Programming in Python – C996

Student Name: David Bailey

Student ID: #000658614

Date: 01/22/2020

Student Mentor Name: Kelly Smith

A, B, C, & D: Code Explanation

The Python program uses BeautifulSoup to parse through the HTML text. This allows the different tags in the HTML text to be searched and separated, as needed. Telling python and BeautifulSoup to look for all “a” tags that include “href” will give us the links that are included on the source webpage. The section of code below executes three of the desired actions by finding the links, saving them as absolute references, and removing duplicates.

links = []

for url in soup.find\_all('a'):

link = urljoin('https://www.census.gov/programs-surveys/popest.html',url.get('href'))

if link not in links:

links.append(link)

“Soup.find\_all(‘a’)” finds all of the data that is included in a tag, and “url.get(‘href’)” makes sure the “a” tag includes “href”, which web links have. It also separates the link from the tag itself.

From the urllib.parse library the urljoin function combines the relative references with the domain name so that the links show the absolute reference when printed.

To remove duplicates, I created a list called links and told python to look at each link and only add it to the output if it wasn’t already accounted for in the file.

# E: Functioning Python Code

See “Web Scraping Python Code.py”

# F: HTML Code

See “HTML from print(soup).txt”

# G: CSV File

See “Links from Census Data.csv”

# H: Screenshot of Results

See below or “Snip of IDLE after running code.png”

A screenshot of a cell phone

Description automatically generated