FIT 100

Assignment 1: Searching the Web

(or, Finding what you want, and no more!)

Required reading for Assignment 1:

Link to and read the sections on **Search Engine Math** and **Boolean Searching** at the Search Engine Watch website. You do not have to read the information that is linked off of each of these pages, only the content at that URL. Review the **Search Engine Features** page to help in your search. There are also two required readings concerning copyright, the web and fair use. Copyright will be an important issue to consider as we move into Project 1.

Search Engine Math:
http://www.searchenginewatch.com/facts/math.html

Boolean Searching:
http://www.searchenginewatch.com/facts/boolean.html

Copyright, Public Domain and the Web:
http://www.copyrightwebsite.com/info/publicDomain/publicDomain.asp

Fair Use:
http://www.copyrightwebsite.com/info/fairUse/fairUse.asp

Introduction:
Many of you have done a fair amount of browsing and searching on the WWW. But have you ever thought about how and where to search in such a way that you get only those sites you want and no more? Constructing a search that does exactly that is very difficult, if not impossible. However, you can learn to search the Web in a way that brings back a smaller set of "hits" (web pages that match your requirements), and improve the chances that these hits are more relevant than not.

So, what exactly IS a Search Engine and why do I care?
A search engine is really just a program, or series of programs, that is designed to try and help users find useful information on the Web. A search engine consists of several components: a crawler (a program), a query processor (another program that deals with user requests), a user interface where a user enters terms for the query processor, and an index database. The basic idea is that a search engine takes terms that you enter and tries to match those terms with documents out on the Web that are most relevant.

Seems simple, doesn't it? Yes, it seems simple… but relevance is hard for a program to determine when it doesn't "know" the person doing the search. This is an exercise for you to see both the ease and difficulty of searching for information on the web.

Objectives:
- To use basic search strategies in a search engine and bring back sites with information on a topic.
- Learn to find the best search method for a particular search engine.

- To develop systematic and precise search skills.

Online Resources (not required reading-just useful links):
Below is a list of some available search engines. There are many more search engines out there.

Google: http://www.google.com/
Uses link popularity as a way to rank a web site. If 50 different sites link to one other site, this is a good indicator that it is a relevant page for the topic it covers.

WiseNut: http://www.wisenut.com/
A newer search engine and possible rival to Google. Wisenut uses an algorithm similar to Google's PageRank.

All The Web: http://alltheweb.com/
Also a newer search engine. Allows searches for MP3s, videos, pictures.

AltaVista: http://av.com/
One of the oldest search engines around. Allows searches just on images and other formats. Also has a translate feature.

Vivisimo!: http://vivisimo.com/
Groups results from a search into hierarchical sets of categories (they show as folders).

Some search engines use a directory structure to organize human-indexed web sites by subject. These indexers decide which category is most appropriate for a web site and place it there:

Yahoo!: http://www.yahoo.com/
Directory setup. Provides email, news, etc.

List of Search Engines by function:
http://www.searchenginewatch.com/links/
A useful page with lists of major and specialized search engines.

More resources on Search Engines:
The Virtual Chase
http://www.virtualchase.com/Search_Engines/

To Do:

1. Go to http://yahoo.com and use the categories/directories to find the web site on Computational Geometry.

A few of the Yahoo categories

   A. What is the most logical starting point?

   B. What is the URL of the site you found on Computational Geometry?

   C. What is the full path in the directory to get to the Yahoo! page that lists this site? For example, one path in Yahoo! to get to the Information School website at the University of Washington is:

Now, go back to the start page at Yahoo [just click on the Yahoo! logo at the top left of the page]

and try to search for the same website using the search box at the top of the page.

   D. How did you construct the search? Write the entire search box contents **exactly** as you enter them in the search box.

E. Where did the Computational Geometry site come up in the first page of results? List the page number and result number.

2. Using a search engine from the list above, (**not the Yahoo Directory**), search for information about the riots that broke out in Seattle in December of 1999 over WTO, the World Trade Organization.
   A. What search engine did you use?

   B. How did you construct your search? Write the entire search box contents exactly as you enter them in the search box.

   C. How many sites pertaining to the riots in Seattle in 1999 were in your first page?

   D. Try 2 other search strategies using the same search engine (see the readings on Boolean and Math searching if you have questions). List them both here:

   E. Look at the search that appears to be most effective - in other words, the search that gave you more sites you considered relevant to the topic in the first 2 pages.

