University of Washington Bothell

CSS 436: Cloud Computing

Program 1: Simple Web Crawl

Purpose

The programmable Web and cloud is built on HTTP. In this lab we will utilize HTTP programming to "GET" and crawl parts of the HTML based Web. The goal is to familiarize students with HTTP and programmatically searching the web.

Problem Statement

Create a java application, WebCrawl.java, which takes two arguments from the command line:

1) A URL as a starting point
2) The number of hops from that URL (num_hops)

Your application will download the html from the starting URL which is provided as the first argument to the program. It will parse the html finding the first <a href > reference to other absolute URLs, for instance https://www.w3schools.com/tags/att_a_href.asp . Make sure that you have not previously visited this page (if you have then skip and find the next reference). Look for only http and https URLs. The application will then download the html from that page and repeat the operation. If that page is not accessible then continue on the current page looking for the next reference and visit that reference. You will do this num_hops times.

Your app should print out to the console the URL of each hop that you visit. If you encounter a page without any accessible embedded references you should stop there and print out the result.

Problem Statement Details

- **Example Invocation:**

  Java WebCrawl http://courses.washington.edu/css502/dimpsey 5

Details

- Please use java for this program. You can use either the HttpConnectionURL or HttpClient java objects
- The JSoup library should not be used for this assignment.
- Make sure that your program takes in two arguments on the command line—*do not query the user for the URL or number of hops.*
- Your program should gracefully handle all input, including bad input.
- An URL with a trailing / should be seen as the same as one without a trailing /
- If a site is unreachable continue down the current page for a URL which is accessible
- You should appropriately handle HTTP requests with return codes in the 300s or 400s
- **We will test your code on the Linux Lab so please test you application there**

Turn In

WebCrawl.java with your name at the top of the file