Analysis report for

Web and Social Computing Assignment 2

By Niranjan

Roll No. 192IT014

Crawler4j uses the BFS Strategy for crawling.The package only supports BFS traversal and implementation of DFS inside the package was not present. Therefore all results observed were using BFS algorithm as traversal strategy.

Politeness delay=100ms

Maximum pages to fetch=500

Maximum Depth of Crawling=10

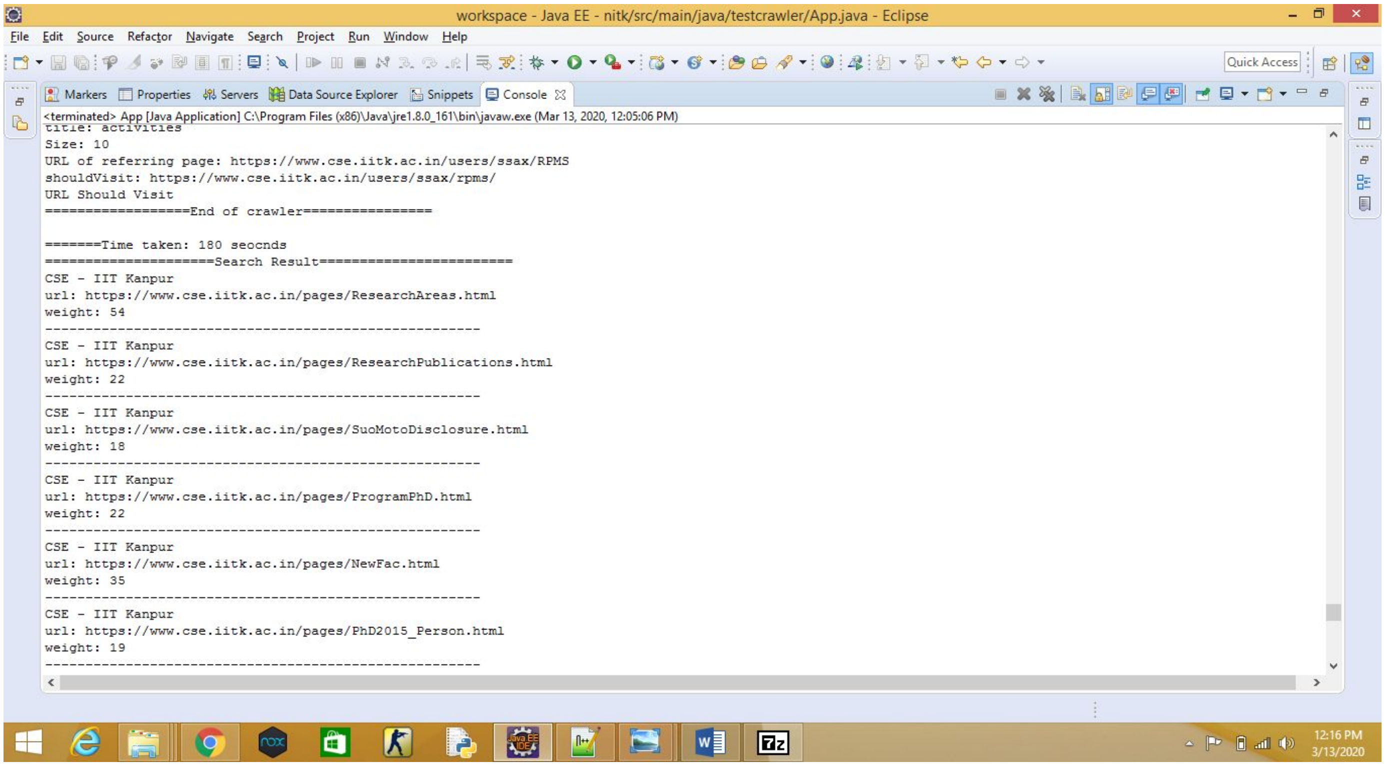
1.1Sequential crawling:

No.of Crawlers=1

Seed url="http://www.cse.iitk.ac.in/"

