Report 01

Seyed Ali Firooz Abadi

Problem

Extract data from the district of choice using the language of choice.
The program should ask for the district number as input and read the content of the district's URL (or local version of the district). The program should report statistics of the content and write the content to an output file.

Solution

1. To solve this problem, I used 3 libraries Requests, BeautifulSoup, and OS. I started the program by asking the user to enter a number for the district the user wants to check. The user's input will then go in a while loop to check if it is the right input. I used try except to account for any unforeseen errors, and gave the user a maximum of 3 tries after which the program will terminate itself.
2. If the user input was 78, the rest of the program will run, but if not, it will show a message informing the user that other districts are not supported yet.
3. If in fact the user input was 78, we proceed to request the webpage HTML file using the Request library and the get() function. Then we parse the HTML code using first "lxml" and then "html5lib", we do this so that we could have a better HTML code since each parsing method is slightly different (lxml is better in handling bad files and html5lib is better organized).
4. After getting a good HTML file of the district webpage, we start by extracting data from it using BeautifulSoup functions. Different filters (queries) were used to find the data (due to the bad structure of the HTML file (where CSS styling elements were implemented in line rather as classes!))
5. After successfully targeting the HTML tags, I used a BS function to get the text inside that tag and save it as a string. Some elements were harder to target and I was left out with a list of strings rather than just one string.
6. String manipulation were then done on some data that needed more attention.

7. After getting all needed data as strings stored in there own variables, I started the process of opening the output file using the “with” keyboard to eliminate the need to close the text file afterwards.
8. I have written all the variables inside the text file with a simple formatting to make it easier to read.
9. After finishing the writing process, I was ready to apply my user-defined function to count the words, lines, characters, and empty spaces and output the statistics to a different text file.
    a. The user-defined function used the OS library to strip the formated text file.
    b. First I used a for loop to manipulate the strings and count the words and characters. Each time the loop runs will increase the lines, words, char, and spaces count.

Output

We get all needed information stored as strings in there own variables and 2 output files that been created by the program.