CS601-01/CS601-02: Lab 5[1]

LinkMatcher: Using regular expressions to find hyperlinks

Due Date: October 25th, 11:59pm (12 pts)

For this homework, you will write a class called `LinkMatcher`, that finds and returns a list of "unique" hyperlinks contained in a given HTML file. A link in an HTML document is specified in the **href** attribute of an **anchor tag** (<a> tag), as in the following example:
<a href="http://www.cs.usfca.edu/">

Note that the format of the anchor tag is actually more complex than the example above shows (say, the following would also be a valid a tag: *<a name="home" target="_top" href="index.html" id="home" accesskey="A">*, so you can **not** assume that it is <a followed by the space and the href attribute. Before you start working on this lab, go over the HTML links tutorial at http://www.w3schools.com/html/html_links.asp
Your goal is to extract only the link, you do not need to capture other attributes of the anchor tag.

You may **not** use any classes or packages for this assignment except `String`, `ArrayList`, `StringBuffer`, `Pattern`, `Matcher`, `BufferedReader`, `OutputStream`, `InputStream`, `PrintWriter`, `Socket`, `IOException`. Before using any other class, please ask the instructor.

You are required to use the starter code provided by the instructor.

Requirements:

1. Fill in the regular expression in the string called REGEX for matching an anchor tag (and capturing a link by using a group)
2. Fill in the code in the `findLinks` method that takes the name of the .html file and returns an ArrayList of hyperlinks in that html. The requirements:
   - The list should not contain "duplicates". For the purpose of this assignment, "duplicates" are the links that are the same, except for the fragment. Example: your code should consider these two links as equal: (because they are the same if you remove the fragment).
     "java/lang/StringBuffer.html#StringBuffer"
     "java/lang/StringBuffer.html#StringBuffer-java.lang.String"

   - Do not include links that take you to the same page (links that start with the fragment).
   - You are required to use a regular expression and classes Pattern and Matcher in

[1] The assignment is modified from the original assignment of Prof. Engle.

this method.

3. Fill in the code in the `fetchAndFindLinks` method that takes a URL, fetches an html page at this URL (using sockets), and finds all unique hyperlinks in that webpage. Again, the list should not contain "duplicates" (see the explanation above) or the links that take you back to the same page.

Javadoc Comments

You are required to add javadoc comments to your code (above each class and above each public method).

Submission:

Your lab5 should be submitted to your private lab5-username repo on github. Please keep pushing your code to github as you work on the lab. No history on github except for the final submission, will result in a 0 for the assignment.

The instructor will provide a test for the lab, however it provides only minimal correctness checks. You are responsible for doing your own thorough testing before submitting the lab.

Please note that you are **not** allowed to search the web for the appropriate regex, you need to come up with it yourself, it is the main point of the assignment. **The instructor will be asking students to come for a quick code review for this lab.** If you are unable to answer questions about the different parts of regex you are using, you will not be able to get any credit for the lab.

Please note that the next lab will come out on Monday night, so the *recommended* submission date for lab 5 is Oct 23.