Time taken to crawl=180sec

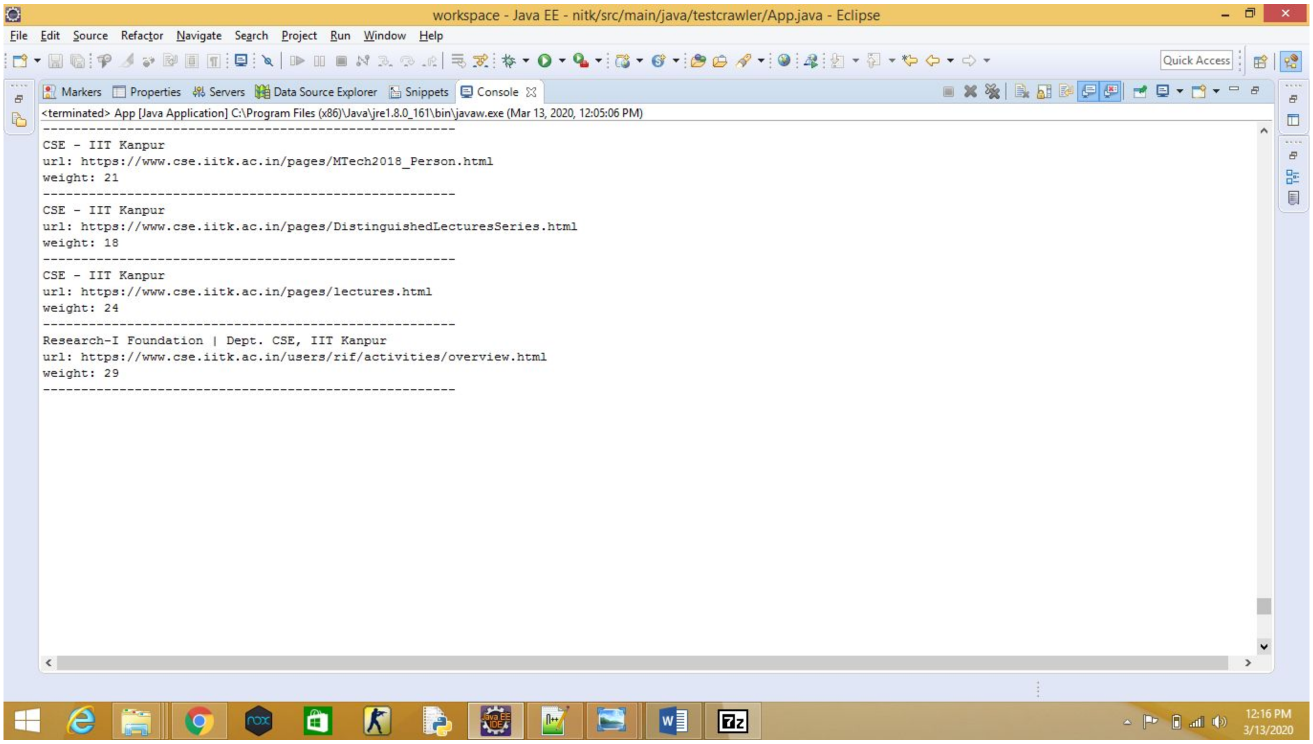
Result:



workspace - Java EE - nitk/src/main/java/testcrawler/App.java - Eclipse
File Edit Source Refactor Navigate Search Project Run Window Help
Quick Access
Markers Properties Servers Data Source Explorer Snippets Console
12:16 PM
3/13/2020

```
<terminated> App [Java Application] C:\Program Files (x86)\Java\jre1.8.0_161\bin\javaw.exe (Mar 13, 2020, 12:05:06 PM)
title: activities
Size: 10
URL of referring page: https://www.cse.iitk.ac.in/users/ssax/RPMS
shouldVisit: https://www.cse.iitk.ac.in/users/ssax/rpms/
URL Should Visit
==================End of crawler================

=======Time taken: 180 seocnds
=====================Search Result========================
CSE - IIT Kanpur
url: https://www.cse.iitk.ac.in/pages/ResearchAreas.html
weight: 54
------------------------------------------------------
CSE - IIT Kanpur
url: https://www.cse.iitk.ac.in/pages/ResearchPublications.html
weight: 22
------------------------------------------------------
CSE - IIT Kanpur
url: https://www.cse.iitk.ac.in/pages/SuoMotoDisclosure.html
weight: 18
------------------------------------------------------
CSE - IIT Kanpur
url: https://www.cse.iitk.ac.in/pages/ProgramPhD.html
weight: 22
------------------------------------------------------
CSE - IIT Kanpur
url: https://www.cse.iitk.ac.in/pages/NewFac.html
weight: 35
------------------------------------------------------
CSE - IIT Kanpur
url: https://www.cse.iitk.ac.in/pages/PhD2015_Person.html
weight: 19
------------------------------------------------------
```


workspace - Java EE - nitk/src/main/java/testcrawler/App.java - Eclipse
File Edit Source Refactor Navigate Search Project Run Window Help
Quick Access
Markers
Properties
Servers
Data Source Explorer
Snippets
Console
12:16 PM
3/13/2020

<terminated> App [Java Application] C:\Program Files (x86)\Java\jre1.8.0_161\bin\javaw.exe (Mar 13, 2020, 12:05:06 PM)

