Analysis report for

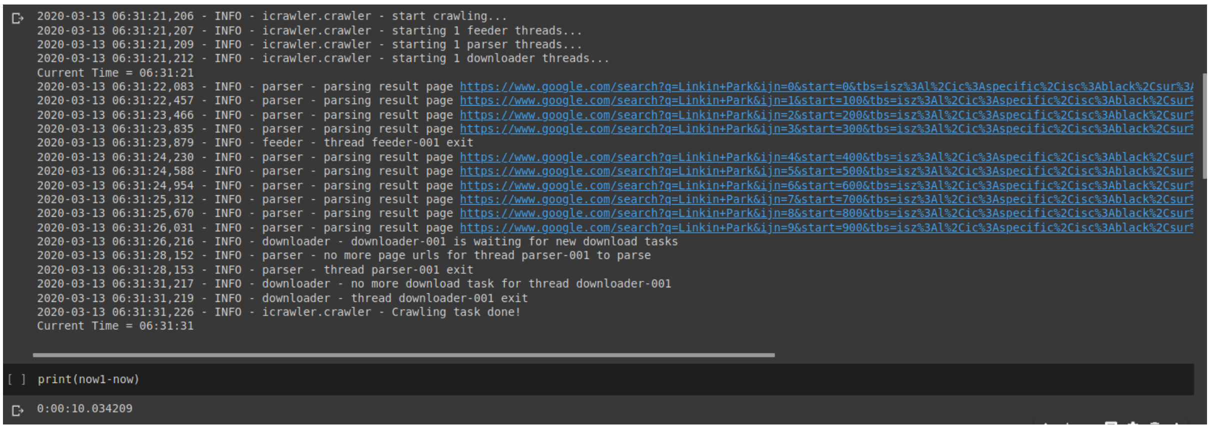
Web and Social Computing Assignment 2

By Sampat Kr Ghosh
Roll No. 192IT020

Introduction:

This reports explains the implementation of web crawling using icrawler package. Performance analysis is doen on number of threads used in multi-threaded crawling.

Performance comparison of sequential and multithreaded crawler:

