My program will crawl the UNAIDS website to scrape information about the National AIDS Spending Assessment (NASA) reports submitted by countries in the form of country reports.

The aim of this project is to identify word counts in these documents related to areas of interest outlined by UNAIDS by topic. Topics will include key populations and community-led response. Each of these topics will have a list of words associated with it, and each report that is scraped (excluding the non-English documents) will generate a count of the mention of each of these words. This program and its findings will be sent to my summer internship manager at UNAIDS headquarters in Geneva. They will use these findings to focus on underperforming countries.

The data source I intend to use is the UNAIDS database of NASA eeports found at <https://www.unaids.org/en/dataanalysis/knowyourresponse/nasacountryreports>). Because I will be ‘Crawling [and scraping] multiple pages in a site I haven’t used before’, this will receive a Challenge Score of 8.

I plan to use plotly to display both a map and a graph of the countries and their corresponding word-use rates. The intent is to tell the users the countries that appear (based simply on word counts) to be most closely aligned with the key words identified by UNAIDS as important points of focus. It may also make sense to represent them as a percentage of overall words or word combinations in order to ensure I am capturing useful information. Non-English reports will be excluded. For example, if one country writes rather short reports, then it makes sense that they would have a lower word count. I am still exploring this aspect of how to make the data generated as meaningful as possible; Any suggestions are greatly appreciated.

Thanks!