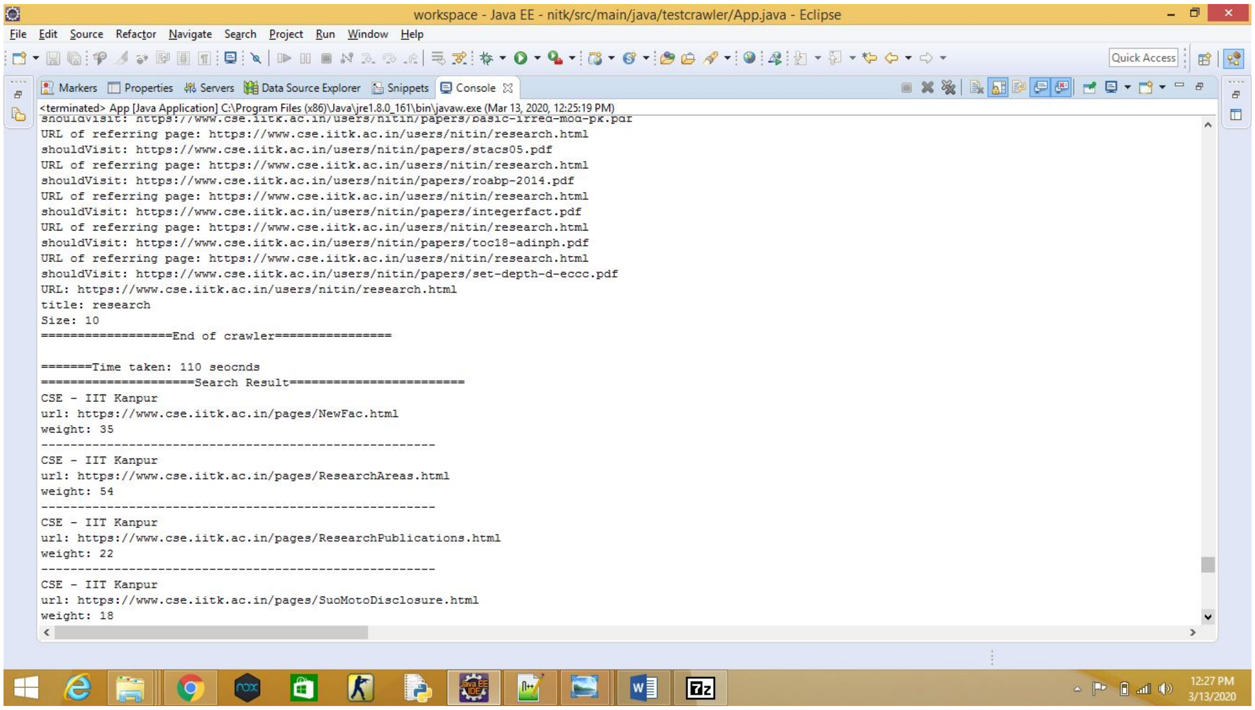
```
------------------------------------------------------
CSE - IIT Kanpur
url: https://www.cse.iitk.ac.in/pages/MTech2018_Person.html
weight: 21
------------------------------------------------------
CSE - IIT Kanpur
url: https://www.cse.iitk.ac.in/pages/DistinguishedLecturesSeries.html
weight: 18
------------------------------------------------------
CSE - IIT Kanpur
url: https://www.cse.iitk.ac.in/pages/lectures.html
weight: 24
------------------------------------------------------
Research-I Foundation | Dept. CSE, IIT Kanpur
url: https://www.cse.iitk.ac.in/users/rif/activities/overview.html
weight: 29
------------------------------------------------------
```

1.2 No.of Crawlers=2

Seed url="http://www.cse.iitk.ac.in/"

