UID: 30353366726

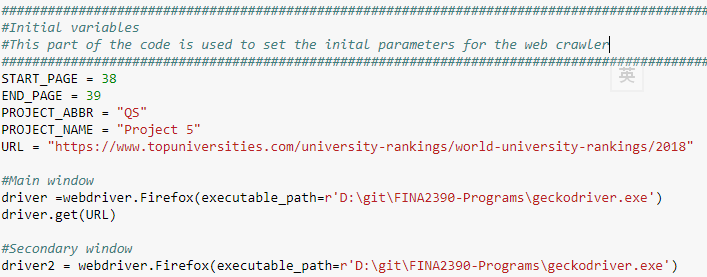
Name: Tsui Sai On

**Project 5.21 University Ranking**

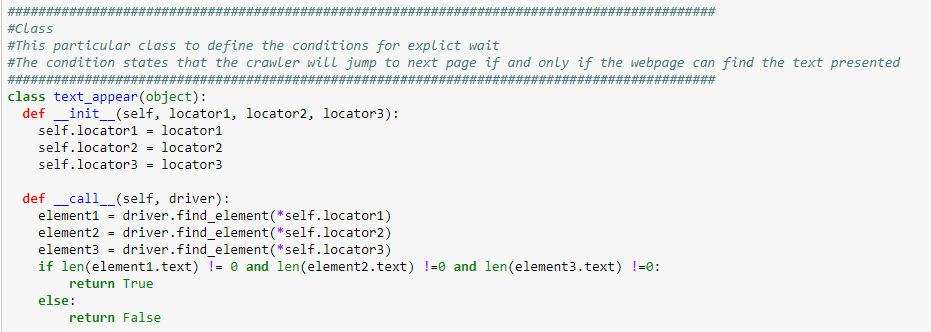
This web crawler contains various part to retrieve the data. I will use the QS crawler as an example to explain my code.



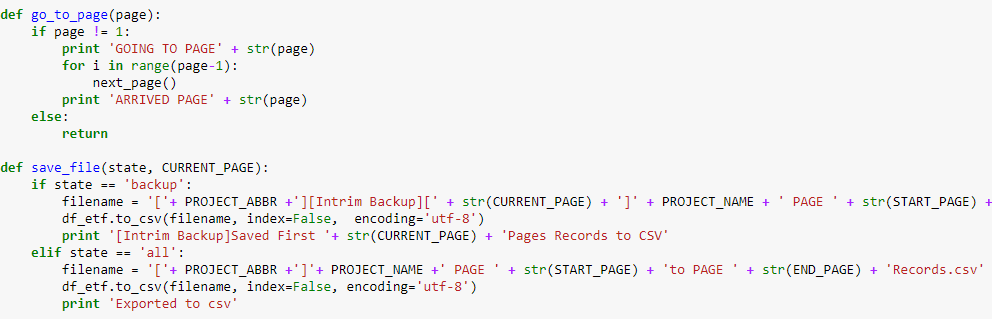
This part the packages import to perform the task. Nothing special about that



This part contains all the initial parameters for the program. This program allow partial retrieval of particular pages by adjusting the start page and end page parameters.



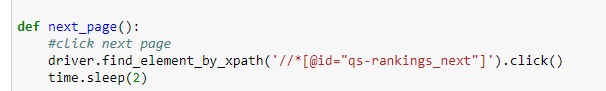
This part is the class definition of the program. This class “text\_appear” is to detect whether are text are loaded and presented in a certain page. There are 3 parameters of the class and these parameters denotes the 3 fields that we would like to assert. For most of the case, 3 parameters will be enough to get all the data we need this time. Of course, the program is scalable and able to detect more fields by adding extra parameters. This function is particularly useful to deal with pages using AJAX where the data and the table are not loaded at the same time



This part contains two function

The go\_to\_page function is used to navigate to start page if we are not starting at page 1. As QS website doesn’t provide button to jump to certain page, the implementation of this function is to press the next page button a numbers of times to reach the start page.

The save\_file function is used to export the dataframe retrieved from the website to a csv file. As there are maybe other language in the university name, we use UTF8 encoding to ensure all the fields can be displayed properly.



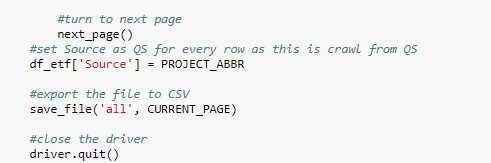
The next page function is used to press the button “next page” on the website, as it may require sometime to load the page, we added a time.sleep there to wait for a moment



As all the required fields may not appear on the table ranking, we need to open a new window to visit the detailed university profile one by one to get others fields. This function is used to open a new window to crawl the extra fields. The try and catch expression wrapped a webdriver wait, which is used to wait until the page is completely loaded and continue the crawling. With the text\_appear condition, we can ensure no erroe such as element not found will appear during runtime.

Again the condition inside “finally” is to ensure there is an element before we actually point to it. There is many case that the website may not have a unified XPATH so that error will occurs and stopped the crawling.





This is the main function of the program. There are two foreach loop in the in the function. The outer loop is used to loop over each page in the range and the inner loop is to loop over ever row in each page. In each row of the page, we extract the required fields by xpath and also extract the link of the detailed profile of the university and visit it by function “dig”. Upon finishing all the rows in a page, we call the next\_page function to turn the page to next page. After finishing all pages in the range, we set the source of every row to QS and export the csv file