CSCI 585: Final exam

Due: Wednesday, May 4, 2022 11:59 pm (Pacific Daylight Time)

Assignment description

Please read the following carefully, before starting the test:

• the exam is open books/notes/devices/minds - feel free to look up whatever you want, from wherever (but not whomever)!

• **WHAT TO ANSWER:**

- you need to answer the **first 5** questions (Q1..Q5), 5 points each
- after that, you can **pick any 5**, out of the remaining 10 questions (Q6..Q15), 2 points each; you CAN answer more questions if you want (6,7,8,9 or all 10)
- we will score everything you answer, add the scores (including partial ones), **CAP** them at 35 [>35 becomes 35]. How cool!

• 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! There are open-ended questions (which means there is more than one right answer), but they are not subjective ones (which means they are not about your opinion/viewpoint). Please do answer carefully: answer just WHAT IS ASKED, otherwise you won't get points (eg. if a question is -ABOUT- column fragmentation, don't DESCRIBE/DEFINE column fragmentation!). It's the quality (of your answer) that counts, not quantity (verbosity), or extraneous details...

• please do NOT cheat - 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 DSP accommodations - your exam duration will be as per DSP determination] - **submitting past the deadline comes with a penalty**, because it is not fair to others if you go over when they don't; note that you need to submit each answer separately (not all of them as a single PDF), this is a Crowdmark requirement

Fun fact: 'data' occurs 24 times (25, including in this line!), at least once in each question :)

Good luck! Hope you enjoy answering the questions, hope you find them to be easy+fun+st

Time left Hide

15:58:46

Submit your assignment

Help

After you have completed the assignment, please save, scan, or take photos of your work and upload your files to the questions below. Crowdmark accepts PDF, JPG, and PNG file formats.

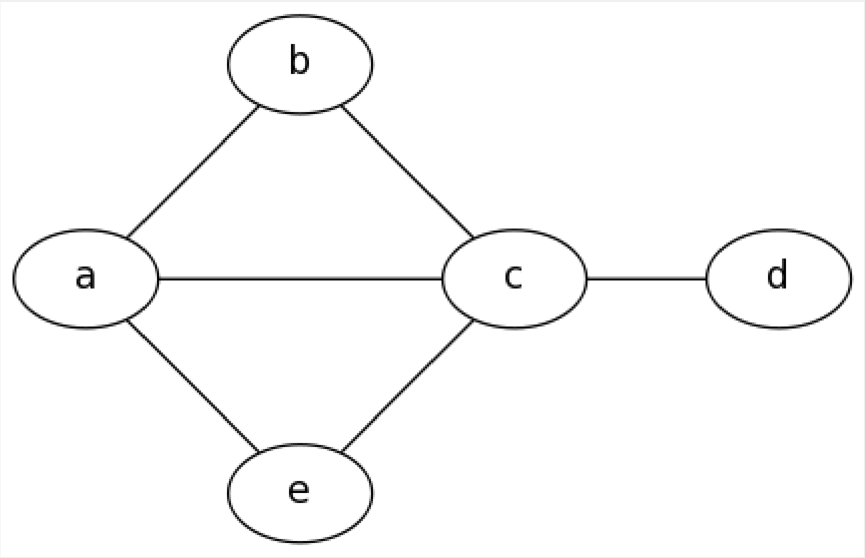
Q1 (5 points)

In the early days of digital data processing, mainframes were used to store and query data, access was through 'dumb' terminals. TODAY, we can access a wealth of data via smartphones.

Pick **five** 'connectivity' technologies, say a few words (a sentence or two) about each. Be sure to make the 5th one be 'MCC'.

Q2 (5 points)

Show how you would represent the following graph data as JSON in **three** ways, and XML in **two** ways (your XML ways can be equivalent to your JSON ones, ie you don't need to come up with 5 ways, just 3).



Time left Hide

15:58:46

Q3 (5 points)

A. What is it about JSON that makes it very powerful/flexible for data representation?

B. How would you represent 7 types of sins that people commit (!), using JSON? Use this as a guide: https://en.wikipedia.org/wiki/Seven_deadly_sins You can simply provide a small, syntactically valid, JSON example to illustrate your 'format' :)

Q4 (5 points)

For your HWs, you handled data in a variety of formats. List and briefly discuss **five** of them - only two of them can come from the spatial HW, others need to come from HW4, HW5.

Q5 (5 points)

A. Supervised ML is data-intensive. There is a human cost involved, as well - what is it? Explain.

B. What are **three** data-caused/related/oriented problems that arise in data-driven AI (ie ML)?

Q6 (2 points)

In the first lecture, we talked about data, as 'raw fact'. During the last lecture, during the brief review, we took a parting look at 'data' - what was it? In other words, what do you now know 'data' to be?

Time left Hide

15:58:46

Q7 (2 points)

Horizontal fragmentation of data can help with backup/recovery, and access (ie via CDNs). WHAT ELSE? Briefly explain.

Q8 (2 points)

Pick **two** apps you use often, and explain what type of data they deal in, and how that data might be stored.

Q9 (2 points)

Dataflow can clearly speed up computation, on account of 'dirty propagation' where only affected downstream nodes get re-executed (as opposed to the entire graph of nodes).

What is a non-technical benefit (a pretty big one in fact) that ensues from using visual dataflow graphs (eg from your HW4)?

Q10 (2 points)

Today's BI data analysis/viz can be done on smartphones (eg Salesforce dashboards). Put on your thinking cap - how is BI likely to evolve (what's next)?

Time left Hide

15:58:46

Q11 (2 points)

ML's backprop involves iteration to minimize errors in the model being generated.

Name and briefly discuss two other data mining algorithms that similarly involve iteration for error reduction.

Q12 (2 points)

A. 'ARFF' is "CSV++" - in what sense? In other words, explain how the ARFF data format augments the good old Excel CSV format.

B. WHY make this improvement, ie. what can it help with?

Q13 (2 points)

What makes RDF triple data representation, powerful?

Q14 (2 points)

What extra piece of data can you add to at each location from your HW4, and how would you visualize it (the extra data) on a map? You can answer in general terms (text), and/or provide a small drawing.

Time left Hide

15:58:46

Q15 (2 points)

What was the point about discussing a WIDE variety of data mining/machine learning programs/tools/APIs/libraries?

Time left Hide

15:58:46