CMSC498O
Fall 2014

Introduction to Data Science I
MIDTERM
Closed Book

Deshpande

- Total points: 45. Weight: 15% of the course grade.
- Show your reasoning. Write partial solutions. You will get a fair amount of the credit if I think you know the concepts.
- Unless otherwise specified, a *Yes/No* answer is *not sufficient* for any question. No points will be given without accompanying explanation.

Your Name:

Miscellaneous Questions 1 (9 questions - 2 pt each)

1. What is "volunteer bias" in sampling? You can use an example.

   Answer: See class notes.

2. What are "protocol buffers"? What are they used for?

   Answer: See class notes.

3. List three different data models.

   Answer: See class notes.

4. What is the difference between "deduplication" and "record linkage" in the context of entity resolution?

   Answer: See class notes.

5. Explain the rule-based approach to *relation extraction* in Information Extraction with an example.

   Answer: See class notes.

6. Briefly explain the notion of "overfitting" in statistical modeling.

   Answer: See class notes.

7. List and briefly explain one classification technique.

   Answer: See class notes.

8. Briefly explain the "local-as-view" approach in data integration.

   Answer: See class notes.

9. Consider the relations: $R(A, B)$ and $S(B, C)$, and the SQL query:

   ```
   select R.a, count(*) from R natural join S group by R.a;
   ```

   Briefly explain the result of this query in words. Why might you want to use *left outer natural join* instead of *natural join* here ? Assume A is a primary key for R.

   Answer: The query returns for each tuple in R, the number of matching tuples in S. The result will not contain any tuples $t = (a, b)$ in R that do not have a match in S, and we might want to use outerjoin to produce the result $(a, b, 0)$ for those tuples.

Miscellaneous Questions 2 (9 questions - 3 pt each)

1. We want to test whether the average cost of the phones purchased by UMD students this year has gone up or down from previous year. Say we know that the average cost of a phone purchased last year by a UMD student was $300. Explain the steps you would take to test the null hypothesis that the average cost this year = $300, including how you would make the decision of rejecting or not rejecting the hypothesis.

   Answer:

   - Take a random sample of the UMD students who have a bought a phone in the current year (you can just take a random sample of the students, and ignore the students who didn't purchase a phone in the current year).
   - Let μ denote the sample mean, and let σ denote the sample deviation. Say $\mu > 300$.
   - Choose a test statistic, in this case z-statistics is appropriate.
   - Compute the z-statistic as: $\mu - 300/\sigma$.
   - Consult standard formulas to decide how low is the probability of obtaining that z-statistic – if it is too low (below 0.05), we reject the null hypothesis.

2. Consider the following schema:
   create table r (a integer primary key, c integer);
   create table s (b integer primary key, a integer references r);
   create table t (c integer primary key, b integer references s);
   alter table r add constraint rreft foreign key (c) references t(c);

   - Why can't I add the foreign key reference directly in the "create table" statement for table "r" ?
     Answer: The table "t" hasn't been created yet.
   - Explain why the statement "drop table r" would be rejected.
     Answer: Because there is a referential integrity constraint from "s".
   - Is there any way I can delete all the tables ? Explain in words.
     Answer: Just reverse what we did to create the "cyclic" integrity constraints above. First alter the table "r" to remove the referential integrity constraint, and then delete "t", and then "s", and finally "r".

3. What does 'I' stand for in ACID properties? Briefly describe one mechanism for ensuring 'I'.

   Answer: See class notes

4. List and briefly describe three single-source data quality problems.

   Answer: See class notes

5. Briefly, at a high level, explain the *k*-means clustering algorithm.

   Answer: See class notes

6. Fill in the pseudocode for a naive implementation of the aggregation operation in the following query using Hashing.

   ```
   select R.A, sum(R.B) from R group by R.A

   ------------------------------------------------------------------------
   HashMap h = new HashMap();
   ```

```
for each tuple r in R:
    if h.contains(r.A):
        h.replace(r.A, h.get(r.A) + r.B)
    else
        h.put(r.A, r.B)

// print out the results
for k in h.keySet():
    print k, h.get(k)
```

---

7. Consider an *email* dataset, containing the data about a collection of emails within a company. For each email, we know the email addresses of the sender and the recipients, as well as the email text. There are however ambiguities since a person may use different accounts, including external (e.g., gmail, or yahoo). Suggest three approaches you would use to disambiguate, i.e., decide which email addresses belong to the same person.

   Answer: Some options are:

   - String similarity of the username part of the email address – should only be used if the username is sufficiently distinct (e.g., long usernames)
   - Analyzing email texts can give fairly good indications about whether they were written by the same user.
   - Users tend to send emails to the same group of users. We can generate some potential candidate pairs of email addresses (based on string similarity), and check if receivers of their emails overlap a lot (equivalently, we can check whether the people they receive emails from overlap.

8. On the following tables, what are the results of the queries listed below?

   R

A	B	C
'a1'	10	10
'a1'	20	20
'a2'	30	30
'a2'	0	NULL

   S

C	D
30	'd1'
NULLL	'd2'

   - `select avg(B) from R group by A`: **Answer:** (15), (15)
   - `select * from R where C != 10`: **Answer:** Second and third rows would be returned, but not the fourth
   - `select * from R, S where R.C = S.C or R.C is null`: The result contains 3 tuples. **Answer:** ('a2', 30, 30, 30, 'd1'), ('a2', 0, NULL, 30, 'd1'), ('a2', 0, NULL, NULL, 'd2')

9. (1) Consider an unbiased coin. Which of the two sequences, HHHHHH and HTTHTH, is more likely to happen if I were to roll the die 6 times? Explain.
   (2) An analyst for a fire department observed that: more firemen sent to control a fire meant higher overall damage. Should he recommend that fewer firemen be sent to fires? Why or why not?

   Answer: (1) Both sequences are equally likely.

   (2) No. Correlation does not equate causation. In this case, both the things (more firement sent to control the fire, and more damage) were likely caused by the same underlying thing – larger fires.