Time taken to crawl=110s

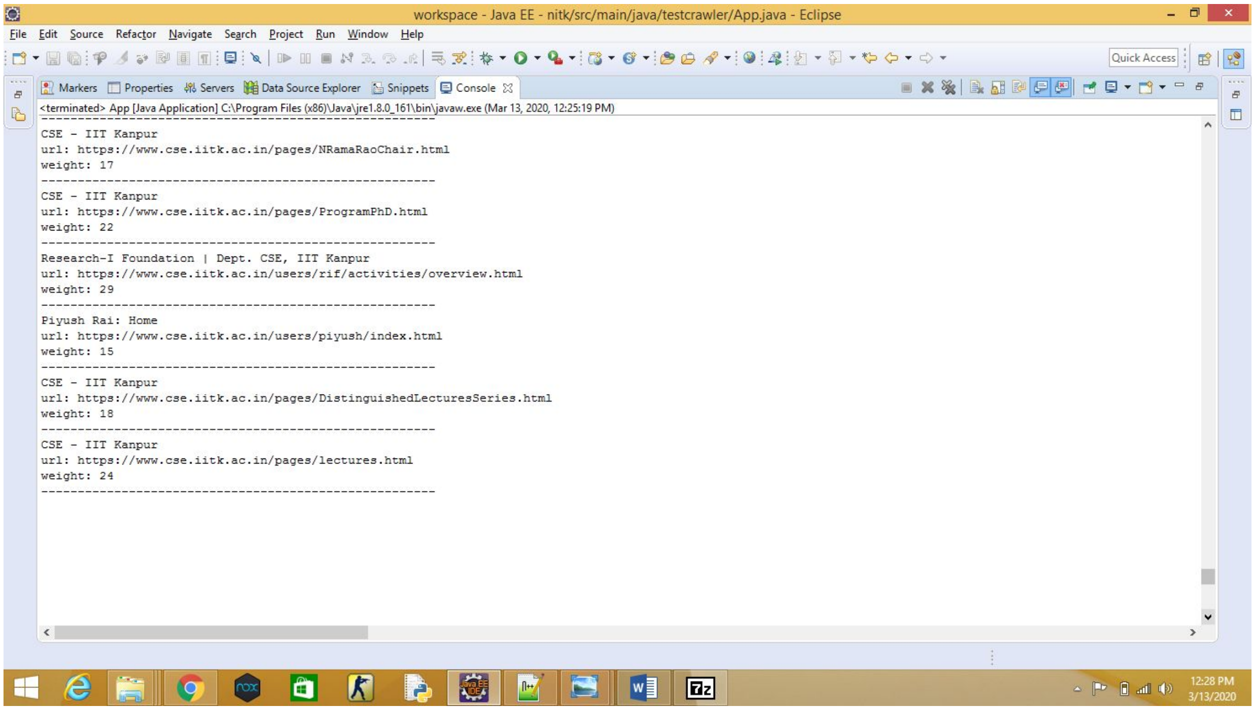
Result:



workspace - Java EE - nitk/src/main/java/testcrawler/App.java - Eclipse
File Edit Source Refactor Navigate Search Project Run Window Help
Quick Access
Markers Properties Servers Data Source Explorer Snippets Console
12:27 PM
3/13/2020

```
<terminated> App [Java Application] C:\Program Files (x86)\Java\jre1.8.0_161\bin\javaw.exe (Mar 13, 2020, 12:25:19 PM)
shouldVisit: https://www.cse.iitk.ac.in/users/nitin/papers/basic-irred-mod-pk.pdf
URL of referring page: https://www.cse.iitk.ac.in/users/nitin/research.html
shouldVisit: https://www.cse.iitk.ac.in/users/nitin/papers/stacs05.pdf
URL of referring page: https://www.cse.iitk.ac.in/users/nitin/research.html
shouldVisit: https://www.cse.iitk.ac.in/users/nitin/papers/roabp-2014.pdf
URL of referring page: https://www.cse.iitk.ac.in/users/nitin/research.html
shouldVisit: https://www.cse.iitk.ac.in/users/nitin/papers/integerfact.pdf
URL of referring page: https://www.cse.iitk.ac.in/users/nitin/research.html
shouldVisit: https://www.cse.iitk.ac.in/users/nitin/papers/toc18-adinph.pdf
URL of referring page: https://www.cse.iitk.ac.in/users/nitin/research.html
shouldVisit: https://www.cse.iitk.ac.in/users/nitin/papers/set-depth-d-eccc.pdf
URL: https://www.cse.iitk.ac.in/users/nitin/research.html
title: research
Size: 10
==================End of crawler================

=======Time taken: 110 seocnds
=====================Search Result========================
CSE - IIT Kanpur
url: https://www.cse.iitk.ac.in/pages/NewFac.html
weight: 35
------------------------------------------------------
CSE - IIT Kanpur
url: https://www.cse.iitk.ac.in/pages/ResearchAreas.html
weight: 54
------------------------------------------------------
CSE - IIT Kanpur
url: https://www.cse.iitk.ac.in/pages/ResearchPublications.html
weight: 22
------------------------------------------------------
CSE - IIT Kanpur
url: https://www.cse.iitk.ac.in/pages/SuoMotoDisclosure.html
weight: 18
```



workspace - Java EE - nitk/src/main/java/testcrawler/App.java - Eclipse
File Edit Source Refactor Navigate Search Project Run Window Help
Quick Access
Markers Properties Servers Data Source Explorer Snippets Console
12:28 PM
3/13/2020

```
<terminated> App [Java Application] C:\Program Files (x86)\Java\jre1.8.0_161\bin\javaw.exe (Mar 13, 2020, 12:25:19 PM)
------------------------------------------------------
CSE - IIT Kanpur
url: https://www.cse.iitk.ac.in/pages/NRamaRaoChair.html
weight: 17
------------------------------------------------------
CSE - IIT Kanpur
url: https://www.cse.iitk.ac.in/pages/ProgramPhD.html
weight: 22
------------------------------------------------------
Research-I Foundation | Dept. CSE, IIT Kanpur
url: https://www.cse.iitk.ac.in/users/rif/activities/overview.html
weight: 29
------------------------------------------------------
Piyush Rai: Home
url: https://www.cse.iitk.ac.in/users/piyush/index.html
weight: 15
------------------------------------------------------
CSE - IIT Kanpur
url: https://www.cse.iitk.ac.in/pages/DistinguishedLecturesSeries.html
weight: 18
------------------------------------------------------
CSE - IIT Kanpur
url: https://www.cse.iitk.ac.in/pages/lectures.html
weight: 24
------------------------------------------------------
```

