**Crawler Engine**

**Assumption:**

The given url should be begin with http, https. All other urls will be rejected.

**About:**

This tool will automatically crawl the home page url and it's related page links of a particular domain to extract all email addresses. The tool is enough intelligent to avoid crawling any external link found during the crawling. After the successful crawling the result will be recorded inside a CSV file with these 3 column names; "input url name (eg. http://recruiterbox.com/)", "crawled domain page link (for eg. http://recruiterbox.com/about)", "email id"

**Work flow:**

This is a command line tool for crawling web pages. The user is capable of submit single or multiple urls at a time. Also the user can determine the output file location where they want to record the crawling result.

If we submitted a url to the crawling tool, the tool will validate the url before submitting it to the crawling engine. If the submitted url is other than http or https, the validator will reject the url. If the entered url have any trailing ‘/’ it will be cleaned out by the validator before submitting it to the crawling engine.

After a successful validation, the validator will pass the url to the crawling engine. In the crawling engine, every urls in the page will be scanned and added to a crawler queue (we are using SET data structure for implementing the crawler queue. Because this kind of data structure will not allow any duplicated entries). If the crawler found any relative url, it will convert into an absolute url before it storing into the Crawler queue, if the crawler found out any external url, it will be omitted.

During the page scanning process the crawler will clean up the page source by replacing any closed html tag and some special characters like single or double quote, semi colon, back or forward slash, back space, tab, carriage return, greater than or less than and a string “mailto” with a space.

Then the crawler will scan the cleaned up pages for the email. If the crawler found out any valid email, then it will log to the CSV file.

Using the same process, all the urls stored inside the crawler queue will be processed until the queue is empty.

**For More Reference:**

Please Read Crawler Document Index.html.

**Run Command:**

Added in separate file.