University of Washington Bothell

CSS 490: Cloud Computing

Program 1: Simple 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 an application which takes two arguments:

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

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). The application will then download the html from that page and repeat the operation. You will do this NumHops times. Your app will print out to the console the value of the final URL you landed on as well as the html. If you encounter a page without any embedded references you should stop there and print out the result.

Problem Statement Details

- **Example:**

  MyWebCrawl.exe http://courses.washington.edu/css502/dimpsey 5

  Prints out to the console the URL and the html from the URM 5 hops from http://courses.washington.edu/css502/dimpsey

Details

- You may build you app either with C#/.Net, or Java.
- Please make sure to factor your application appropriately
- You may use packages from the web but do make sure to make direct HTTP calls (for instance using HTTPClient)

Turn In

A .**zip file** which the module named:

- Executable of application
- All code and clear instructions or Makefile on how to build and run the application