CSC 215-05 Project 3 – Web Scraper

April 11, 2018

Due

- May 4th 11:59pm. Canvas will not let you submit assignments after this date.
- Submit your C++ source file and a screenshot showing a portion of your output.

Tasks / Skills

- String parsing.
- Using an array, vector or recursion.

Steps

1. Retrieve the web page found at http://www.msn.com/
2. Report all the URLs that are found on that page.
3. For each found URL, retrieve its web page and report all the URLs that are found on that page.
4. Only report URLs to a depth of 2.
5. When reporting the URLs the hierarachy should be clear. Here is one way to format the output:

```
link 1: www.msn.com
        link 1: www.abc.com/index.html
                link 1: www.abc.com/contact_us.html
                link 2: www.abc.com/about_us.html
        link 2: www.xyz.com/index.php
                link 1: www.xyz.com/contact_us.html
```

Notes

- You only need to find hyperlinks defined by the <a> HTML tag.
- Do not report or follow any hyperlink that is a bookmark on the same webpage. A bookmark is a href value starting with a '#' character.
- Do not report or follow any hyperlink that starts with "javascript".

Grading

- 30 points – Display all the URLs found on the www.msn.com web page. Indent them similar to the above example.
- 30 points – For each URL found, display all the URLs found on that web page. Indent them similar to the above example.
- 30 points – Correct use of an array, vector or a recursive function in your algorithm. You could also accomplish this project with sets of variables. (Choose one)
- 10 points – Proper heading comment, overview comment, source comments, and indentation.