C996 Proficiency Assessment

Jonathan DiQuattro

Programming in Python C996

Emelda Ntinglet, PhD

May 28, 2020

Abstract

This paper is being submitted as a final assessment for the C996 Programming in Python class, and demonstrates the skills learned throughout the course by developing a web links scraper program in Python that extracts all the unique web links from the “US Census Bureau” website.

C996 Proficiency Assessment C996

**Develop a web links scraper program in Python that extracts all of the unique web links that point out to other web pages from the HTML code of the “Current Estimates” web link, both from the “US Census Bureau” website (see web link below) and outside that domain, and that populates them in a comma-separated values (CSV) file as absolute uniform resource indicators (URIs).**

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

This program utilizes the Beautiful Soup 4 and Requests python packages to pull the html from <https://www.census.gov/programs-surveys/popest.html> via Requests, and pass to the BeautifulSoup() function. The Beautiful soup 4 package comes with a findAll() function that can take the anchor tag (findAll(‘a’)) as an argument and return an array of every web link from the page. The array was then iterated over in a for loop, where the href text was extracted using the Beautiful Soup 4 function get(‘href’), which can be applied to an html tag to pull the attributes.

**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.**

When the href was retrieved from each anchor tag the first 8 characters of the href are checked to determine if the URL starts with a ‘/’, which indicates a local. If the url starts with a ‘/’ it is appended to ‘<https://ww.census.gov>’, and passed to a function that checks the first 22 characters of a string for the ‘<https://ww.census.gov>’ text, and is the second indicator that a URL is local. If the URL meets those criteria it is added to the unique set, otherwise it is added to the external set.

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

Each URL is checked for a leading ‘/’, the ‘<https://ww.census.gov>’ strings, and a trailing ‘/’ with the stripSlash function before storing in a set. Combined with appending ‘<https://ww.census.gov>’ to each URL that starts with a ‘/’, ensures that each URI is absolute.

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

The python set() object prevents any duplicates from being stored, and utilizing the set after the URL’s were stripped of trailing and starting ‘/’ characters ensured that only one entry for each unique value was stored. In the isExternal function there is an externalUrls and uniqueUrls that are implemented after all URL’s have been cleaned.

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

The code is provided in the attached scraper.py file.

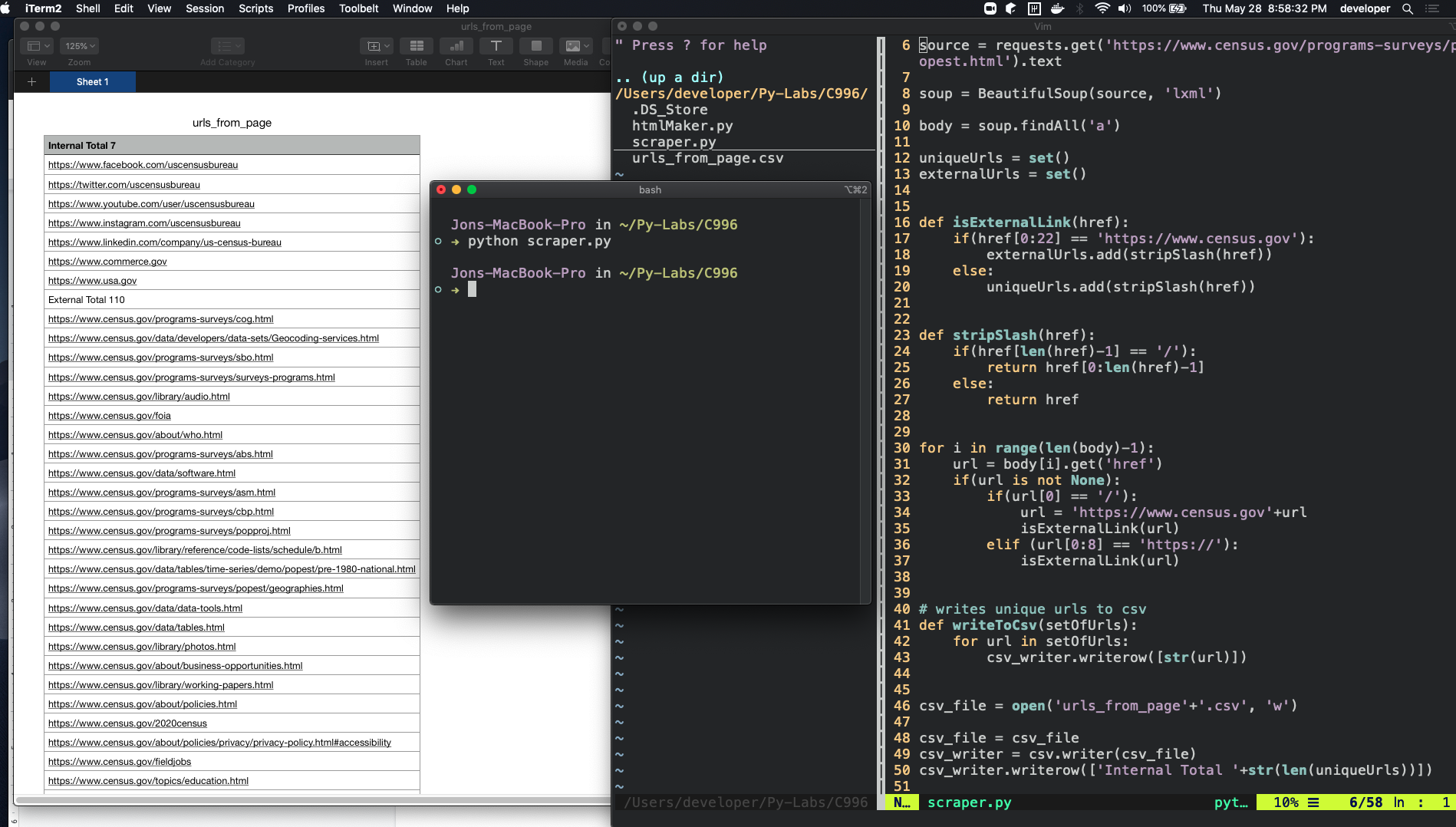
**F. Provide the HTML code of the “Current Estimates” web page scrapped at the time when the scraper was run and the CSV file was generated.**

The HTML code is located in the attached govhtml.htm file.

**G. Provide the CSV file that your script created.**

The created CSV is located in the attached urls\_fom\_page.csv file.

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

****

**I. Acknowledge sources, using in-text citations and references, for content that is quoted, paraphrased, or summarized.**

The citations for this paper were generated using the APA citation wizard from the writing center.

References

Beautiful Soup Documentation. (n.d.). Retrieved from https://www.crummy.com/software/BeautifulSoup/bs4/doc/