CS380 Introduction to Data Structures

Lab 4

Crawling by Hand

Monday Feb 17

The goal of this lab is to simply do some crawling of the web with the goal of finding the following information.

1. How many distinct, indexable web pages are there in the brynmawr.com domain. Indexable generally means that the page must have its own unique, static URL that contains static content. So, for instance, most pages that you get to via search fail on this requirement.

2. How many subdomains of brynmawr.com are there? A subdomain is an Internet domain which is part of a primary domain. For instance, www.brynmawr.com is one. cs.brynmawr.com is a second. There are others.

3. How many completely different domains can you get to from a page within brynmawr.com ?

Rules:

1. Work in teams of 2-3. Perhaps the teams for homework 3? A reason for working in teams is to reduce bandwidth requirements. If all of you are crawling independently we would overwhelm the local WiFi.
    1. Turning off the download of images will help your efforts. (And the efforts of everyone else. Again, by reducing bandwidth requirements.)
2. Crawl for about 45 minutes.
3. Write NO code.
4. Do not crawl any page more than once.
5. Pick a single starting location and crawl from there. Pick your starting location wisely.
6. It will probably be good to have one person “crawling” and another “recording”, but I leave the approach to you. In a team of 3 this leaves one person for the vaunted role of “critic” or “back seat driver”.

One way to get an estimate for question one is to at the end of your crawl compare results with some other group. The size of the intersection of the two crawls compared to the size of the union of the crawls can give an estimate (under some independence assumptions that are clearly false in this case.)

What to turn in:
A single sheet of paper (perhaps this one) with the names of the people in your team and your answers for the 3 questions posed above.