Run the python script:

1. Scraper for <https://training.gov.au/Search/SearchOrganisation>

How to run the script:

Command: python rtoscraper.py -o [excel\_filename]

ex: python rtoscraper.py -o rtofile.xlsx

By default the file result will be saved in "fileresult" folder.

we can change the folder location by changing the "RESFOLDER" value in settings.py

1. Scraper for <https://www.seek.com.au/jobs-in-education-training>

How to run the script:

To scrape this site, we have 4 steps:

* 1. set subclassification in setting.py

We can set subclassification by changing the "SUBCLASSIFICATIONS" value in setting.py

There are 16 subclassifications that you can include in the search result. Please check setting.py file.

* 1. run getseekurls.py script to to get all the urls related to the subclassification we have specified in setting.py

commad: python getseekurls.py

the file result will be saved in fileresult/urls.json

* 1. run seekscraper.py to scrape the data from urls.json that we have saved.

command: python seekscraper.py -s [start\_index] -e [last\_index]

ex: python seekscraper.py -s 1 -e 3000

It means, the script will scrape from index 1 to 3000.

We cannot retrieve all the data at once (if the search result more than 3000) because it will result in a failure in data retrieval. we are only allowed to retrieve data for 3000 records in one execution. for example, the number of urls that we will scrape is 7000, then we have to do 3 script executions:

* python seekscraper.py -s 1 -e 3000
* python seekscraper.py -s 3001 -e 6000
* python seekscraper.py -s 6001 -e 7000

NOTED: we have to restart the system after running seekscraper.py with the number of records above 3000. so if the amount of data is 7000, we need to restart the system 2 times.

the execution results of the four scripts above produce 4 json files stored in the "fileresult" folder (jsondata1.json, jsondata2.json, jsondata3.json).

1. run saveseekdata.py to save and join all seek data in excel format.

this script will join all jsondataxx.json to an excel file.

command: python saveseekdata.py -o [excel\_filename]

ex: python saveseekdata.py -o seekfile.xlsx