

## Part 1: Python

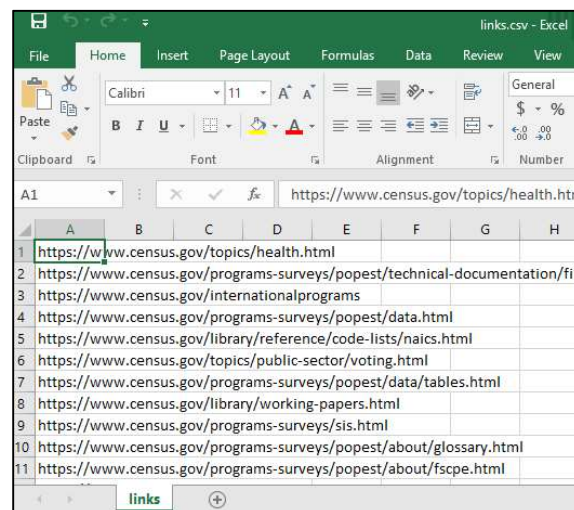
- A. The Python program extracts the links from the HTML code by using the html parser in the BeautifulSoup module to find all of the 'a' tags that have 'href' attributes.
- B. The criteria I used to determine if a link is a locator to another HTML page is by using the html parser to find all 'a' tags that have the 'href' attribute.
- C. Starting at line 32 of the Python script, a 'for loop' is run that checks each link to see if it begins with a forward slash ('/') or a pound symbol ('#'). If so, it appends the beginning of the link to include the main website. According to W3Schools (HTML a href Attribute), there are several different types of links. Forward slashes or pound symbols are considered relative links. The updated links are put into a new list and later outputted to a csv file after duplicates are removed.
- D. On line 56 of the Python script, one line of code is used to remove duplicates. By definition (8.7. Sets), a set cannot contain duplicates. So the final list of links is converted into a set and then back into a list, having had all of the duplicate entries removed when it was converted to a set.
- F. The HTML code of the "Current Estimates" web page scraped at the time when the scraper was run is located within the Part 1 – Python folder in a text file.
- H. Test of the script and screenshots of the successfully executed results are below.

```
Administrator: Command Prompt

C:\Users\rbirmi016\Downloads\C742\C742 Project\Part 1 - Python>python webScraper.py
Pulling and exporting of links complete!

C:\Users\rbirmi016\Downloads\C742\C742 Project\Part 1 - Python>
```

Figure 1 - The script being run in Command Prompt



	A	B	C	D	E	F	G	H
1	https://www.census.gov/topics/health.html							
2	https://www.census.gov/programs-surveys/popest/technical-documentation/file							
3	https://www.census.gov/internationalprograms							
4	https://www.census.gov/programs-surveys/popest/data.html							
5	https://www.census.gov/library/reference/code-lists/naics.html							
6	https://www.census.gov/topics/public-sector/voting.html							
7	https://www.census.gov/programs-surveys/popest/data/tables.html							
8	https://www.census.gov/library/working-papers.html							
9	https://www.census.gov/programs-surveys/sis.html							
10	https://www.census.gov/programs-surveys/popest/about/glossary.html							
11	https://www.census.gov/programs-surveys/popest/about/fscpe.html							

Figure 2 - The csv file that is created as output

## References

HTML a href Attribute. (Retrieved January 20<sup>th</sup>, 2019). Retrieved from:

[https://www.w3schools.com/tags/att\\_a\\_href.asp](https://www.w3schools.com/tags/att_a_href.asp)

Population and Housing Unit Estimates. (Retrieved January 17<sup>th</sup>, 2019). Retrieved from:

<https://www.census.gov/programs-surveys/popest.html>

8.7. Sets — unordered collections of unique elements. (Retrieved January 20<sup>th</sup>, 2019).

Retrieved from: <https://docs.python.org/2/library/sets.html>