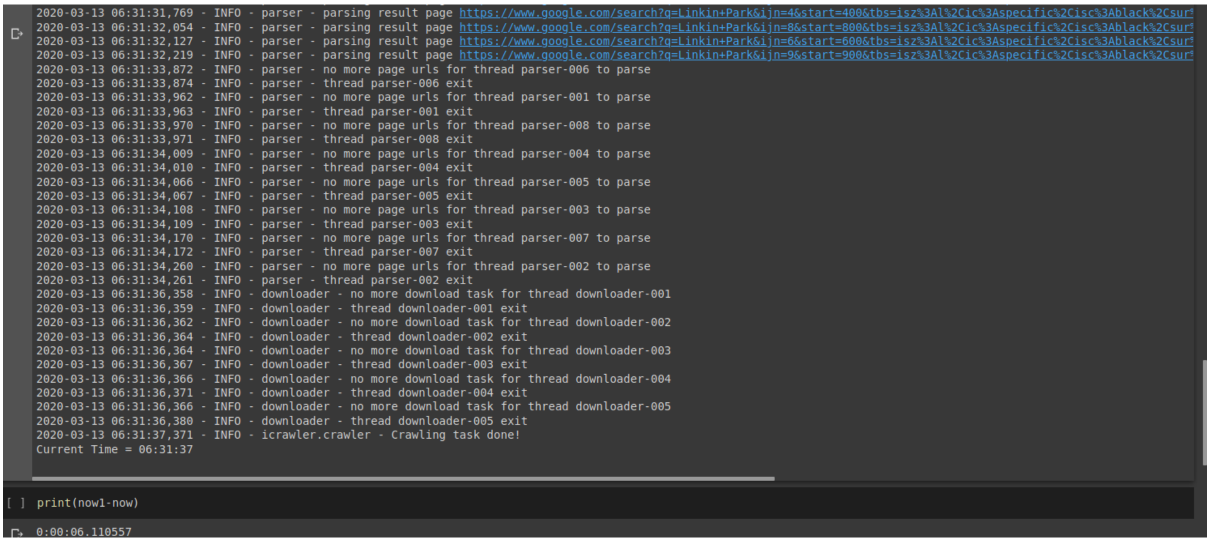


```
2020-03-13 06:31:21,206 - INFO - icrawler.crawler - start crawling...
2020-03-13 06:31:21,207 - INFO - icrawler.crawler - starting 1 feeder threads...
2020-03-13 06:31:21,209 - INFO - icrawler.crawler - starting 1 parser threads...
2020-03-13 06:31:21,212 - INFO - icrawler.crawler - starting 1 downloader threads...
Current Time = 06:31:21
2020-03-13 06:31:22,083 - INFO - parser - parsing result page https://www.google.com/search?q=Linkin+Park&ijn=0&start=0&tbs=isz%3Al%2Cic%3Aspecific%2Cisc%3Ablack%2Csur%3
2020-03-13 06:31:22,457 - INFO - parser - parsing result page https://www.google.com/search?q=Linkin+Park&ijn=1&start=100&tbs=isz%3Al%2Cic%3Aspecific%2Cisc%3Ablack%2Csur
2020-03-13 06:31:23,466 - INFO - parser - parsing result page https://www.google.com/search?q=Linkin+Park&ijn=2&start=200&tbs=isz%3Al%2Cic%3Aspecific%2Cisc%3Ablack%2Csur
2020-03-13 06:31:23,835 - INFO - parser - parsing result page https://www.google.com/search?q=Linkin+Park&ijn=3&start=300&tbs=isz%3Al%2Cic%3Aspecific%2Cisc%3Ablack%2Csur
2020-03-13 06:31:23,879 - INFO - feeder - thread feeder-001 exit
2020-03-13 06:31:24,230 - INFO - parser - parsing result page https://www.google.com/search?q=Linkin+Park&ijn=4&start=400&tbs=isz%3Al%2Cic%3Aspecific%2Cisc%3Ablack%2Csur
2020-03-13 06:31:24,588 - INFO - parser - parsing result page https://www.google.com/search?q=Linkin+Park&ijn=5&start=500&tbs=isz%3Al%2Cic%3Aspecific%2Cisc%3Ablack%2Csur
2020-03-13 06:31:24,954 - INFO - parser - parsing result page https://www.google.com/search?q=Linkin+Park&ijn=6&start=600&tbs=isz%3Al%2Cic%3Aspecific%2Cisc%3Ablack%2Csur
2020-03-13 06:31:25,312 - INFO - parser - parsing result page https://www.google.com/search?q=Linkin+Park&ijn=7&start=700&tbs=isz%3Al%2Cic%3Aspecific%2Cisc%3Ablack%2Csur
2020-03-13 06:31:25,670 - INFO - parser - parsing result page https://www.google.com/search?q=Linkin+Park&ijn=8&start=800&tbs=isz%3Al%2Cic%3Aspecific%2Cisc%3Ablack%2Csur
2020-03-13 06:31:26,031 - INFO - parser - parsing result page https://www.google.com/search?q=Linkin+Park&ijn=9&start=900&tbs=isz%3Al%2Cic%3Aspecific%2Cisc%3Ablack%2Csur
2020-03-13 06:31:26,216 - INFO - downloader - downloader-001 is waiting for new download tasks
2020-03-13 06:31:28,152 - INFO - parser - no more page urls for thread parser-001 to parse
2020-03-13 06:31:28,153 - INFO - parser - thread parser-001 exit
2020-03-13 06:31:31,217 - INFO - downloader - no more download task for thread downloader-001
2020-03-13 06:31:31,219 - INFO - downloader - thread downloader-001 exit
2020-03-13 06:31:31,226 - INFO - icrawler.crawler - Crawling task done!
Current Time = 06:31:31
```

```
[ ] print(now1-now)
```

```
0:00:10.034209
```

