C996 Task Prompt Responses

Elizabeth Sweet

Western Governors University

C996 Task Prompt Responses

A. Explain how the Python program extracts the web links from the HTML code of the “Current Estimates,” found in the weblinks section.

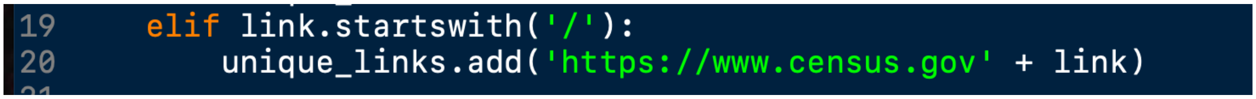
Python uses the request module to get the requested website and return a response object. This response object contains all the information about the website (Reitz, 2019). The response object passes into the BeautifulSoup constructor, which converts it into Unicode. Then BeautifulSoup parses the HTML tags from the document (Richardson, 2020).

B. Explain the criteria you used to determine if a link is a locator to another HTML page. Identify the code segment that executes this action as part of your explanation.

HTML tags <a> indicate a hyperlink in the code. The href attribute gives the link’s destination (HTML Links, 2020). Using find_all, BeautifulSoup searches through all the HTML code and returns these links. They became the variable all_links.

C. Explain how the program ensures that relative links are saved as absolute URLs in the output file. Identify the code segment that executed this action as part of your explanation.

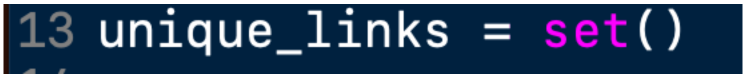
The relative links began with a /. Any link that started this way was converted to an absolute URL by adding the prefix “https://www.census.gov.”   Lines 19 and 20 in the program test each link and add the prefix as needed.



```
19      elif link.startswith('/'):
20          unique_links.add('https://www.census.gov' + link)
```

D. Explain how the program ensures that there are no duplicated links in the output file. Identify the code that executed this action as part of your explanation.

To ensure that there were no duplicates in the output file, I chose to save the links as a set. The set data type does not allow for duplicate elements. The main drawback of using a set is that the order of the objects in the set is not maintained. In this situation, that was not a problem, but for others, using a dictionary might be a better choice (Singh, n.d.). The relevant code is located in line 13.



```
13 unique_links = set()
```

E. Provide the Python code you wrote to extract all the unique links from the HTML code of the “Current Estimates” (in the weblinks section), that point out to other HTML pages.

Included as C996 Final.py

F. Provide the HTML code of the “Current Estimates” web page scrapped at the time when the scraper was run and the CSV file was generated.

Included as FlyBy.html

G. Provide the CSV file that your script created.

Included as unique_links.csv

H. Run you script and provide a screenshot of the successfully executed results.

Included as Screen Shot Successful Run.png

Bibliography

HTML Links. (2020). Retrieved from w3schools.com: https://www.w3schools.com/html/html_links.asp

Reitz, K. (2019). *Quickstart*. Retrieved from Requsets: HTTP for Humans: https://requests.readthedocs.io/en/master/user/quickstart/#make-a-request

Richardson, L. (2020). *Beautiful Soup Documentation*. Retrieved from https://www.crummy.com/software/BeautifulSoup/bs4/doc/

Singh, M. (n.d.). *Python - Ways to remove duplicates from list*. Retrieved from Geeks for Geeks: https://www.geeksforgeeks.org/python-ways-to-remove-duplicates-from-list/