Stamurai
Data Scraping internship

By-Mohammed Sufiyan Abdullah Ghori

I made use of Selenium with Python to extract the data from 3 Schools.
Before going into the details with respect to each school I would like to describe few of the considerationsI took while extracting:

1. For all 3 schools instead of going in each of their schools, I filtered the schools and scraped data for
   1.1. High Schools in Fairfax County Public Schools.
   1.2. Middle Schools in Miami Dade County Public Schools
   1.3. Middle Schools in Prince William County Public Schools
2. The Code I wrote in
   2.1. Fairfax Schools will work for all of its schools as the website template is common
   2.2. Miami Schools doesn't have a common template at large (Individual Extraction is better)
   2.3. Prince William Schools have common template for same schools like middle schools had common template.
3. All the data and code has been uploaded to Github.
4. There are 3 folders with each named as the schools in the assignment. Each Folder contains
   4.1. Code in Python
   4.2. List of all the schools in that school in 2 formats (CSV , HTML)
   4.3. List of Individual Schools Staff is again stored in both formats and stored in 2 separate folders. Each School has a data file with its name with scraped data like name, position, email etc.

Steps Involved

1. **Fairfax County Public Schools**
   1.1. Loaded the required Libraries
   1.2. Opened the website using webdriver
   1.3. Clicked on the Schools in menu
   1.4. Changed items per page to max possible(50)
   1.5. Selected High schools from filters to be applied
   1.6. Read total number of records from the text before records
   1.7. Hiding the side filter to get max view
   1.8. Scrolling till we reach page navigator
   1.9. Clicking next page after reading current page until reaching last page
   1.10. While iterating these records, storing school name,link,principal name,email,address,staff site link in DB List of schools.
   1.11. Then looping through the staff link column and storing the name and position of each staff in the school with file name as school name.
   1.12. Closing webdrivers

2. **Miami Dade County Public Schools**
   2.1. Loaded the required Libraries

2.2. Opened the website using webdriver
2.3. Clicked on the Schools in menu then School Directory
2.4. Scrolling a bit to select Middle Schools
2.5. Selecting all items in items per page
2.6. Iterating through all the records and storing information like School name,LOC,Administrator name,email,phone no,address,school reference links .
2.7. Iterating through school reference links to get school website links.
2.8. Iterating through website links to get staff directory links (In this process we came across a few sites which were completely different from others in finding staff details.)
2.9. Followed sites with 2 common website templates and staff directory links
2.10. While iterating we came to know only 1 type of template has a common way of storing staff details for that a loop was used and data was extracted i.e. name,email.
2.11. While for other common template of staff link the way of storing staff details is unique every time so had to write different codes for them.
2.12. So I wrote code for a few individual schools and stored data (Individual School Codes folder).
2.13. Closing webdrivers

3. Prince William County Public Schools

3.1. Loaded the required Libraries
3.2. Opened the website using webdriver
3.3. Clicked on the Select schools in menu
3.4. Storing only Middle School names with their respective site links
3.5. While viewing the school sites I observed the pathname used to store staff details was the same. So just appending path name to school links to get staff links
3.6. Iterating through staff links and storing staff details with school name as file names.
3.7. An interesting observation here was I was able to get all staff details without scrolling or changing pages in the page navigator at once.
3.8. Closing webdrivers

Summary

1. **Fairfax County Public Schools**
    a. High Schools Scraped
    b. 28 Schools Separate staff files
    c. 1 file with all Schools names and other details

2. **Miami Dade County Public Schools**
    a. Middle Schools Scraped
    b. 13 Schools Separate staff files
    c. 1 file with all Schools names and other details

3. **Prince William County Public Schools**
    a. Middle Schools Scraped
    b. 17 Schools Separate staff files
    c. 1 file with all Schools names and other details