Above figure shows sequential crawling using icrawler



```
2020-03-13 06:31:31,769 - INFO - parser - parsing result page https://www.google.com/search?q=Linkin+Park&ijn=4&start=400&tbs=isz%3Al%2Cic%3Aspecific%2Cisc%3Ablack%2Csur
2020-03-13 06:31:32,054 - INFO - parser - parsing result page https://www.google.com/search?q=Linkin+Park&ijn=8&start=800&tbs=isz%3Al%2Cic%3Aspecific%2Cisc%3Ablack%2Csur
2020-03-13 06:31:32,127 - INFO - parser - parsing result page https://www.google.com/search?q=Linkin+Park&ijn=6&start=600&tbs=isz%3Al%2Cic%3Aspecific%2Cisc%3Ablack%2Csur
2020-03-13 06:31:32,219 - INFO - parser - parsing result page https://www.google.com/search?q=Linkin+Park&ijn=9&start=900&tbs=isz%3Al%2Cic%3Aspecific%2Cisc%3Ablack%2Csur
2020-03-13 06:31:33,872 - INFO - parser - no more page urls for thread parser-006 to parse
2020-03-13 06:31:33,874 - INFO - parser - thread parser-006 exit
2020-03-13 06:31:33,962 - INFO - parser - no more page urls for thread parser-001 to parse
2020-03-13 06:31:33,963 - INFO - parser - thread parser-001 exit
2020-03-13 06:31:33,970 - INFO - parser - no more page urls for thread parser-008 to parse
2020-03-13 06:31:33,971 - INFO - parser - thread parser-008 exit
2020-03-13 06:31:34,009 - INFO - parser - no more page urls for thread parser-004 to parse
2020-03-13 06:31:34,010 - INFO - parser - thread parser-004 exit
2020-03-13 06:31:34,066 - INFO - parser - no more page urls for thread parser-005 to parse
2020-03-13 06:31:34,067 - INFO - parser - thread parser-005 exit
2020-03-13 06:31:34,108 - INFO - parser - no more page urls for thread parser-003 to parse
2020-03-13 06:31:34,109 - INFO - parser - thread parser-003 exit
2020-03-13 06:31:34,170 - INFO - parser - no more page urls for thread parser-007 to parse
2020-03-13 06:31:34,172 - INFO - parser - thread parser-007 exit
2020-03-13 06:31:34,260 - INFO - parser - no more page urls for thread parser-002 to parse
2020-03-13 06:31:34,261 - INFO - parser - thread parser-002 exit
2020-03-13 06:31:36,358 - INFO - downloader - no more download task for thread downloader-001
2020-03-13 06:31:36,359 - INFO - downloader - thread downloader-001 exit
2020-03-13 06:31:36,362 - INFO - downloader - no more download task for thread downloader-002
2020-03-13 06:31:36,364 - INFO - downloader - thread downloader-002 exit
2020-03-13 06:31:36,364 - INFO - downloader - no more download task for thread downloader-003
2020-03-13 06:31:36,367 - INFO - downloader - thread downloader-003 exit
2020-03-13 06:31:36,366 - INFO - downloader - no more download task for thread downloader-004
2020-03-13 06:31:36,371 - INFO - downloader - thread downloader-004 exit
2020-03-13 06:31:36,366 - INFO - downloader - no more download task for thread downloader-005
2020-03-13 06:31:36,380 - INFO - downloader - thread downloader-005 exit
2020-03-13 06:31:37,371 - INFO - icrawler.crawler - Crawling task done!
Current Time = 06:31:37
```

```
[ ] print(now1-now)
```

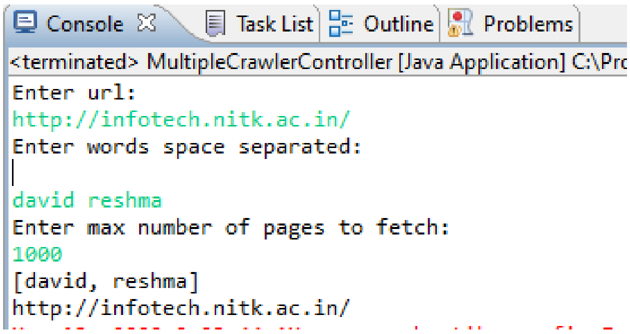
```
0:00:06.110557
```

Above figure shows multi-threaded crawling.

From the time taken, we can say that as the number if threads increases, the time decreases for crawling.

