FINAL EXAM, 839 SPRING 2019

Write down your name here (last name first) AND student ID number:

Please wait until you are told to start working on the exam.

75 minutes, closed books and notes, no electronics, one page cheat sheet allowed.

If a question is ambiguous, please state your assumptions and then answer based on those assumptions. Please keep the assumptions reasonable.

You can only use the space provided in this exam. Adding more space is not necessary and may result in points being deducted.

Some questions may be harder than others. You may want to handle easier questions first, delaying the more difficult questions to later.

15 Questions, 5 Points Each, for a Total of 75 Points.

1) Suppose you want to match tuples from two tables A and B, where each tuple describes a person. Suppose the schemas of these tables are the same: A(name,city,state,age) and B(name,city,state,age). Define the purpose of the blocking step, and give at least two examples of blocking that you can do on these two tables.

2) Suppose you want to do blocking on "city", that is, if two tuples do not agree on "city", then the pair consisting of these tuples does not survive the blocking step. What happens if you have a lot of missing values in "city"? What would you do?

3) Continuing with Question 2, besides missing values, are there any other issues you should worry about before trying to do blocking on "city"? Explain your answer.

4) Suppose you want to go ahead to do blocking on "city", as described in Question 2. Describe an algorithm to do this blocking step fast.

5) Suppose you have a table where each tuple describes a book. Suppose also that the attribute "publisher" has a lot of values such as "Springer", "Springer-Verlag", "Springer Publisher", etc., in short, many values referring to the same publisher but using different variations. You want to clean up this attribute, that is, normalize all the values so that all strings referring to the same publisher will be normalized to a single canonical string (e.g., "Springer-Verlag"). Briefly discuss a solution to do this.

6) Suppose you want to perform entity matching using a logistic regression rule. Define this rule, and give an example where it is more appropriate to use than using a linearly weighted combination of individual similarity scores.

7) Why schema matching is difficult?

8) Consider two tables X(name, phone, city, state) and Y(pname, pphone, address1, address2). Describe a solution that matches the two schemas using the data instances of Tables X and Y. (Note: this is a schema matching question.)

9) Continuing with Question 8, suppose you want to build a classifier that when given as input a data instance of an attribute in Table Y would label that data instance as "name" or not. Describe how you would build this classifier. In particular, what would be the training data for it? What are example features that you would use? What learning method would you consider?

10) Give a small example to demonstrate that by exploiting both the attribute names and the data instances, you may be able to achieve higher matching accuracy in schema matching, than if you exploit only the attribute names or only the data instances.

11) Consider two tables X(id, name, loc) and Y(id, pname, address, salary), both of which describe persons. Suppose you have performed schema matching and discovered that name = pname and loc = address. Explain the subsequent steps you need to do so that you can finally write a SQL query to retrieve the tuples of X and Y and combine them to populate a table Z(id, name, loc, salary). In other words, the SQL query combines the data of Tables X and Y into a single integrated table Z. Show the SQL query.

12) What are the five basic data analysis operations that data scientists often do in the data analysis stage?

13) In association rule mining, suppose the confidence of the rules A->B and B->C are larger than some threshold t. Is it possible that rule A->C has a confidence less than t? If yes, show an example. If no, prove that the confidence of rule A->C is also at least t.

14) Briefly describe a star schema (in OLAP)

15) Explain and contrast OLTP and OLAP queries.