i206 Spring 2013: Assignment 9

Name_____

Question 1: Networking, Operating Systems, and Distributed Systems

(a) Explain why during normal network operations, IP packets may be delivered out of order, even if all routers process packets on a First-In-First-Out (FIFO) basis.

(b) In the operating system simulation discussed in class and viewable at the link below, what is happening when the penguin switches in and out of being the process running on the CPU (represented by the green circle)?

http://courses.cs.vt.edu/~csonline/OS/Lessons/Processes/index.html

(c) Say you want to process a huge collection of web data. In particular, you have an enormous file of urls, where each line of data lists the url, the size in bytes of the web page at that url, and the date crawled. Say you wanted to compute the total number of bytes for every host (domain name in the urls). Why would mapreduce be a better way to go than simply computing this on a single computer?

(d) **(Extra credit**) Referring to question (c) above, what would be the map and reduce steps for computing the number of bytes for each host?

Question 2: Relational Databases (RDBMS)

(a) Either use SQLite3 as already installed on iSchool machines, or download and install on your own machine (see http://www.sqlite.org/download.html).

(b) In Assignment 8 you took HTML data from Yelp and extracted various information using regular expressions. In this assignment you will use that same data (and code – assuming you were successful) to extract the information about restaurants and to store it in an SQLite3 database. (See http://zetcode.com/db/sqlitepythontutorial/ for a tutorial on using SQLite3 from Python).

a) First examine the data extracted in Assignment 8 again and decide how each of the elements should be stored in the database.
b) Use the commandline SQLite3 to create a database and a table to hold the elements of the restaurant data. The database should have, as a minimum, the name, neighborhood, star rating, and phone number of each restaurant. (For extra credit include the categories)

(c) Using your program for Assignment 8 as a basis, use the regular expressions you created there to extract from the yelp_listings.html each of the elements in your database for each of restaurants in the listing.

a) Connect to your database created in Question 2
b) Set up a cursor
c) For each restaurant, use the cursor execute method to submit SQL INSERT commands to take each extracted element from the listings for a given restaurant

and insert the data into the database (hint: use one insert per restaurant to insert all data at once).

d) Run your program to insert the data
e) Then go to the SQLite commandline and use the SQLite3 ".dump" command to output the table definition and contents for your data.

(d) Write a simple python program to connect to your database, run SELECT queries to answer the following questions and **print the SELECT command you used and the results of the queries**.

a) What are the names of the restaurants in the SOMA neighborhood?
b) What are the average ratings by neighborhood? (use the AVG aggregate function and GROUP BY)