Web and Social Computing (IT752)
Lab Assignment 3

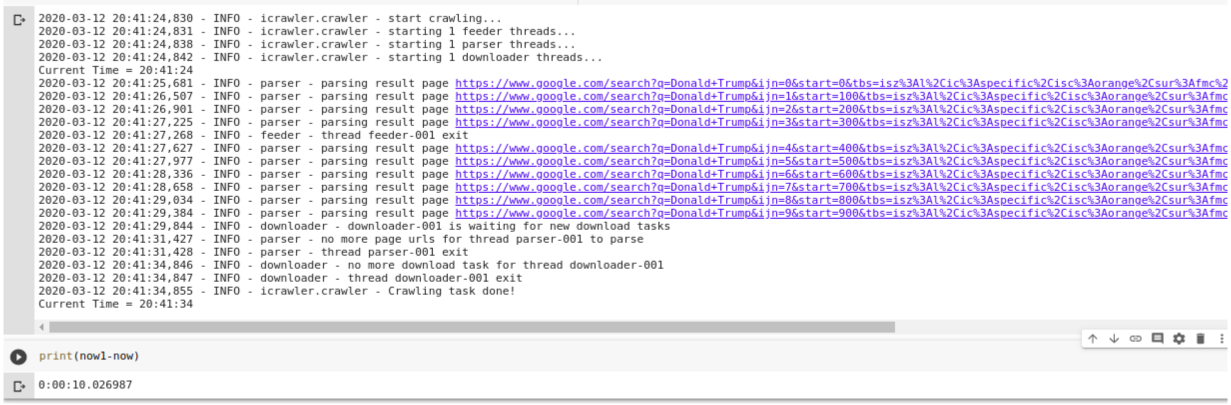
Submitted to : Dr. Sowmya Kamath
Submitted by : Shaikh Sahil Ahmed
Roll number: 192IT022

1. Crawler

A Web crawler, sometimes called a spider or spiderbot and often shortened to crawler, is an Internet bot that systematically browses the World Wide Web, typically for the purpose of Web indexing.
Library used as spider for web crawling. The basic web crawler is uploaded under crawler directory. The main.py file contains the crawler program where it is crawling the website URL: 'http://infotech.nitk.ac.in/'.

2. Single Threaded and Multi Threaded Crawler.

The package used is icrawler to implement crawler and imported googleImageCrawler. For single threaded, I have chosen the thread value as 1 and searched for the keyword “Donald Trump”, so related to the keyword the pages are crawled within the range of date 1 Jan 2016 to 31 Jan 2016.

