EDPSM KPI Data Scraper

Glenn Harper

Python dependencies:

Selenium

Beautiful soup 4 –bs4

lxml

Regex

Request

Problems:

- When grabbing data from the Catalog task list table, its very difficult to understand what data you're grabbing, as blank spaces

in the table are td tags with no text, however when grabbing those tags and checking for empty tags they aren’t considered “empty” and are just ignored. I Tried using bs4’s find(text=False) but that returned tags with and without text.

- for 8 tasks on page, got a count of

- Prints worknotes in console but were showed shortened in original html

# Solution: was screening cell’s tags in table as using the line - “if task\_data[k].get\_text == None or task\_data[k].get\_text == “” “ – the get\_text function was being **referenced** not **called**. Also was doing the actual check based on the **tag** values, which was just checking if the string values I just stored were empty, which they would never be.

- When trying to access the username box on login page, kept throwing timeout errors with selenium because the webdriver couldn’t find any elements on the page. The issue was: 1) that I needed to put in timed waits inbetween any actions that selenium takes to make sure all web elements were loaded before interacting with anything, 2) I needed to switch selenium’s focus to the iframe gsft\_main. This iframe contains a second html document which exists to contain the actual site data and web elements. This requires a very specific pattern of waits

- I kept running into random errors with the browser driver which would only appear the first couple of times it tries to run, so I needed to create a loop at runtime which will make sure that the browser actually gets started. I included a max restart limit to stop the loop if the thrown error is actually fatal

- the table cell’s class names don’t work