**Web Scraping Project**

**Author:** Xiaoran Huang

**Aim:** This project is for collecting the retracted articles’ information and the articles that cite them.

**Install:** No need for specific installation, just make sure python and selenium already installed.

**How to run**: just click the web\_scraping.py

*The program provides 4 different modes as the user can decide to use which one by change the configuration file:*

**Mode 1: CITATION\_LIST\_ONLY**: This parameter can be either 1 or 0, if the user choose to set it as 1, the program will consider the user only want to do the citation list generation *(Please pay attention that if you want to use this mode there must be a retracted article list file that already existed in the directory path that same as the script)*

**MAX\_CITATION\_NUMBER:** This parameter can only be set to an integer that greater than 0, it shows that maximum title number (within the retraction list) that the user wants to search for citation list generation.

**Mode 2: RETRACTION\_LIST\_ONLY**: This parameter can be either 1 or 0, if the user choose to set it as 1, the program will consider the user only want to do the retraction list generation

**MAX\_RETRACTION\_NUMBER:** This parameter can only be set to an integer that greater than 0, it shows that maximum retracted article number that the user wants to get for retracted article information list generation. (For example, if it is setting to 500 means you want to have 500 rows of records in your generated retracted article list file)

**Mode 3:** If the above 2 parameters (**CITATION\_LIST\_ONLY and RETRACTION\_LIST\_ONLY**) both set to 0, the program will just execute both file’s generation without any upper bound limitation( in this case, the variable **MAX\_CITATION\_NUMBER** and **MAX\_RETRACTION\_NUMBER** won’t affect the program any more)

**Mode 4:** If the above 2 variables both set to 1. This is for the test only.

*Other parameters within the configuration file:*

**RETRACTION\_LIST\_NAME**: How the user wants to name the retraced article information list file (must end by .csv)

**CITATION\_LIST\_NAME**: How the user wants to name the cited article information list file (must end by .csv)

**WEB\_SCIENCE\_USERNAME & WEB\_SCIENCE\_PASSWORD:** User’s email address and password that used to log into Web of Science

**TITLE\_WITH\_DATE:** For the cited article information collection only, the user can decides to include the retraction/correction date within the article title for search (set to 1 then) or not(set to 0 then)

**CONTINUE\_WRITE:** Do you want to overwrite the existing information or not, if set to be 0, the program will clean all the existing information within the generated file and replace with the new data.(Only available for citation list only mode or retraction list only mode)