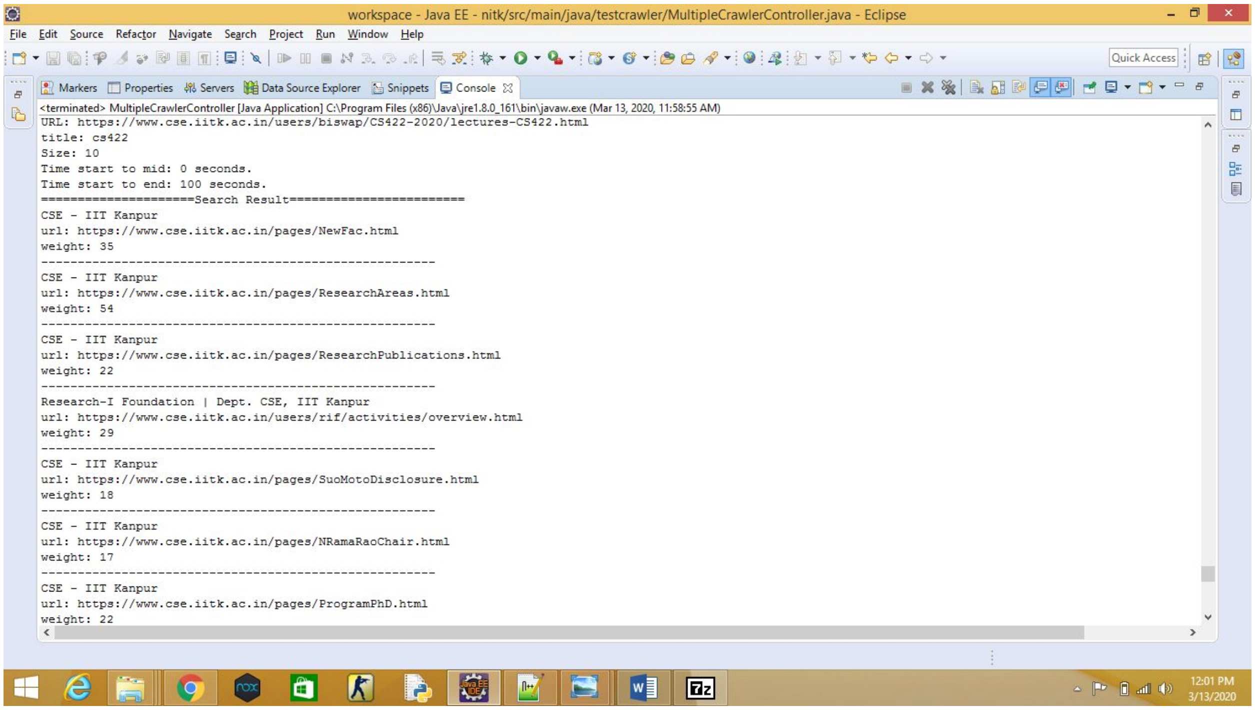
2.1Multithreaded crawling:

No.of Threads=5

Seed url=""http://www.cse.iitk.ac.in/""

