I427 Lab 6, October 3 2013

This lab will help you get started with Assignment 2, in which you’ll build a web crawler. At the end of the lab period, please submit your work by running the following command:

/u/djcran/submit lab6

We will use the submissions to record attendance and to measure the general progress of the class. If you do not finish during class, it would be a good idea to complete these exercises on your own, as they are good practice for writing Perl programs.

1. Login to burrow.soic.indiana.edu using putty, and create a new folder called lab6. You can keep all your files for this lab in this folder.

2. (download.pl) Write a perl program that asks the user for a URL, downloads the specified web page, and writes the HTML code to disk. Check out Perl’s LWP::UserAgent module to help you do this.

3. (links.pl) Building on your solution to question 2, write a program that asks the user for a URL, downloads that page, finds the URLs of all the hyperlinks on that page, and then prints them out. You might want to check out the HTML::LinkExtor Perl classes to do this.

4. (counts.pl) Building on your code from the last question, count the number of times each host name appears in the list of hyperlinks. (For example, the host name of

   http://www.iub.edu/student/index.shtml

is www.iub.edu). Output the list of < domain, count > sorted by the domain name alphabetically. Then output the list sorted by frequency.

*Hint*: How to sort a hash in Perl...

- by hash key? $sorted_keys = sort (keys %hashes)
- by hash value? $sorted_keys = sort { $hashes{$a} <=> $hashes{$b} } (keys %hashes)