Rules. This is to be completed no later than the start of class on the day it is due. Earlier submission during office hours is also possible. A step of this assignment requires that you fill-in an online form by noon of the due date.

Due Date: No later than start of class on Mon Oct 3, 2011.

Topic: Crawlers and freshness

1 Objectives of the assignment

You will be asked to create your own web-page on AFS. The top-level file in your web-page will be an `index.html` file. Within that page/HTML file, provide an anchored link to a second page/HTML file appropriately named `cs485f11.html` of yours. In that second file you will minimally provide a link to the course web-page with anchored text as described (provided) in the provided template file. The location of `cs485f11.html` is that of `index.html`. The location of `cs485f11.html/index.html` will be forwarded to the instructor by filling an online form available in the Handouts section of the course web-page (Section B) related to Assignment 2 (this homework).

You are thus free to deal with `index.html` any way you like, but minimally it will include a link to `cs485f11.html`. The latter file must be created and will minimally include a link to the course's web-page. The file `cs485f11.html` will serve a variety of purposes including a testing of how often Google visits (if it ever visits) your newly created web-page and the `cs485f11.html` file in particular.

In addition you will perform three queries on Google and Bing and will determine through probabilistic consideration an estimate of the corpus size of those two search engines. This is different from the methods followed in Assignment 1.

2 Requirements

1. The first part of the assignment is completed when an online form is filled no later than noon time, the day the assignment is due. This serves as a confirmation that you have achieved the objectives of Part 1 described below.

2. The second part of the assignment requires the submission of a written report (in paper or electronic form) that is no more than one page long.

3 Part 1 (60 points)

(a) **Login and Logistics.** Login on AFS through any AFS machine. Possible choices are `afs10.njit.edu`. Note that you need to use ssh (secure shell) to connect to these machines from within NJIT or outside NJIT. In fact you are going to be connected to one of `afsconnect1.njit.edu` or `afsconnect2.njit.edu` which are aliases for `afsXY.njit.edu` where `X,Y` are numeric digits. Your AFS login is usually your UCID. Your AFS password might be different or the same as the UCID password. One way to reset the former to the latter is by going to `http://afspassword.njit.edu`

Additional information can be found at `http://ist.njit.edu` by clicking on the Services tab or directly following the link `http://ist.njit.edu/accounts/index.php`.

The link `http://ist.njit.edu/webhosting/afs.php` shows how to create a web-page.

(b) **File index.html in `public_html`.** In the AFS home directory of yours, a subdirectory `public_html` will be the home directory of your web-page.

Create inside that directory an `index.html` file (note that the suffix is not `htm` but `html`). A template that can be used for this `index.html` is the CS 485 web-page (you know the URL) or the `template.html` file available with this Assignment 2 from the Handouts section of the course web-page (section B). Naturally you are expected to edit that file accordingly and in at least the following manner.

Step b.1

- Clean `index.html` up by removing CS 485-related info specific to the instructor. Personalize your page minimally by providing at least your name.
- Edit the `TITLE` head to remove the instructor's information, and add your own name in a title that might read `Web-page of` etc.
- Update the `META` tag keyword list appropriately to reflect your design.
- Create/Update an anchor that will point to a newly created file named `cs485f11.html`. In the bottom of this `index.html` file include a `This file was last updated on` with date and time information of your choice (the `template.html` provides an example).
- You are free and welcome to further customize this file to your taste.

Step b.2

Repeat the same steps for the newly created `cs485f11.html` with the only difference that an anchor in this file will point to the web-page of the course. You may reuse `template.html` if you so wish.

This way each one of you will have created

Step b.3

- An `index.html` file with appropriate TITLE, KEYWORDS information and at least one link pointing to a `cs485f11.html` file.
- A `cs485f11.html` file with appropriate TITLE, KEYWORDS information and at least one link pointing to the web-page of the course.
- Both files will be time stamped as explained earlier, and time stamps must be updated often and monitored (daily or weekly).

(c) Deliverables.

Whenever you feel ready, but no later than **noon time** of the due date submit some information of the location of `index.html` file by using form CS485B colocated with this Assignment in section B1 of the course web-page. We will then check whether everything is ok. (Beware of the existence of two forms in section B.) **We will use this information to retrieve `index.html`, connect through it to `cs485f11.html` and connect through it to the course web-page. If any link is missing you will be penalized accordingly.**

This completes this part of the assignment. (Those two pages created must remain in place throughout the semester. You can delete the `cs485f11.html` file and the link to it from `index.html` after the end of the semester in December.)

In the meatime (before, during, or after the submission) you can also do the following.

- Search the Web (eg. Google or Bing) to check whether you have been indexed!
- If you have been indexed by a search engine look at the cached copy of your web-page maintained by that search engine and record the time stamp of `This file was last updated on`.
- Update the date/time information of your page (and if you wish, add material to your web-page), and repeat the first two items above occasionally (once a day) to determine the freshness/recency rules of the search engine in question, i.e. how often it will visit the newly stamped version of your web-page.

4 Part 2 (65 points)

We are asking you to search in queries 1-6 certain uncorrelated terms, and in queries 7-9 to improvise and determine yourselves something better than 1-6. Choose two words for `w1` and `w2` that are not correlated. For a counterexample a poor choice for the pair (w1,w2) would be (tropical, fish) or (tropical,storm). The textbook suggests (tropical,Lincoln). Use something else (eg. your last name for w1 or w2). Because this is not a verbatim query for words `w1, w2` we use the $\lceil\rceil$ symbols in queries 7-9 instead of $\langle\rangle$.

No	Query Term(s)	GOOGLE No of hits	BING No of hits
1	$\langle\ tiger\ \rangle$		
2	$\langle\ algorithm\ \rangle$		
3	$\langle\ tiger\ algorithm\rangle$		
4	$\langle\ falcon\ \rangle$		
5	$\langle\ algorithm\ \rangle$		
6	$\langle\ falcon\ algorithm\rangle$		
7	$\lceil w1\rceil$		
8	$\lceil w2\rceil$		
9	$\lceil w1\ w2\rceil$		

Table 1: Table for Q3.7

Based on the hits reported for 1-3, 4-6 and 7-9 estimate through probabilistic reasoning as specified in class, textbook, and the notes what your estimate of the corpus size of Google and Bing would be. Are these estimates consistent with those reported in Assignment 1 (solutions)? Explain. [Solutions for Assignment 1 would be made available late Friday/Saturday Sep 30.]