A. V. Gerbessiotis CS 345 Fall 2015
HW 1 Sep 14, 2015 66 points

CS 345: Homework 1 (Due: Sep 28, 2015)

Rules. Individual homeworks; see Handout 1 (aka Syllabus). Hardcopies may be submitted no later than start of class the day they are due; electronic copies by 23:59:59 the same day.

Problem 1. (21 points) **Create your Web-page and establish Freshness of it**

1.A. Objectives. The objective of this problem is for you to create (a) a web-page if you don't already have one, (b) establish a link from your web-page to another page to be named `cs345f15.html`, and (b) then from within `cs345f15.html` create at a minimum a link to the course web-page as it appears in the syllabus.

(1.A.1.) If you do not have a web-page, you will have the opportunity to create a web-page for yourself. This file is usually named and referred to as `index.html`. If you have one, create an anchor (link) to another file of yours to be named `cs345f15.html`. (Note that nowhere in `index.html` should a link to the course page appears; if it does you will lose lots of points!)

(1.A.2.) File `cs345f15.html` will minimally contain a link to the course web-page `http://www.cs.njit.edu/∼alexg/courses/cs345/index.html`. You are free to add another information though it is not required. A template for either (1.A.1) or (1.A.2) is available under `template.html` in Section B of the course web-page. Download and edit it at will.

(1.A.3.) After you are done with (1.A.1) and (1.A.2) you submit the link information of your web-page to us through an online form recording the transaction number and time. The latter information will become part of your homework submission. We will do the grading through that page!

1.B. UCID account: login and password. Usually the password you will need is the UCID login and password combination. If it does not work you might need to go to `ist.njit.edu` and under `Services.Accounts.UCID` or `Cybersecurity@NJIT.StrongPasswordManagement` locate either the GlobalPasswordChange or Unattended-PasswordReset to change all of your passwords as needed to a common one. If you still have problems, go to the basement of the Parking Deck.

1.C. Instructions to create a Web-page. In `ist.njit.edu` click on `Services.WebpagesAndSocialNetworking` and on the right (main) area the first link there for instructions.

1.D Your Web-page (index.html). At `http://web.njit.edu/∼UCID/index.html` after you complete the creation of your web-page there would be a file named `index.html` that you have just created. It must minimally include and conform to the following requirements.

1.D.1 File `index.html` has no explicit references to `CS 345` or to the intructor. It minimally lists your name, NJIT email address or UCID.

1.D.2 File `index.html` has a time stamp at the bottom recording some date and time information. This might be of the form `This file was last updated on`. See also template.html

1.D.3 File `index.html` has a properly formulated `TITLE` tag that includes text along the lines `Web-page of` including your name.

1.D.4 File `index.html` has a properly formulated `META` tag keyword list appropriate to your web-page content and your design. See template.html for an example.

1.D.5 File `index.html` has an anchor link with some relevant anchor text to a `cs345f15.html` file that would co-exist in the same directory with this `index.html` file.

1.E File cs345f15.html.
In the same location/directory as `http://web.njit.edu/∼UCID/index.html` create the second html file using or not `template.html` of Section B of the course web-page as guidance. Its name should be `cs345f15.html` and minimally include and conform to the following requirements.

1.E.1 File `cs345f15.html` has an anchor that points to the course web-page.

1.E.2 File `cs345f15.html` has a time stamp at the bottom recording some date and time information. This might be of the form `This file was last updated on`. See also template.html.

1.E.3 File `cs345f15.html` has a properly formulated `TITLE` tag that includes text along the lines `Course Taking in Fall 2015` .

1.E.4 File `cs345f15.html` has a properly formulated `META` tag keyword list appropriate to your web-page content and your design and the link it provides. See template.html for an example.

1.E.5 File `cs345f15.html` has an anchor link with some relevant anchor text to the course web-page.

1.F Future Actions.

Check the Web (Google, Bing, etc) to find out whether you have been indexed. If you have try to retrieve a cached copy of it (click on the link and a Javascript pop up menu will appear) and capture it in a jpg file for future references (including time accessed by the search engine).

Update periodically your web-page's `This file was last updated on` time stamp, so that you can find out when it was cached from the time-stamp information of yours, if Google or Bing do not provide their own information.

1.G Grading and Deliverables.

(a) You have created your web-page i.e. file `index.html` and then modified it accordingly per section 1.D to include certain information including an anchor/link to `cs345f15.html` .

(b) You have created `cs345f15.html` and modified it according to section 1.E.

(c) You have visited Section B of the course Web-page, and submitted link information to your web-page, recorded the submission's time and transaction number that you provided.

(d) Submit the transaction number and submission time, with the rest of the homework in writing (electronic or paper submission).

(e) **Future.** Between now and later homeworks, track your page on the Web and record (screenshots) how often it gets updated using the cached (by a search engine) copy and the timestamp information recorded by you and regularly updated there. This would help you in a later assignment to collect more points by showing how often a page gets indexed by a search engine (aka freshness).