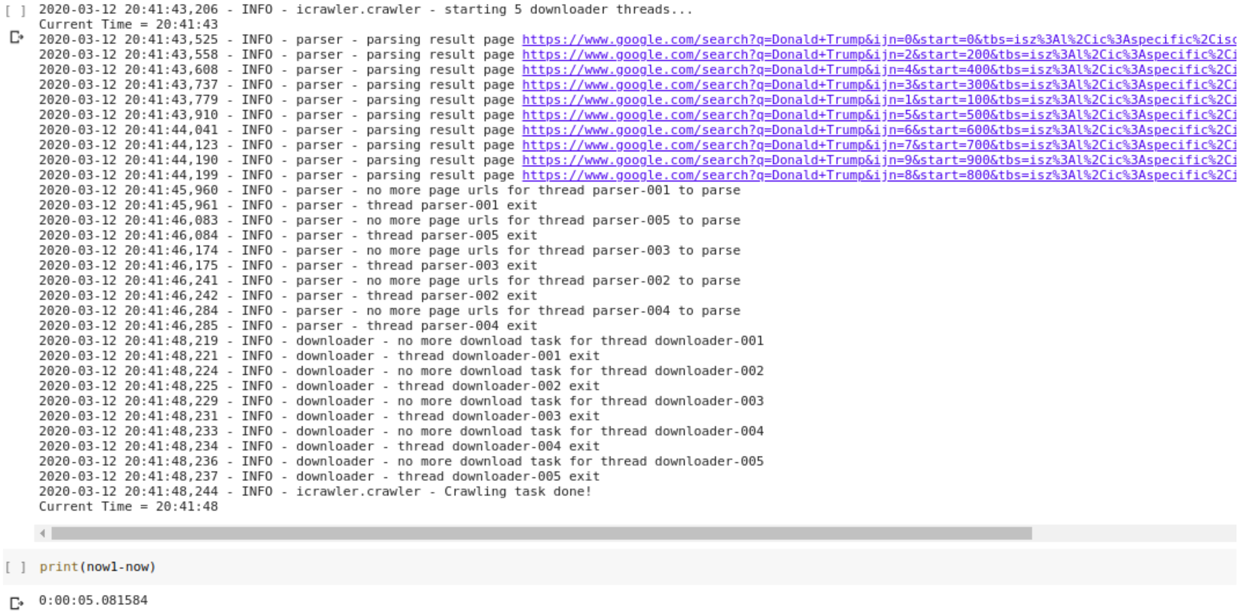


```
2020-03-12 20:41:24,830 - INFO - icrawler.crawler - start crawling...
2020-03-12 20:41:24,831 - INFO - icrawler.crawler - starting 1 feeder threads...
2020-03-12 20:41:24,838 - INFO - icrawler.crawler - starting 1 parser threads...
2020-03-12 20:41:24,842 - INFO - icrawler.crawler - starting 1 downloader threads...
Current Time = 20:41:24
2020-03-12 20:41:25,681 - INFO - parser - parsing result page https://www.google.com/search?q=Donald+Trump&ijn=0&start=0&tbs=isz%3Al%2Cic%3Aspecific%2Cisc%3Aorange%2Csur%3Afmc%2
2020-03-12 20:41:26,507 - INFO - parser - parsing result page https://www.google.com/search?q=Donald+Trump&ijn=1&start=100&tbs=isz%3Al%2Cic%3Aspecific%2Cisc%3Aorange%2Csur%3Afmc
2020-03-12 20:41:26,901 - INFO - parser - parsing result page https://www.google.com/search?q=Donald+Trump&ijn=2&start=200&tbs=isz%3Al%2Cic%3Aspecific%2Cisc%3Aorange%2Csur%3Afmc
2020-03-12 20:41:27,225 - INFO - parser - parsing result page https://www.google.com/search?q=Donald+Trump&ijn=3&start=300&tbs=isz%3Al%2Cic%3Aspecific%2Cisc%3Aorange%2Csur%3Afmc
2020-03-12 20:41:27,268 - INFO - feeder - thread feeder-001 exit
2020-03-12 20:41:27,627 - INFO - parser - parsing result page https://www.google.com/search?q=Donald+Trump&ijn=4&start=400&tbs=isz%3Al%2Cic%3Aspecific%2Cisc%3Aorange%2Csur%3Afmc
2020-03-12 20:41:27,977 - INFO - parser - parsing result page https://www.google.com/search?q=Donald+Trump&ijn=5&start=500&tbs=isz%3Al%2Cic%3Aspecific%2Cisc%3Aorange%2Csur%3Afmc
2020-03-12 20:41:28,336 - INFO - parser - parsing result page https://www.google.com/search?q=Donald+Trump&ijn=6&start=600&tbs=isz%3Al%2Cic%3Aspecific%2Cisc%3Aorange%2Csur%3Afmc
2020-03-12 20:41:28,658 - INFO - parser - parsing result page https://www.google.com/search?q=Donald+Trump&ijn=7&start=700&tbs=isz%3Al%2Cic%3Aspecific%2Cisc%3Aorange%2Csur%3Afmc
2020-03-12 20:41:29,034 - INFO - parser - parsing result page https://www.google.com/search?q=Donald+Trump&ijn=8&start=800&tbs=isz%3Al%2Cic%3Aspecific%2Cisc%3Aorange%2Csur%3Afmc
2020-03-12 20:41:29,384 - INFO - parser - parsing result page https://www.google.com/search?q=Donald+Trump&ijn=9&start=900&tbs=isz%3Al%2Cic%3Aspecific%2Cisc%3Aorange%2Csur%3Afmc
2020-03-12 20:41:29,844 - INFO - downloader - downloader-001 is waiting for new download tasks
2020-03-12 20:41:31,427 - INFO - parser - no more page urls for thread parser-001 to parse
2020-03-12 20:41:31,428 - INFO - parser - thread parser-001 exit
2020-03-12 20:41:34,846 - INFO - downloader - no more download task for thread downloader-001
2020-03-12 20:41:34,847 - INFO - downloader - thread downloader-001 exit
2020-03-12 20:41:34,855 - INFO - icrawler.crawler - Crawling task done!
Current Time = 20:41:34
```

```
print(nowl-now)
```

```
0:00:10.026987
```

