A. V. Gerbessiotis | CS345
Sep 13, 2016 | Fall 2016
HW 1 | 67 points

CS 345: Homework 1 (Due: Before midnight of Sep 27, 2016)

Rules. Individual homeworks; see Handout 1 (aka Syllabus).

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

1.A. Objectives. The objective of this problem is for you to create (a) a web-site if you don't already have one, (b) establish a link from your web-site to another web-page to be named `cs345f16.html` , and (c) from within `cs345f16.html` create at a minimum a link to the course web-site.

(1.A.1.) If you do not already have a web-site, you will create one. The default web-page is usually named and referred to as `index.html`. If you have one, minimally edit it by creating a single anchor (link) to another web-page of yours to be named `cs345f16.html` plus few other things. (Note that nowhere in your `index.html` should a link to the course web-site should appear; if it does you will lose lots of points!)

(1.A.2.) File `cs345f16.html` will minimally contain a link to the CS345 web-site plus few other things `http://web.njit.edu/~alexg/courses/cs345/index.html`. You are free to add addtional information in your files though this is not required. A template for either (1.A.1) or (1.A.2) is available in Section B as `template.html`. Use it, edit it at will, or do not uset it all.

(1.A.3.) After you are done with (1.A.1) and (1.A.2) you submit the link information of your web-site to us through an online form (cs345b in section B.1) recording the transaction number and time. Be careful what you type, as the form information will be your official submission and used to retrieve the two pages and grade your work.

1.B. Logistics: UCID account login and UCID account password. Your NJIT UCID account is associated with a login and a password. 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 UnattendedPasswordReset 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-site 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 `CS345` 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-site/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 `cs345f16.html` file that would co-exist in the same directory with this `index.html` file.

1.E File `cs345f16.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-site as guidance. Its name should be `cs345f16.html` and minimally include and conform to the following requirements.

1.E.1 File `cs345f16.html` has a anchor link that points to the course web-site.

1.E.2 File `cs345f16.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 `cs345f16.html` has a properly formulated `TITLE` tag that includes text along the lines `Course Taking in Fall 2016` .

1.E.4 File `cs345f16.html` has a properly formulated `META` tag keyword list appropriate to the web-page content and your design and the link it provides. See template.html for an example.

1.E.5 File `cs345f16.html` has anchor link text relevant to the CS345 web-site.

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-site through web-page `index.html` and then modified it accordingly, per section 1.D, to include certain information including an anchor/link to `cs345f16.html` .

(b) You have created `cs345f16.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 electronic form (email).

(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. (10 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 WILL give it such would be `web.njit.edu` RATHER THAN `www.cs.njit.edu`. 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 ALL of the manual carefully to the end). 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-site as directed. The total size of files that you would transfer should be at least 6MiB if you do this before Sep 14, 2016. Use appropriate options to get the desired result! (Do not try to access the protected area C by providing logins/passwords to wget.)

Then answer the following questions

2.B.1 Give the size in KiB of the downloaded files and also in MiB (Be careful with your kilos).

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 kind of a message for [2.B.4]?

Problem 3. (12 POINTS) **(Google and Bing corpus size?)**
DO NOT use names (including politicians' or historical figures, alive or not) nor names of Institutions (eg NJIT) nor the following words: a, an, and, the, this, that, of, in, for, what, who, which.
(a) Write the same query in Google and Bing that contains one or two disjunctions and neither negations or conjunctions (implied or otherwise) that that returns a number of documents that is higher than 17,500,000,000. Provide a screenshot. The shorter the query to size 2 (terms) the more points you will get. The higher a number gets the better.
(b) Write the same query in Google and Bing that contains one disjunction and one negation and no conjunctions (implied or otherwise) that that returns a number of documents that is higher than 17,500,000,000. Provide a screenshot. The shorter the query to size 2 (terms) the more points you will get. The higher a number gets the better.

Problem 4. (12 POINTS) **(Do Search Engines ”cheat”?)**
DO NOT use names (including politicians' or historical figures, alive or not) nor names of Institutions (eg NJIT) nor the following words: a, an, and, the, this, that, of, in, for, what, who, which.
Formulate an appropriate query in Google and then in Bing, take a screenshot of the original answers in Google and Bing including the count of the result, and walk through several pages until the count changes into something that is at least 10 times smaller and take screenshots as well! If you can't figure out such a query it might mean that you did not try enough...
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. (12 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. If the query system is accepting Boolean queries, answer the following questions. (OR, NOT, AND, - are all operators for this imaginary engine; order of evaluation as in Boolean Algebra). Case for search terms is ignored. Note the double quotation marks in the last query.

(a) What do the four queries 5-8 mean? (Be brief)

(b) Is there any inconsistency observed in Queries 5-8 assuming Queries 1-4 are consistent? Explain the inconsistency and provide a corrected number of Hits consistent with the remaining results that you deem correct.

No	Query Terms	Hits
01	⟨to⟩	800
02	⟨be⟩	600
03	⟨to be⟩	500
04	⟨to be or not⟩	400
05	⟨”to be or not to be”⟩	100
06	⟨TO BE OR NOT TO BE⟩	600
07	⟨to be or not to be⟩	400
08	⟨to be OR not to be⟩	100

Table 1: Table for Problem 5

Date Prepared: 9/9/2016,Updated 9/14/2016