (2017). *Web scrape in practice.* Retrieved from https://www.linkedin.com/learning/python-for-data-science-essential-training/web-scrape-in-practice?u=2045532

United States Census Bureau (n.d.). *Population and Housing Unit Estimates.* Retrieved from https://www.census.gov/programs-surveys/popest.html