So, execution time taken by the single threaded crawler is 10.02 seconds.
Now I go for the Multi Threaded Crawler. Same as above I have chosen the thread value as 5 and searched for the keyword “Donald Trump”, so related to the keyword the pages are crawled within the range of date 1 Jan 2016 to 31 Jan 2016.



```
2020-03-12 20:41:43,206 - INFO - icrawler.crawler - starting 5 downloader threads...
Current Time = 20:41:43
2020-03-12 20:41:43,525 - INFO - parser - parsing result page https://www.google.com/search?q=Donald+Trump&ijn=0&start=0&tbs=isz%3Al%2Cic%3Aspecific%2Cisc
2020-03-12 20:41:43,558 - INFO - parser - parsing result page https://www.google.com/search?q=Donald+Trump&ijn=2&start=200&tbs=isz%3Al%2Cic%3Aspecific%2Ci
2020-03-12 20:41:43,608 - INFO - parser - parsing result page https://www.google.com/search?q=Donald+Trump&ijn=4&start=400&tbs=isz%3Al%2Cic%3Aspecific%2Ci
2020-03-12 20:41:43,737 - INFO - parser - parsing result page https://www.google.com/search?q=Donald+Trump&ijn=3&start=300&tbs=isz%3Al%2Cic%3Aspecific%2Ci
2020-03-12 20:41:43,779 - INFO - parser - parsing result page https://www.google.com/search?q=Donald+Trump&ijn=1&start=100&tbs=isz%3Al%2Cic%3Aspecific%2Ci
2020-03-12 20:41:43,910 - INFO - parser - parsing result page https://www.google.com/search?q=Donald+Trump&ijn=5&start=500&tbs=isz%3Al%2Cic%3Aspecific%2Ci
2020-03-12 20:41:44,041 - INFO - parser - parsing result page https://www.google.com/search?q=Donald+Trump&ijn=6&start=600&tbs=isz%3Al%2Cic%3Aspecific%2Ci
2020-03-12 20:41:44,123 - INFO - parser - parsing result page https://www.google.com/search?q=Donald+Trump&ijn=7&start=700&tbs=isz%3Al%2Cic%3Aspecific%2Ci
2020-03-12 20:41:44,190 - INFO - parser - parsing result page https://www.google.com/search?q=Donald+Trump&ijn=9&start=900&tbs=isz%3Al%2Cic%3Aspecific%2Ci
2020-03-12 20:41:44,199 - INFO - parser - parsing result page https://www.google.com/search?q=Donald+Trump&ijn=8&start=800&tbs=isz%3Al%2Cic%3Aspecific%2Ci
2020-03-12 20:41:45,960 - INFO - parser - no more page urls for thread parser-001 to parse
2020-03-12 20:41:45,961 - INFO - parser - thread parser-001 exit
2020-03-12 20:41:46,083 - INFO - parser - no more page urls for thread parser-005 to parse
2020-03-12 20:41:46,084 - INFO - parser - thread parser-005 exit
2020-03-12 20:41:46,174 - INFO - parser - no more page urls for thread parser-003 to parse
2020-03-12 20:41:46,175 - INFO - parser - thread parser-003 exit
2020-03-12 20:41:46,241 - INFO - parser - no more page urls for thread parser-002 to parse
2020-03-12 20:41:46,242 - INFO - parser - thread parser-002 exit
2020-03-12 20:41:46,284 - INFO - parser - no more page urls for thread parser-004 to parse
2020-03-12 20:41:46,285 - INFO - parser - thread parser-004 exit
2020-03-12 20:41:48,219 - INFO - downloader - no more download task for thread downloader-001
2020-03-12 20:41:48,221 - INFO - downloader - thread downloader-001 exit
2020-03-12 20:41:48,224 - INFO - downloader - no more download task for thread downloader-002
2020-03-12 20:41:48,225 - INFO - downloader - thread downloader-002 exit
2020-03-12 20:41:48,229 - INFO - downloader - no more download task for thread downloader-003
2020-03-12 20:41:48,231 - INFO - downloader - thread downloader-003 exit
2020-03-12 20:41:48,233 - INFO - downloader - no more download task for thread downloader-004
2020-03-12 20:41:48,234 - INFO - downloader - thread downloader-004 exit
2020-03-12 20:41:48,236 - INFO - downloader - no more download task for thread downloader-005
2020-03-12 20:41:48,237 - INFO - downloader - thread downloader-005 exit
2020-03-12 20:41:48,244 - INFO - icrawler.crawler - Crawling task done!
Current Time = 20:41:48
```

```
print(nowl-now)
```

```
0:00:05.081584
```

So, execution time taken by the single threaded crawler is 10.02 seconds. Multi thread is more efficient.

2. Data structure used for indexing (optimal for searching, insertion...)

In the case of a crawler though, it would not be needed because starting from the root, you can maintain a "list" of visited URLs and every time you are about to follow a link, check if it has been encountered before. If it has not been encountered then add it to the list and follow it.

It doesn't have to literally be a list (i.e. array) though, it can be a dictionary or other data structure that would help in speeding up the search.
It could also be an SQL database or something like a key-value storage like redis. If you use something like this then all the indexing and querying will be handed for you by the database system with which you could communicate through a standard method (SQL, special API, other).

3. BFS and DFS crawler

BFS and DFS crawler is implemented and uploaded under the “bfs dfs” folder.
Giving an input, crawler crawls on the basis of bfs and dfs.