CSC 455: Database Processing for Large-Scale Analytics
Assignment 4

Due 11:00pm, Sunday, November 8th.

Supplemental reading: Python for Data Analytics, Chapter 4, Chapter 5

1. Implement the OR function in python that can combine two boolean NumPy matrices (do not use any built-in operators such as | or + for your answer, although you can use these operators to test your function). If you prefer, you can use regular list of lists in python instead of the NumPy matrix (e.g., like the example at the end of this paragraph). Your function should work for one-dimensional and two-dimensional matrices and it should error check to verify input matrix compatibility. Only matrices of the same size can be OR-ed together, so if the input to the function is two incompatible matrices, the function should return an error message (using return, not print).

   For example ORFunction([[True, False], [False, False]], [[False, True], [True, False]]) should return [[True, True], [True, False]].
   And ORFunction([[True, False, True], [False, False, True]], [[False, True], [True, False]]) should return an error message.
   Hint: Once you check for size compatibility, you will probably want to use a double for-loop.

2. We are going to work with a small extract of tweets (about 200 of them), Assignment4.txt available in dropbox in Assignment 4.

   For now, we are going to extract a few columns only. Do not forget that you can use json.loads(OneTweetString) to parse a tweet entity into a python dictionary. Don't forget to add import json to your code.

   a. Create a SQL table to contain the following attributes of a tweet:
      "created_at", "id_str", "text", "source", "in_reply_to_user_id",
      "in_reply_to_screen_name", "in_reply_to_status_id", "retweet_count", "contributors".
      Please assign reasonable data types to each attribute (e.g., VARCHAR(10000) is a bad idea).
      Use SQLite for this assignment.

   b. Write python code to read through the Assignment4.txt file and populate your table from part a. Make sure your python code reads through the file and loads the data properly (including NULLs).
      NOTE: The input data is separated by a string "EndOfTweet" which serves as a delimiter. The text itself consists of a single line, so using readlines() will still only give you one row which needs to be split by the tweet delimiter.

3. Write SQL queries to do the following:

a. Count the number of iPhone users (based on “source” attribute)

b. Create a view that contains only tweets from users who are not replying ("in_reply_to_user_id" is NULL)

c. Select tweets that have a "retweet_count" higher than the average "retweet_count" from the tweets in the view in part b

d. Create a view that contains only “id_str”, “text” and “source” from each tweet that has a “retweet_count” of at least 5

e. Use the view from part-d to find how many tweets have a “retweet_count” of at least 5

f. Write python code to compute the answer from 3-e without using SQL, i.e., write code that is going to read data from the input file and answer the same question (find how many tweets have a “retweet_count” of at least 5).

4. Write a python function that takes the name of a SQL table as parameter and then does the following:
   Select all rows from that table (you can assume that the table already exists in SQLite) with all attributes from that table and output to a file a sequence of corresponding INSERT statements, one for each row from the table. Think of this as an exporting tool, since these INSERT statements could now be executed in Oracle (you do not need to actually do that).

   This is similar to the question from the end of Part-2 in Assignment 1, only the values will have to be extracted form a SQLite table first. For example:
   generateInsertStatements('Students') should write to a file an insert statement from each row contained in the Students table:

   inserts.txt:
   INSERT INTO Students VALUES (‘1’, ‘Jane’, ‘A-‘);
   INSERT INTO Students VALUES (‘1’, ‘Mike’, ‘B‘);
   INSERT INTO Students VALUES (‘1’, ‘Jack’, ‘B+‘);

   I will be sure to post sample code for Assignment1.

   Hint: as you iterate through the rows of the given table, instead of printing the output to screen using print as you have done before, you will want to write an INSERT statement to an output file each time.

Be sure that your name and “Assignment 4” appear at the top of your submitted file.