Time taken to crawl=100sec

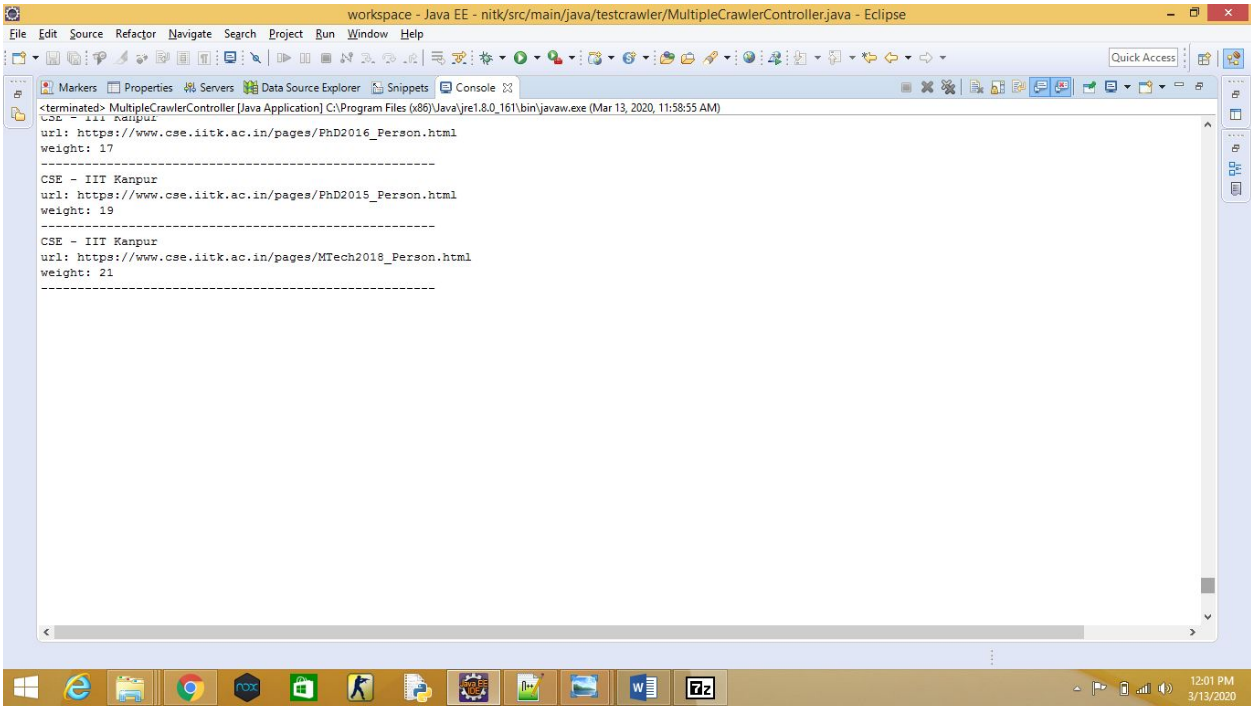
Result:



workspace - Java EE - nitk/src/main/java/testcrawler/MultipleCrawlerController.java - Eclipse

File Edit Source Refactor Navigate Search Project Run Window Help
Quick Access
Markers Properties Servers Data Source Explorer Snippets Console
12:01 PM
3/13/2020

```
<terminated> MultipleCrawlerController [Java Application] C:\Program Files (x86)\Java\jre1.8.0_161\bin\javaw.exe (Mar 13, 2020, 11:58:55 AM)
URL: https://www.cse.iitk.ac.in/users/biswap/CS422-2020/lectures-CS422.html
title: cs422
Size: 10
Time start to mid: 0 seconds.
Time start to end: 100 seconds.
=====================Search Result========================
CSE - IIT Kanpur
url: https://www.cse.iitk.ac.in/pages/NewFac.html
weight: 35
------------------------------------------------------
CSE - IIT Kanpur
url: https://www.cse.iitk.ac.in/pages/ResearchAreas.html
weight: 54
------------------------------------------------------
CSE - IIT Kanpur
url: https://www.cse.iitk.ac.in/pages/ResearchPublications.html
weight: 22
------------------------------------------------------
Research-I Foundation | Dept. CSE, IIT Kanpur
url: https://www.cse.iitk.ac.in/users/rif/activities/overview.html
weight: 29
------------------------------------------------------
CSE - IIT Kanpur
url: https://www.cse.iitk.ac.in/pages/SuoMotoDisclosure.html
weight: 18
------------------------------------------------------
CSE - IIT Kanpur
url: https://www.cse.iitk.ac.in/pages/NRamaRaoChair.html
weight: 17
------------------------------------------------------
CSE - IIT Kanpur
url: https://www.cse.iitk.ac.in/pages/ProgramPhD.html
weight: 22
```



workspace - Java EE - nitk/src/main/java/testcrawler/MultipleCrawlerController.java - Eclipse

File Edit Source Refactor Navigate Search Project Run Window Help
Quick Access
Markers Properties Servers Data Source Explorer Snippets Console
12:01 PM
3/13/2020

```
<terminated> MultipleCrawlerController [Java Application] C:\Program Files (x86)\Java\jre1.8.0_161\bin\javaw.exe (Mar 13, 2020, 11:58:55 AM)
CSE - IIT Kanpur
url: https://www.cse.iitk.ac.in/pages/PhD2016_Person.html
weight: 17
------------------------------------------------------
CSE - IIT Kanpur
url: https://www.cse.iitk.ac.in/pages/PhD2015_Person.html
weight: 19
------------------------------------------------------
CSE - IIT Kanpur
url: https://www.cse.iitk.ac.in/pages/MTech2018_Person.html
weight: 21
------------------------------------------------------
```

