USERDOC_M4.PDF By Nialls Chavez
How to use Crawler.java
Preface:

This is the USERDOC for the webalyzer and crawler files. These programs provide the ability for the user to do general purpose web crawling and gives them the ability to input their own command line arguments to be able to not only customize the starting place of the web crawler but also to provide a save file for the information in the crawl to be saved to, and the ability for the user to be able to load in a previously started /or completed crawl to continue or to analyze the results given from the crawl.
In addition to this the webalyzer gives the user the ability to analyze the output gotten from the crawl and to be able to give quantitative results based on the information given to the wehablyzer.

Implementation of Crawler.

As Previously stated, crawler operates on command line args that can be set from inside of Eclipse or from the command line. a list of arguments given can be seen in the -h(help) command but also can be viewed here for convenience. Here are the commands given to you by web crawler.

```
-s      (String) Specifies the start URL — the URL at which the crawl begins.
        Overridden if a previously saved crawl state is loaded from file via -L.
-L      (String) Specifies a file name from which to load a previous crawl state.
        Overrides -s. If the specified filename is non-existent, illegal, does not contain a
        legal CrawlState object, or is otherwise erroneous, that is be considered to be an
        UNRECOVERABLE ERROR.
-S      (String) Specifies a file name into which to save the current crawl after it completes.
         If the specified filename is illegal or erroneous, that is be considered to be a
         RECOVERABLE ERROR and the file will be saved as crawl_dump.dat
-m      (int) Specify the CRAWL-MAX parameter. In accordance with the Crawler Safety Requirements,
        Crawler does NOT allow a value of greater than 10,000 here. A user-specified value
        greater than 10,000 (or less than 1) any values outside this ranger are considered to be
        UNRECOVERABLE.
-d      (int) Specify the delay time between page requests, in milliseconds. In accordance with the
        Crawler Safety Requirements, Crawler MUST NOT allow a value of less than 1000 here. A
         user- specified value less than 1000 is considered to be UNRECOVERABLE.
-r      (flag; no argument required) Generate a human-readable report at the end of the crawl.

-h      (flag; no argument required) Print a short help message and exit.
```

These commands are pretty self explanatory and with a basic knowledge of how command line arguments can easily be implemented to give the user complete control of crawler.

Implementation of Webalyzer.

Once you have run a crawl, you can then analyze the results you obtain by running the webalyzer.java file.
The operation of this file is very similar to the operation of crawler. It runs off of command line arguments which are referenced not only in your -h command but also here:

```
-L          Specifies file to load.");
-o          Print report on the average OUTDEGREE of the graph.
-O          Print report on the full OUTDEGREE of the graph.
-n          Print report on the minimum OUTDEGREE of the graph.
-N          Print report on the maximum OUTDEGREE of the graph.
-i          Print report on the average INDEGREE of the graph.
-I          Print report on the full INDEGREE of the graph.
-j          Print report on the minimum INDEGREE of the graph.
-J          Print report on the maximum INDEGREE of the graph.
-d          Print report on the average shortest-path distance between mutually-reachable nodes in the graph.
-D          Print report on the complete shortest-path distance matrix between all pairs of nodes in the graph.
-r          Print report on the diameter of the graph.
-c          Print report on the number of strongly connected components in the graph.
-C          Print report on the size of the largest strongly connected component in the graph.
-h          Print a short help message and exit.
```

And with that you are given complete control of webalyzer. and with this information in mind you are ready to start parsing the web grabbing data, and going on your merry little way.

I hope that this guide finds you well and has been a tremendous help.

-Nialls Chavez