## **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.

## 

## Fig 2. Scraping a webpage and saving your results.

## 

**B.**

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

Fig 3. For loop finding relative links.

**C.**

I used a for loop to find the relative links that started with a backslash and appended “[https://www.census.gov](https://www.census.gov/)” in front of the link.

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

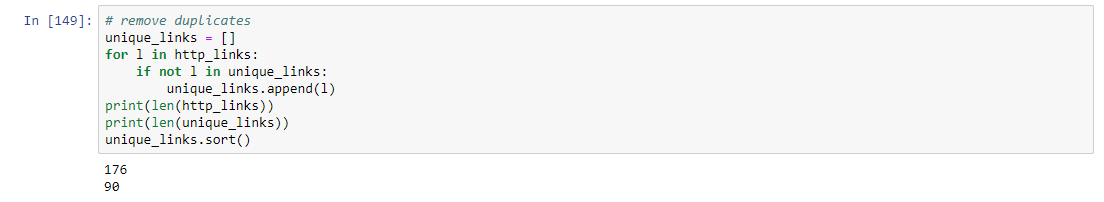
## 

## **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.

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



**F.**

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

**G.**

Provided in a separate document: unique\_links.csv

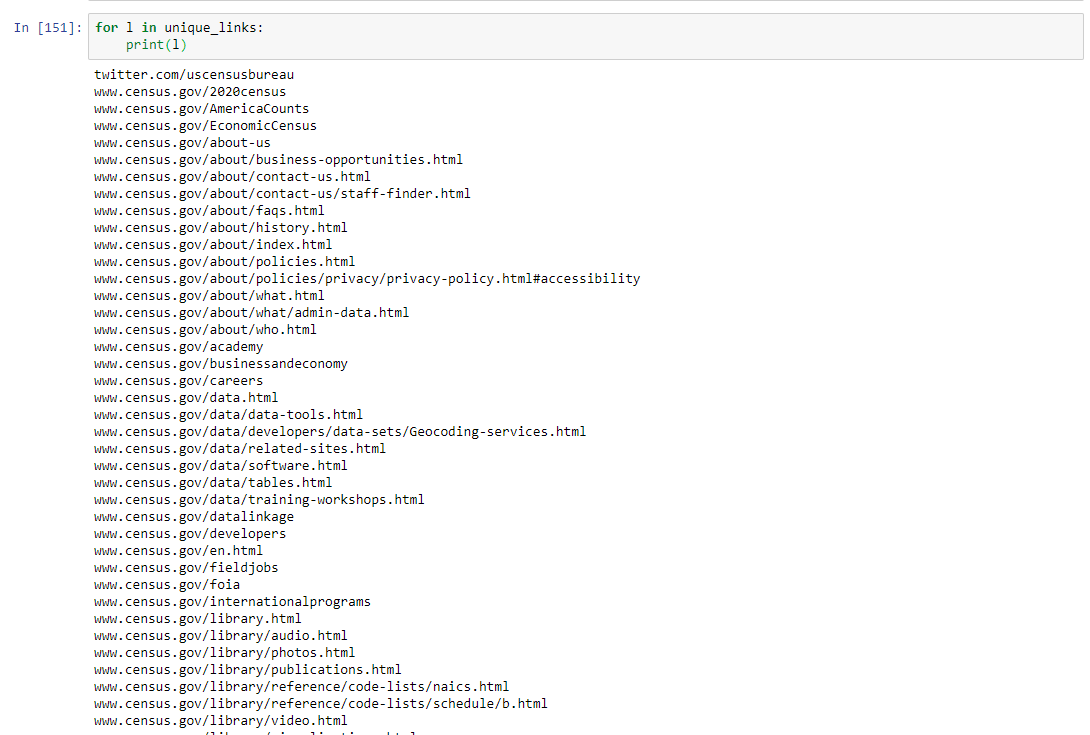
I used the following code to create a .csv file (Martelli, 2010):

“with open('unique\_links.csv', 'w', newline='') as myfile:

wr = csv.writer(myfile, quoting=csv.QUOTE\_ALL)

wr.writerow(unique\_links)”

**H.**

****

**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>