      Why do you believe your results were better with this search?

3. Use a search engine you haven't tested yet. Look at the help sheet at Search Engine Watch [http://www.searchenginewatch.com/facts/ataglance.html] or find the help page for the search engine to see what search operators to use.

   Now, find a site dedicated to the victims of the terrorist attack in September, 2001.

   A. What is the URL?

   B. How did you construct your search? Write the entire search box contents exactly as you enter them in the search box.

   C. Did you adjust your search at all during this process? Show the different searches used to finally arrive at the site you chose.

4. **This question (#4 A-E) is EXTRA CREDIT!!!**
   Use web searches to find the answers to the questions below. With each answer, write the following:

   - the search engine used
   - the series of queries you used to find the answer
   - an explanation of your search strategy, identifying which of the above query refinement strategies you applied and why
   - the URL of the page where you found the answer to the question

   The first question shows you appropriate answers to guide your work:

   A. IBM produced the first hard disk drive. Find a page on an official IBM web site describing this disk. When was it made, what was it called, and how much storage space did it offer?

      (1) http://altavista.com
      (2) search on term **IBM** to figure out IBM's domain name, which is ibm.com
      (3) search on terms **+site:ibm.com +"hard disk"** for pages on IBM's web site(s) with phrase "hard disk" in them
      (4) refine search on terms **+site:ibm.com +"hard disk" +first"** with additional keyword "first" to narrow search answer ("1956, 305 RAMAC, 5 MB") found on second hit:
      http://www.storage.ibm.com/hdd/firsts/n1956.htm

   B. Who is often called the "Father of Computing" and when did he live?

C. The ENIAC is often called the first general-purpose computer, though some question this distinction. In kilograms, how much did ENIAC weigh?

D. What was ENIAC designed to be used for?

E. Who is considered the father of the concept of hyptertext?

Searching for Images on the WWW

A note on copyright and public domain images:
Images and other files and content on the Internet are protected in the same way as print materials and photographs. Use of digital images for purposes of alteration and display on the Internet has limited coverage under the conditions of fair use. [http://www.templetons.com/brad/copymyths.html].

Public Domain
[http://www.copyrightwebsite.com/info/publicDomain/publicDomain.asp] items are those in which the copyright has been lost, has expired, or the author of the work makes no copyright claims to reproductions or enhancements of the work.

When Do Items Become Public Domain?

If you use an image of a person for reasons of making a profit, you are responsible for obtaining permission from the person or their heirs. If you use a trademark image, you must also get permission.

5. Using the Search Engine Math you read about, construct a search to find sites that contain images in the public domain. Use Google for this first search.

   A. How did you construct your search? Write the entire search box contents exactly as you enter them in the search box.

6. Do that same search across in AltaVista and Vivismo as well (or 2 other search engines of your choice). Compare your top 10 hits.

   A. What 2 search engines are you using?

B. Are the top 10 returns the same in each Search Engine?

C. Do all top 10 sites state, very clearly, that they DO contains images in the Public Domain-free for anyone to use?

7. Try changing the search technique and see if you get different results.

   A. How did you construct your new search? Write the entire search box contents exactly as you enter them in the search box.

   B. How do your results change?

C. Can you say, without any doubt, that all pages returned contain images within the public domain? What information from the site leads you to believe this?

8. Now do a search for sites that contain images free of copyright.

   A. What search engine are you using?

   B. How did you construct the search? Write the entire search box contents exactly as you enter them in the search box.

   C. Of the top 10 results back, how many sites [give the number] were actually very plainly "copyright free"? How can you tell?

9. Search for images of your favorite city on the web. Alta Vista has a way to just search for image media on the web.

   A. Give the URLs to your 2 favorite images from the list.

   B. What other search engines have this same feature? Find 2 and list them.

10. Find an image of the New York skyline.

    Give the URL of the ORIGINAL image location (not just the search engine page that listed it).

    What is the copyright information about the use of this image on your personal web site? (You may need to search the site a bit to find this)

11. Search for images you would like to use in a website of misinformation (Project 1) and save them. (You will not turn in the URL for your image, this is just to start you on your image search for the project.)

    NOTE: Make sure that any image you select is in the Public Domain OR the copyright policy on the site where you find it states that you are allowed to use it for non-commercial purposes!!!!! If there is no copyright policy evident, then try to contact the site owner and explain why you would like to use the image.

    You must provide adequate proof why you are allowed to use and manipulate any images that you find.