CSCI 585: final exam, 12/14/21

Duration: 2 hours: 90 minutes for answering, 30 minutes for submitting
[DO follow this 90/30 split!]

Please read the following carefully, before starting the test:

• the exam is **open** books/notes/devices - feel free to look up whatever you want (OMG)!

• **you NEED to do Q0** (worth 0 points, but -5 if you skip!); from Q1 through Q10 (worth 5 points each), you can choose **any 7**, for a total of 35; if you want you can do **8, 9, or even, all 10** - we will ADD up ALL your scores from all the questions you answer, then CAP it at 35, meaning, **a total > 35 will be set to 35, and if <=35, will stay as-is** - the most amazing deal ever!

• 'data' occurs in all 10 questions, look for it :) (it's a DATAbase course after all!)

• there are no ‘trick’ questions, or ones with long calculations or formulae, and there's certainly nothing to memorize [it's all OPEN, duh :)] It doesn't mean the questions are trivial! Please do answer carefully: in other words, **just answer WHAT IS ASKED**, otherwise you won't get points (eg if a question is ABOUT TM, don't DESCRIBE/DEFINE TM!)

• **please do NOT cheat [not kidding]** - this means NOT communicating with anyone via any device/medium/channel - **you will get a 0**, and be reported to SJACS, if you are found to have cheated; ANY attempt to get help from others in any form is a VIOLATION, as per https://policy.usc.edu/scampus-part-b/, sections 11.11 through 11.14 [read it, if you are not familiar with it]

• when the time is up (90 minutes), stop your work, then spend the rest of time (30 minutes) on submission [students with ASAS accommodations - your exam duration will be as per ASAS determination] - **submitting past the deadline comes with a penalty of 10 points [off your net score]**, because it is not fair to others if you go over when they don't

Good luck! Hope you enjoy answering the questions.

Q0 [0 points]. DO turn this in - DO NOT omit doing so [you will LOSE 5 points if you skip this].

Please write the following line, and sign it - it is your acknowledgment of having read USC's policies on academic misconduct (https://policy.usc.edu/scampus-part-b/, 11.11-11.14) and agreement to honor them.

I have read USC's standards on academic integrity, and agree to abide by them.

Q1 [2+2+1 = 5 points].

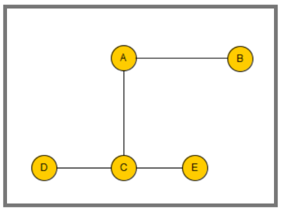
a. To help with indexing relational data stored as columns (and rows) in tables, the usual advice to SQL programmers is to avoid mixing columns and literals, eg. do not write (PRICE*1.1 > 50), for example (if we wonder which products in our store will be more than $50, if we increase all current prices by 10%). BUT - suppose that we DO want to write such 'mixed' (literals and columns) expressions, eg. (RAWSCORE+5 > 85), for readability purposes. How would you still make use of indexing in such 'mixed' expression cases (eg. the two examples in this question)? Don't worry about the technicalities or specific product details/syntax etc; instead, simply answer the question at a conceptual level (note: the answer is not a vague or subjective one!).

b. How is a bitmap index different from a simple column index such as for GPA, salary, etc? The answer is NOT that 'multiple columns are used in a bitmap index'.

c. How would we index a column that has long (eg. upto 100 chars), unique, strings?

Q2 [2+2+1 = 5 points].

How would you represent the following graph data, using XML, JSON, and, plaintext (using your own arbitrary format)?



Q3 [1*5 = 5 points].

Pick one of these four domains: an airport, an airline, a farm, a retail store (eg. 7-11). For your choice, list 5 pieces of data analysis (mining) you'd perform on each. Eg. if 'USC' was a choice, one of the five answers would be 'outgoing GPAs vs incoming (high school) GPAs, fitted using linear regression'. For each, be sure to clearly list the data (eg GPA) as well as the analysis (eg linear regression fitting).

Q4 [1+2+1+1 = 5 points].

a. Of all the data modeling/mining algorithms we went through, which is the most powerful?

b. Specifically what makes it [your choice in a. above] powerful? Explain, using a sentence or two.

c. Conversely, which one is the simplest algorithm to understand, implement, use, and interpret?

d. Why do we use 'edge' hardware for processing (certain kinds of) data, eg. in ML?

Q5 [1*5 = 5 points].

There are clear/unmistakable/irreversible 'trends' in the way data is stored, analyzed, interpreted etc. Name 5 of them (trends) we looked at, say a line or two about each. In other words, where is all this going, compared to what used to be, and, what is practiced now?

Q6 [5 points].

How would 'data viz' play out in the coming (eventually!) "metaverse"? In other words, today's data viz is on colorful screens, with animation and/or interaction (eg the bar chart races, D3, Shiny etc examples we looked at) - what will the future look like? FYI, 'metaverse' is a fancy word for VR (neither the term 'metaverse' nor VR is Facebook's/Meta's invention btw), so think in terms of that - describe what visualizations, interactions... would be like, in the metaverse: you are being asked to prognosticate/foretell/prophesy/imagine/... what will eventually be commonplace!

Q7 [4+1 = 5 points].

a. You dealt with FOUR different spatial data representations, in your HWs! What are they?

b. And, what would a 5th one be, in your own plaintext format? Assume you want to represent (just) long,lat,label,popularity for each location. You can just describe your format, or even better, provide a small, simple example as well (eg. with 3 locations from your HWs).

Q8 [5 points].

What would happen if the entire notion/practice of 'BI' (ie data warehousing) just went away?! That's the future that companies such as Snowflake [https://www.snowflake.com/] are proposing. What would take the place of ETL? Discuss, using a few lines. PS: we did briefly go through this in class :)

Q9 [1+4 = 5 points].

'Data flow' came up again and again, during the second half of the course.

a. What makes dataflow powerful?

b. List 4 instances (examples, ie. tools etc) we talked about (we discussed/encountered 5 of them).

Q10 [1*5 = 5 points].

Data analysis, without doubt, is transformational - in Kai Fu Lee's words, "data is the new oil". But, as we saw, it has its dark side, too. Discuss, using a sentence or two for each, five misuses/abuses of data [think: by individuals, corporations, governments].