Problem 2. (15 points) Crawling and Wget

2.A Wget.

The UNIX command wget available on an AFS machine allows you to fetch copies of web-pages. This includes a copy of the course web-page. However it is idiosyncratic and thus the prefix you give to it such as `cs.njit.edu` or `web.njit.edu` or `www.cs.njit.edu` will determine what you get or not. Login on afs (use program ssh and choose one of afsconnect1.njit.edu or afsconnect2.njit.edu) and read information about wget by doing `man wget` under AFS or `wget --help`. On UNIX typing `du -s dirname` you can get info about the file size of all files in the directory (and subdirectories) named `dirname`. The size may be in kilobytes or mutliples of 512B (read the manual). Certain details are missing because you are expected to read the manual pages and also to improvise. Wget generates a log file while transferring file. You can capture it into a file by redirecting its output to a file named `capturefilename` as partially shown in the following invocation fragment

```
wget ....   alexg/courses/cs345/index.html > & capturefilename
```

Moreover, it creates a directory that stores all the relevant files.

2.B Wget and a copy of the course Web-page.

Grab a copy of the course web-page. Among the three various prefixes of a URL listed above one would get what you wish for. The total size of files that you would transfer should be at least 2MiB if you do this before Sep 12, 2015, but more than 3MiB if you do it later. (KiB, MiB is SI notation, as you may recall.) Use appropriate options to get the desired result! (Do not provide login/password information to access protected areas.)

Then answer the following questions

2.B.1 Give the size in MiB of the downloaded files.

2.B.2 Give the number of downloaded documents (as reported in the log files). Note relevant requests by wget are reported accordingly so this information is available in the log file captured.

2.B.3 How many URLs were not found, i.e. even if there was a link to them, the corresponding file was not found. (In the capture file, this generates a relevant error.)

2.B.4 How many URLs resulted in rejects because of password authentication issues? (In the capture file, this generates a relevant error.)

2.B.5 What kind of error was generated for [2.B.3] and what for [2.B.4]?

Problem 3. (12 POINTS) **(Google and Bing corpus size?)**
For parts (a), (b), (c) below the following words should not be used: a, an, and, the, this, that, of, in, for, what, who, which.
(a) Write one query in Google with two or more terms that are different from each other that contains only disjunctions and no conjunctions or negations that returns a number of documents that is higher than 15,000,000,000 and as close to 30,000,000,000 as you can get. (If it happens to be higher, all the better.) Provide a screenshot. The shorter the query to size 2 the more points you will get.
(b) Do as in (a) but for Bing this time.
Note. If you write `a OR -a` not only term a appears twice but you also have a negation.
(c) Write the same query in Google and Bing with two or more terms that are different from each other that contains at least one disjunction and no conjunctions that returns a number of documents that is higher than 25,000,000,000 in both engines. Provide two screenshots. The shorter the query to size 2 the more points you will get.

Problem 4. (8 POINTS) **(Do Search Engines "cheat"?)**
(a) Formulate an appropriate query in Google, take a screenshot of the answer including the count of the result, and walk through several pages until the count changes into something that is at least 10 times smaller! If you can't figure out such a query it might mean that either you did not try enough or Google "does not estimate original query response size (aka you know how we call it)".
(b) Formulate an appropriate query in Bing, take a screenshot of the answer including the count of the result, and walk through several pages until the count changes into something that is at least 10 times smaller! If you can't figure out such a query it might mean that either you did not try enough or Google "does not estimate original query response size (aka you know how we call it)".
Note. The higher the ratio of the decrease is the more points you will get. Ie. an original result of 10000 that shrinks down to 1000 is not as good as a 100 that goes down to 1, or a 100000 that goes down to 100.

Problem 5. (10 POINTS) **(TO BE OR NOT TO BE capitalized...)**
The brackets below ⟨ and ⟩ indicate the start and end of a query. They are not typed in the query. We queried a search engine sometime on Thu Sep 10, 2015 and we got the following results. Number of hits reported are in thousand millions (aka billions) and rounded.

If the query system of the two engines is accepting Boolean queries, answer the following questions

5a. Is there an inconsistency between (the results for) query 08 and query 09? Explain.

5b. What would the answer to query 10 be? Explain and justify based on the information available. One of the numbers listed or a function of it is the answer. Explain and justify your answer. (We will provide screenshots of the asnwer as a justification.)

No	Query Terms	Hits in Billions
01	⟨`to`⟩	25
02	⟨`be`⟩	18
03	⟨`or`⟩	17
04	⟨`not`⟩	15
05	⟨`to be`⟩	18
06	⟨`to be or`⟩	8
07	⟨`to be not`⟩	15
08	⟨`to be or not`⟩	7
09	⟨`to be or not to be`⟩	7
10	⟨`to be OR not to be`⟩	?

Table 1: Table for Problem 5

Date Posted: 9/10/2015