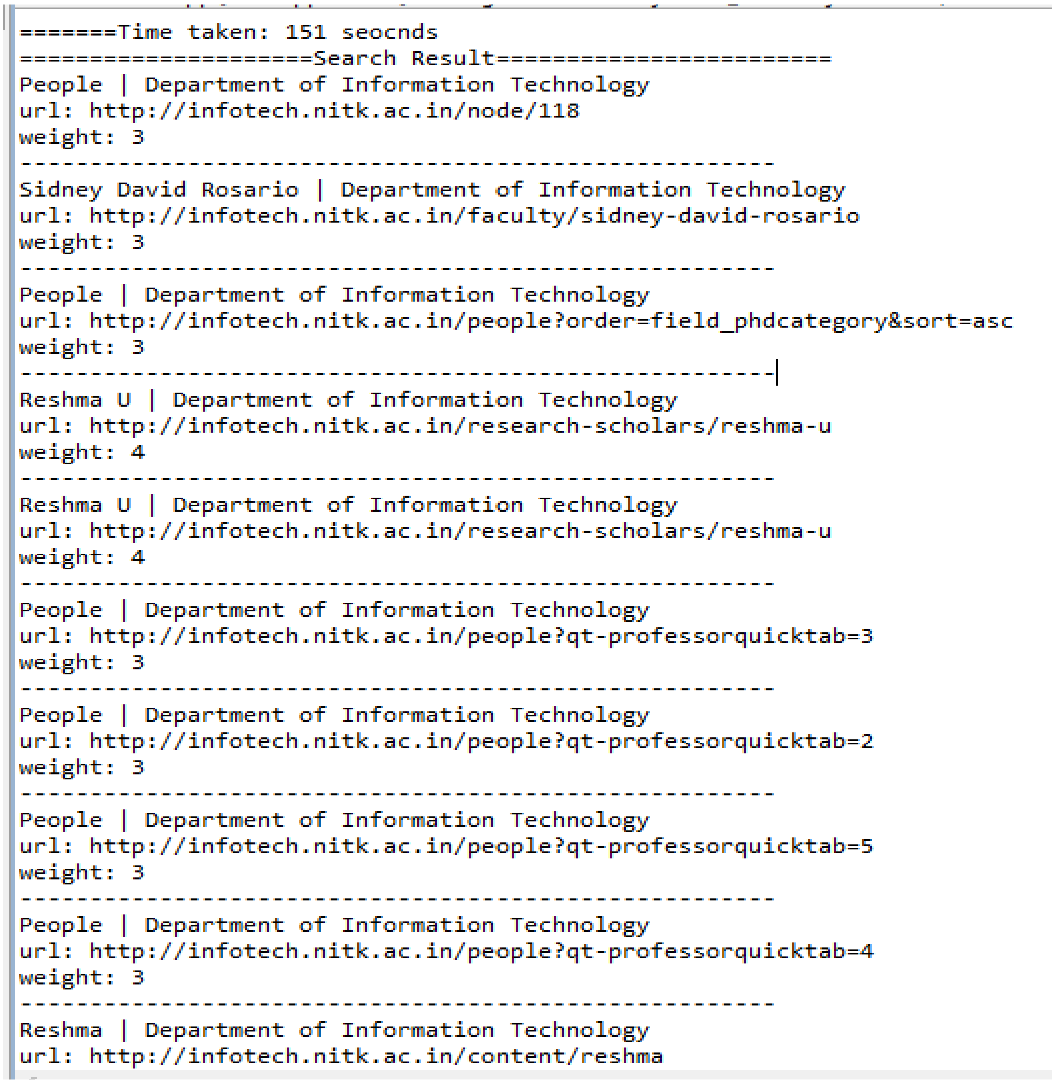
Below are the screenshot where
Input: Starting URL, list of words, max number of pages to be download.

Try each one of the below variants, while designing crawling strategies.
a. Crawl only from pages whose body text includes one of the words in the list.
b. Crawl only from pages whose title includes one of the words in the list.
c. Give priority to page with most words from list (either most different words or most occurrences of words)
Output: Search results page



```
<terminated> MultipleCrawlerController [Java Application] C:\Prc
Enter url:
http://infotech.nitk.ac.in/
Enter words space separated:

david reshma
Enter max number of pages to fetch:
1000
[david, reshma]
http://infotech.nitk.ac.in/
```



```
=======Time taken: 151 seocnds
=====================Search Result========================
People | Department of Information Technology
url: http://infotech.nitk.ac.in/node/118
weight: 3
------------------------------------------------------
Sidney David Rosario | Department of Information Technology
url: http://infotech.nitk.ac.in/faculty/sidney-david-rosario
weight: 3
------------------------------------------------------
People | Department of Information Technology
url: http://infotech.nitk.ac.in/people?order=field_phdcategory&sort=asc
weight: 3
------------------------------------------------------
Reshma U | Department of Information Technology
url: http://infotech.nitk.ac.in/research-scholars/reshma-u
weight: 4
------------------------------------------------------
Reshma U | Department of Information Technology
url: http://infotech.nitk.ac.in/research-scholars/reshma-u
weight: 4
------------------------------------------------------
People | Department of Information Technology
url: http://infotech.nitk.ac.in/people?qt-professorquicktab=3
weight: 3
------------------------------------------------------
People | Department of Information Technology
url: http://infotech.nitk.ac.in/people?qt-professorquicktab=2
weight: 3
------------------------------------------------------
People | Department of Information Technology
url: http://infotech.nitk.ac.in/people?qt-professorquicktab=5
weight: 3
------------------------------------------------------
People | Department of Information Technology
url: http://infotech.nitk.ac.in/people?qt-professorquicktab=4
weight: 3
------------------------------------------------------
Reshma | Department of Information Technology
url: http://infotech.nitk.ac.in/content/reshma
```