CS380 Information Retrieval and Web Search

Lab 5

Robot Exclusions, Tracking and Forms

Monday, Feb 24

It will be helpful if you sit more or less in your presentation groups as I will be talking with each group about their projects. Do not leave until I have gotten a chance to talk to your group.

Robot Exclusions

In class we discussed three ways to exclude "robots" from crawling a site, or at least particular pages on a site. Find example of each. Hint, look for pages that the authors would not want to have appearing in a search engine. While you are doing this, also look for other forms of exclusion. For instance, tags exists that indicate that the page should not be included in the yahoo directory hierarchy (sites still have this tag although the yahoo directory hierarchy no longer exists).

Some of the exclusion formats are difficult to find. Do not look for more than 15 minutes.

Tracking

Sample pages from the web looking for ways in which you are being tracked. What sites do not track you? Find examples of each. It will be very helpful to look at the request and response headers for a page. Tracking is not limited to these.

Writing Forms

Hand write your own html page with a single form. The form should contain multiple type of inputs: text boxes, radio buttons, check boxes, etc. You can read about creating forms on any number of web pages.

After reading about creating web pages with forms, do the following:

1. Create your own website on cs.brynmawr.edu! To do so:
    1. Make a public readable directory named “public_html”.
    2. Put a file named “hello.txt” into this directory. This file should just contain “hello YOUR NAME”.
    3. Make sure that this file is public readable
    4. In a browser goto http://cs.brynmawr.edu/~YOURUNIXLOGIN/hello.txt
2. Create a web page in your public_html directory that contains a form with at least 3 different types of input. For an example of an html form go to http://cs.brynmawr.edu/~gtowell/UC/former.html
3. Note that if you are logged into our UNIX system, this is a publicly readable file /home/gtowell/public_html/UC/former.html.
4. Make the action of your form a link to “http://russell.cs.brynmawr.edu/~gtowell/UC/hello.php". This page just echoes back all of the request headers and the form elements.
5. Try your form with both get and post as the “method”.

What to hand in:

A list of sites where you found robot exclusions (and the type of exclusion.)

URL for the page containing your form and a typical response from your form by my hello.php. (The response should change whenever you change the inputs on your form.) You may hand write / copy this information from a browser.