**Programming in Python Task: Web Scraping**

Heidi L. Petersen

College of Information Technology, Western Governors University

C996: Programming in Python

Dr. Emelda Ntinglet

August 10, 2020

Programming in Python Task: Web Scraping

My program “Task” collects the weblinks from the website census.gov, gets rid of the duplicate links, and puts them in an excel file for easy viewing. See the Jupyter notebook screenshot to follow along with each of the inputs. The program first (Input 2) imports the libraries needed to complete the assignment. After (Input 3), it reads the webpage and creates a parse tree. This is the variable called “Soup”. Next (Input 4), it creates list variables for the absolute and relative links. Following that, Input 5 has Python look for the <a> html tag that begin with the string “http” for the absolute links within the webpage. It then (Input 6) looks for the html tags that begin with a backslash. Input 6 of my program takes the list of relative links and adds the landing page address as a string at the beginning of each of the relative links. It then appends the list of relative links so that it now contains an absolute link for each link. It adds each of the links to the individual lists of absolute links and relative links. Input 7 combines the absolute and relative lists into one list. Input 8 removes the duplicate links by combining the lists into one and then converting it into a set. When it is a set, the duplicate stings are eliminated. Then I changed the set back into a list. Input 9, takes each link and puts each on their own line. Input 10, creates the excel file.