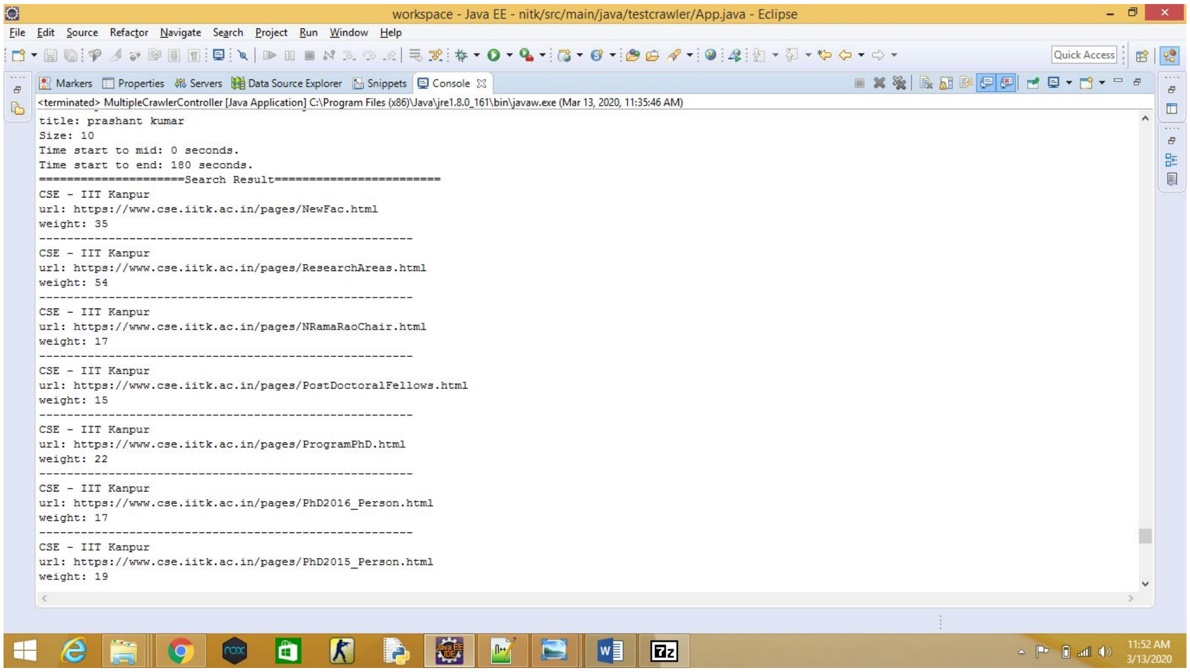
2.2Multithreaded crawling:

No.of Threads=3

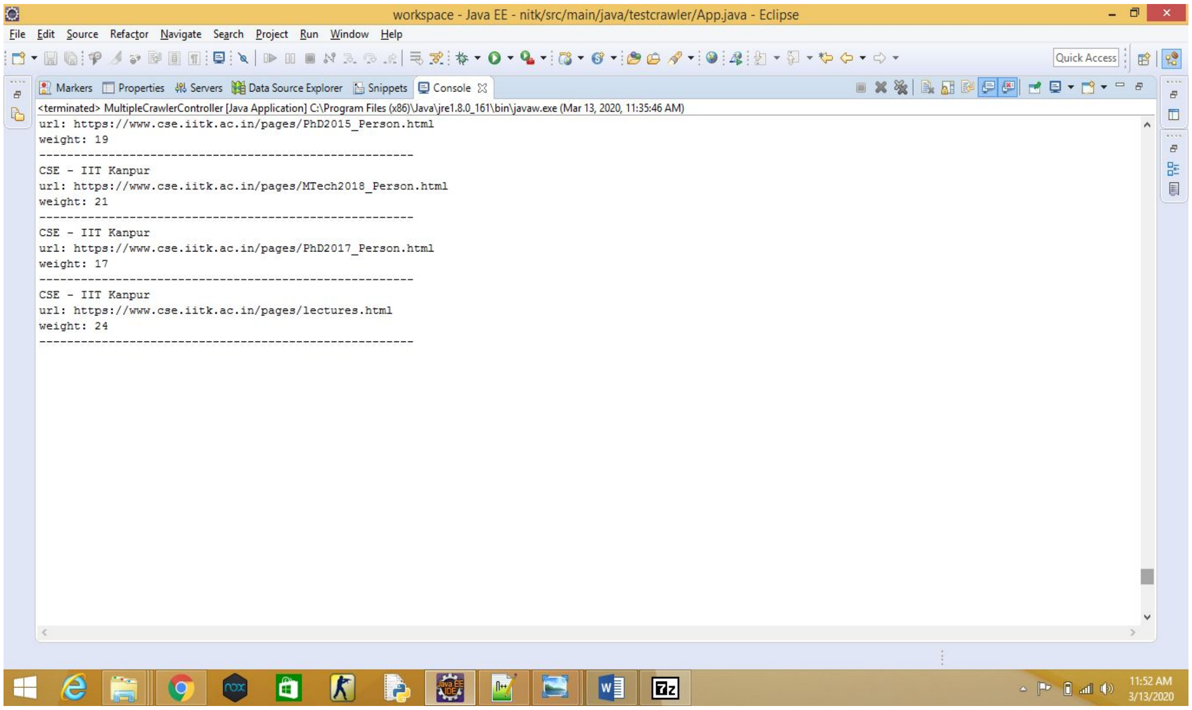
Seed url=""http://www.cse.iitk.ac.in/""

Time taken to crawl=180s.

