C996: PYTHON PROGRAMMING

In this assignment, I will use Jupyter Notebook to demonstrate my Python skill in gathering and extracting information.

A. At first, I import all the packages needed for this exercise. Secondly, I use the 'request' method to grab content from the URL. Then, BeautifulSoup is used to parse the links and a loop is used to find all links that lead to Html pages.

```
# importing packages
import requests
from bs4 import BeautifulSoup, SoupStrainer
import csv
import re
from urllib.parse import urljoin
```

```
# request method is used to grab content from the url
url = "https://www.census.gov/programs-surveys/popest.html"
r = requests.get(url)
raw_content = r.text
#print(r.text)
```

```
# using BeautifulSoup to parse the page
soup = BeautifulSoup(raw_content, 'html.parser')
#print(soup.prettify())
```

```
# innitializing a list
links = list()

# gathering all web links pointing to html page and add to a list, also removing duplicates
for item in soup.find_all('a', href=re.compile(r'http')):
    if item not in links:
        links.append(item.get('href'))
```

B. Look at the code below

```
# innitializing a list
links = list()

# gathering all web links pointing to html page and add to a list, also removing duplicates
for item in soup.find_all('a', href=re.compile(r'http')):
    if item not in links:
        links.append(item.get('href'))
```

I use 'find_all' method to search for links that having a 'href' tag because they will be connected to another Html webpages. Then I stored all the links in a list. This step also removed duplicated links.

C. My program needs to make sure that relative URLs are stored as absolute URLs in the csv file.

```
for link in links:
    if 'https://www.census.gov/' not in link:
        link = 'https://www.census.gov' + link
        edited_links.append(link)
    else:
        edited_links.append(link)
```

This line of code will identify if the links start with 'https://www.census.gov/', then add 'https://www.census.gov' if it is missing, so we are able to get all the URLs as absolute.

D. My program needs to make sure that all the URLs are unique in the CSV file. It used a Python loop to go through all the elements inside the list to pick up unique links and stored them in a list.

```
# innitializing a list
links = list()

# gathering all web links pointing to html page and add to a list, also removing duplicates
for item in soup.find_all('a', href=re.compile(r'http')):
    if item not in links:
        links.append(item.get('href'))
```

Also, my program checked the list one more time to make sure if Task C may create some duplicates

```
# checking if the newly created edited_links list containing any duplicated links
final_list = list()
for fl in edited_links:
    if fl not in final_list:
        final_list.append(fl)
```

E. The Python code that gathers all the unique URLs pointing out to other web pages is attached below.

```
# importing packages
import requests
from bs4 import BeautifulSoup, SoupStrainer
import csv
import re
from urllib.parse import urljoin
```

```
# request method is used to grab content from the url
url = "https://www.census.gov/programs-surveys/popest.html"
r = requests.get(url)
raw_content = r.text
#print(r.text)
```

```
# using BeautifulSoup to parse the page
soup = BeautifulSoup(raw_content, 'html.parser')
#print(soup.prettify())
```

```
# innitializing a list
links = list()

# gathering all web links pointing to html page and add to a list, also removing duplicates
for item in soup.find_all('a', href=re.compile(r'http')):
    if item not in links:
        links.append(item.get('href'))
```

```python
# saving the relatives links as absolute links if not absolute
edited_links = list()

for link in links:
    if 'https' not in link:
        link = 'https://www.census.gov' + link
        edited_links.append(link)
    else:
        edited_links.append(link)

#print(links)
```

```python
# checking if the newly created edited_links list containing any duplicated links
final_list = list()
for fl in edited_links:
    if fl not in final_list:
        final_list.append(fl)

#print(final_list)
```

```python
# Outputing data to a csv file names C996.csv
with open("C996.csv", "w") as f:
    wr = csv.writer(f, delimiter ="\n")
    wr.writerow(final_list)
```

F. The Html code was saved under Current_Estimate.txt (attached in this assessment)

G. The CSV file which created after my program executed is C996.csv

H. My program is saved as C996 Coding File.ipynb
Screenshot:

```python
# checking if the newly created edited_links list containing any duplicated links
final_list = list()
for fl in edited_links:
    if fl not in final_list:
        final_list.append(fl)

print(final_list)
```

```
['https://www.census.gov/en.html', 'https://www.census.gov/topics/population/age-and-sex.html', 'http
s://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.c
ensus.gov/topics/families.html', 'https://www.census.gov/topics/population/migration.html', 'https://
www.census.gov/programs-surveys/geography.html', 'https://www.census.gov/topics/health.html', 'http
s://www.census.gov/topics/population/hispanic-origin.html', 'https://www.census.gov/topics/housing.ht
ml', 'https://www.census.gov/topics/income-poverty.html', 'https://www.census.gov/topics/internationa
```

My output file is C996.csv
Screenshot:

jupyter C996.csv 8 minutes ago

File Edit View Language

```
https://www.census.gov/en.html
https://www.census.gov/topics/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/topics/families.html
https://www.census.gov/topics/population/migration.html
https://www.census.gov/programs-surveys/geography.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/topics/international-trade.html
https://www.census.gov/topics/population.html
https://www.census.gov/topics/population/population-estimates.html
https://www.census.gov/topics/public-sector.html
```

Or you can open the file in the Microsoft Excel program.

	A	B	C	D	E	F
1	https://www.census.gov/en.html					
2	https://www.census.gov/topics/population/age-and-sex.html					
3	https://www.census.gov/businessandeconomy					
4	https://www.census.gov/topics/education.html					
5	https://www.census.gov/topics/preparedness.html					
6	https://www.census.gov/topics/employment.html					
7	https://www.census.gov/topics/families.html					
8	https://www.census.gov/topics/population/migration.html					
9	https://www.census.gov/programs-surveys/geography.html					
10	https://www.census.gov/topics/health.html					
11	https://www.census.gov/topics/population/hispanic-origin.html					
12	https://www.census.gov/topics/housing.html					

I. Reference.
None

J. Presentation as above