Networks and Distributed Systems	**Homework 6: HTTP and Web**
CS3700 Spring 2019	**February 4, 2019**

This homework is due at the beginning of class on March 14, 2019 and is worth 2% of your grade.

Name: _____

CCIS Username: _____

Problem	**Possible**	**Score**
1	20	
2	20	
3	30	
4	25	
Total	95	

1. In this question, we're going to use the `telnet` program at the command line to manually make a HTTP request. It should be available on most modern machines, as well as all of the CCIS Linux machines. If you're not familiar with `telnet`, read the Linux manual page before beginning this question.

1a. Using `telnet`, connect to the machine `toastytech.com` on the HTTP port. Issue a HTTP v1.0 request for the file "/evil/", providing no request headers. What is the response code, and what is the server telling you to do? Why might this be the case? (10 pts)

1b. The previous request did not work out the way we expected. Now, make a HTTP v1.1 request for "/evil/" to the same site. What response code do you get now? How much content does the server give you? (*Hint: There is one important header you need to be sure to include*) (10 pts)

2. Many services "crawl" the web in order to provide useful services; the most common example of this is web search engines like Google. However, the operators of websites often wish to express what parts of their website should and should not be crawled. We're going to explore this functionality in this question.

2a. The `robots.txt` file is one way this can be accomplished. What is the format of this file, and where should it be placed on your website so that Google et al. can find it? (10 pts)

2b. Locate the `robots.txt` file for `github.com`. Give two examples of pages on Github's web site that Google would be allowed to index, and two examples that Google would not be allowed to index. (10 pts)

3a. What is the `User-Agent` HTTP header? How is it used by web servers? (5 pts)

3b. Suppose that we built a custom web browser, and desired to only allow users who were running this particular web browser to visit our site (i.e., we did not want to allow users on Google Chrome to access it for security reasons). Would the `User-Agent` header be a good way of accomplishing this goal? Why or why not? (10 pts)

3c. Recall that the HTTP `Referer` header tells the server which web page "referred" it to the current request. However, this header field has raised a number of privacy concerns. Give an example of privacy issues on sites like Facebook or Twitter that is caused by the `Referer` header. (10 pts)

3d. If you were an operator of a site like Facebook, how might you ensure that the users who click on links to external sites from your site are not subject to these privacy issues? (5 pts)

4. For the next set of questions, you'll be using the developer tools in your browser. First, you should clear your browsers cache; typing CTRL-SHIFT-Delete opens the clear dialog in Firefox and Chrome. Next, in Firefox, go to Menu -> Web Developer -> Toggle Tools. In Chrome, go to Menu -> More Tools -> Developer Tools. Alternatively, in both browser you can type CTRL-SHIFT-I. Once the developer tools are open, go to the "Network" tab, then browse to `cnn.com`.

4a. Wait a few moments for CNN to finish loading. How many total requests did it take for the CNN website to load? (5 pts)

4b. How many requests were for HTML? How many for images? How many for JavaScript? (Hint: the filters in the "Network" tab will help.) (5 pts)

4c. Many of the requests on CNN are to third-party advertisers and tracker. Find the request to *hpr.outbrain.com* and click on it to inspect the HTTP request and response. Did Outbrain set cookies in your browser? Did your browser send cookies to Outbrain? What is the purpose of these cookies? (10 pts)

4d. Switch to the JavaScript "Console" tab of the developer tools. Look around, then complete this sentence: "Send your ideas to: _____" (5 pts)