RESULT:


workspace - Java EE - nitk/src/main/java/testcrawler/App.java - Eclipse
File Edit Source Refactor Navigate Search Project Run Window Help
Quick Access
Markers Properties Servers Data Source Explorer Snippets Console
11:52 AM
3/13/2020

```
<terminated> MultipleCrawlerController [Java Application] C:\Program Files (x86)\Java\jre1.8.0_161\bin\javaw.exe (Mar 13, 2020, 11:35:46 AM)
title: prashant kumar
Size: 10
Time start to mid: 0 seconds.
Time start to end: 180 seconds.
=====================Search Result========================
CSE - IIT Kanpur
url: https://www.cse.iitk.ac.in/pages/NewFac.html
weight: 35
------------------------------------------------------
CSE - IIT Kanpur
url: https://www.cse.iitk.ac.in/pages/ResearchAreas.html
weight: 54
------------------------------------------------------
CSE - IIT Kanpur
url: https://www.cse.iitk.ac.in/pages/NRamaRaoChair.html
weight: 17
------------------------------------------------------
CSE - IIT Kanpur
url: https://www.cse.iitk.ac.in/pages/PostDoctoralFellows.html
weight: 15
------------------------------------------------------
CSE - IIT Kanpur
url: https://www.cse.iitk.ac.in/pages/ProgramPhD.html
weight: 22
------------------------------------------------------
CSE - IIT Kanpur
url: https://www.cse.iitk.ac.in/pages/PhD2016_Person.html
weight: 17
------------------------------------------------------
CSE - IIT Kanpur
url: https://www.cse.iitk.ac.in/pages/PhD2015_Person.html
weight: 19
```


workspace - Java EE - nitk/src/main/java/testcrawler/App.java - Eclipse
File Edit Source Refactor Navigate Search Project Run Window Help
Quick Access
Markers Properties Servers Data Source Explorer Snippets Console
11:52 AM
3/13/2020

```
<terminated> MultipleCrawlerController [Java Application] C:\Program Files (x86)\Java\jre1.8.0_161\bin\javaw.exe (Mar 13, 2020, 11:35:46 AM)
url: https://www.cse.iitk.ac.in/pages/PhD2015_Person.html
weight: 19
------------------------------------------------------
CSE - IIT Kanpur
url: https://www.cse.iitk.ac.in/pages/MTech2018_Person.html
weight: 21
------------------------------------------------------
CSE - IIT Kanpur
url: https://www.cse.iitk.ac.in/pages/PhD2017_Person.html
weight: 17
------------------------------------------------------
CSE - IIT Kanpur
url: https://www.cse.iitk.ac.in/pages/lectures.html
weight: 24
------------------------------------------------------
```

I didn’t conduct the experiment by increasing the number of thread as the time is not getting affected much as the number of threads are increasing.

My Observation:

By increasing number of crawlers in sequential crawling the time reduced to 100sec from 1809 sec.

In multithreading crawler,number of crawlers is taken as 1.

The time for 3 threads is 180sec and for 5 threads it is 110sec.

The time in multithreading in comparison with sequential with no. of crawlers as1 is 180 sec

And 3 threads is 180 sec. So time is not changing.

With 1 crawler and changing threads speed up can be achieved.

3.Data structure used for indexing :

The data structures used are namely the following :

> **3.1Set** : A set, by definition, contains unique entries. In other words, no duplicates. All the pages we visit will be unique (or at least their URL will be unique). We can enforce this idea by choosing the right data structure, in this case a set
>
> **3.2List** : This is just storing a bunch of URLs we have visited. When the crawler visits a page it collects all the unique URLs on that page through above mentioned set data structure and we just append all of them from all the pages to this list. Lists have special methods that sets ordinarily do not, such as adding an entry to the end of a list or adding an entry to the beginning of a list.

4.Searching using list of words:

The input parameters are as follows:

> Seed url("https://www.cse.iitk.ac.in/")
>
> maximum pages to fetch-500

The weight of words like research and faculty is used and cumulative value is displayed in each strategy mentioned above.

Weight is the number of times that word came in that page.

From the search result performance also advantages of concurrency is identifiable. That’s why to design an efficient strategy for crawling an optimal number of threads are chosen and that performs the best with full utilization.