USERDOC.PDF

By Nialls Chavez

How to use Crawler.java

Preface:

This [program is a general purpose web crawler that takes an initial start point specified by the user and then with that information begins to parse the web for a user defined number of cycles. the idea is that each node that is added into the graph as x number of edges attached to that particular node that all branch off eventually going through the entire interwebs :) This documentation will show you how you can use this powerful tool to analyze the we in new and exciting ways!

HOW TO:
something to be known about this program is that it is allure from the command line using command line arguments. this how to document will assume that a the user already has a basic understanding of linux command line controls and a basic understanding of how java and the compiler works.
WIth that in mind lets begin .

First thing you as the user will need to know is the different functions that are given to you by the crawler. these argument are documented not only in the -h help output but also for quick reference is posted here

-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 default.txt

-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.

Now that you as the user have the basic understanding of how the crawler's interface works it is time to know how to compile this program.
if you are using terminal first you must link all the files together then add in whatever commands you wish to implement. if you are running this program from inside eclipse or some other IDE(this is the easier way to do it) all you need to do is go into run configurations then in the

arguments tab you just place whatever command line arguments you would like to implement and the compiler will take care of all the linking and all the dirty java compile work for you.

This completes the Crawler user documentation.

Happy Crawling!

- Nialls Chavez