Guideline to Run The Web Crawler

Requirements:

Java, Firefox Browser

How to Run:

First, please download the webCrawler.zip file. After unzipping the file, in the terminal, change your directory to webCrawler/src folder. Next, use the following two commands to run the code. The first command compiles the code and the second one runs it:

```
javac -cp .:../selenium/client-combined-3.1.0-nodeps.jar:../selenium/lib/cglib-nodep-3.2.4ar:../
selenium/lib/commons-codec-1.10.jar:../selenium/lib/commons-exec-1.3.jar:../selenium/lib/
commons-logging-1.2.jar:../selenium/lib/gson-2.3.1.jar:../selenium/lib/guava-21.0.jar:../selenium/lib/
hamcrest-core-1.3.jar:../selenium/lib/hamcrest-library-1.3.jar:../selenium/lib/httpclient-4.5.2.jar:../
selenium/lib/httpcore-4.4.4.jar:../selenium/lib/httpmime-4.5.2.jar:../selenium/lib/jna-4.1.0.jar:../
selenium/lib/jna-platform-4.1.0.jar:../selenium/lib/junit-4.12.jar:../selenium/lib/neko-
htmlunit-2.23.jar:../selenium/lib/phantomjsdriver-1.3.0.jar:../selenium/lib/sac-1.3.jar:../selenium/lib/
serializer-2.7.2.jar:../selenium/lib/websocket-api-9.2.15.v20160210.jar:../selenium/lib/websocket-
client-9.2.15.v20160210.jar:../selenium/lib/websocket-common-9.2.15.v20160210.jar:../selenium/lib/
xalan-2.7.2.jar:../selenium/lib/xercesImpl-2.11.0.jar:../selenium/lib/xml-apis-1.4.01.jar crawler/
Crawler.java
```

```
java -cp .:../selenium/client-combined-3.1.0-nodeps.jar:../selenium/lib/cglib-nodep-3.2.4ar:../
selenium/lib/commons-codec-1.10.jar:../selenium/lib/commons-exec-1.3.jar:../selenium/lib/
commons-logging-1.2.jar:../selenium/lib/gson-2.3.1.jar:../selenium/lib/guava-21.0.jar:../selenium/lib/
hamcrest-core-1.3.jar:../selenium/lib/hamcrest-library-1.3.jar:../selenium/lib/httpclient-4.5.2.jar:../
selenium/lib/httpcore-4.4.4.jar:../selenium/lib/httpmime-4.5.2.jar:../selenium/lib/jna-4.1.0.jar:../
selenium/lib/jna-platform-4.1.0.jar:../selenium/lib/junit-4.12.jar:../selenium/lib/neko-
htmlunit-2.23.jar:../selenium/lib/phantomjsdriver-1.3.0.jar:../selenium/lib/sac-1.3.jar:../selenium/lib/
serializer-2.7.2.jar:../selenium/lib/websocket-api-9.2.15.v20160210.jar:../selenium/lib/websocket-
client-9.2.15.v20160210.jar:../selenium/lib/websocket-common-9.2.15.v20160210.jar:../selenium/lib/
xalan-2.7.2.jar:../selenium/lib/xercesImpl-2.11.0.jar:../selenium/lib/xml-apis-1.4.01.jar crawler/
Crawler
```

- The crawling takes around 14 hours. While it is running, Firefox browser pops up and disappears frequently. Therefore, we suggest to run this over night or when you are not using the machine.

- Output files will be generated in the webCrawler/results folder. Once the crawling is done, Firefox does not pop up anymore and you should be able to find roughly 1900 files in the results folder.
- The webCrawler/results folder could be compressed and sent to google drive or dropbox or any other available online tool

Thanks for your help. Please let me know if anything is unclear or not working as explained.
Hooman Mostafavi - hoomanm@cs.uoregon.edu