CS585 Final

Spring term, 5/6/2020
Duration: 90+30 minutes

Instructions/notes

- the exam is OPEN ANYTHING/EVERYTHING (notes, cheatsheets, devices, Internet...)!!
- the exam is NOT a 'collaborative' one - 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].
- you are required to answer the following questions: **Q1 (0 points), Q2 (5 points), and ANY FIVE of Q3-Q12** (6 points each); you can answer more if you like (and have the time) - the additional questions will bring you additional points, with a cap of 35 points total.
- please answer each question on a separate sheet [you will be uploading each answer separately].
- [for fun: look for the word 'data' in each question]
- you have a LOT of latitude in answering the questions! That said, please do keep the answers relevant.
- DO finish on time (within 90 minutes), and take up to 30 minutes to submit the answers; you will NOT be allowed to send us answers after the test, at all.
- GOOD LUCK! Hope you do well.

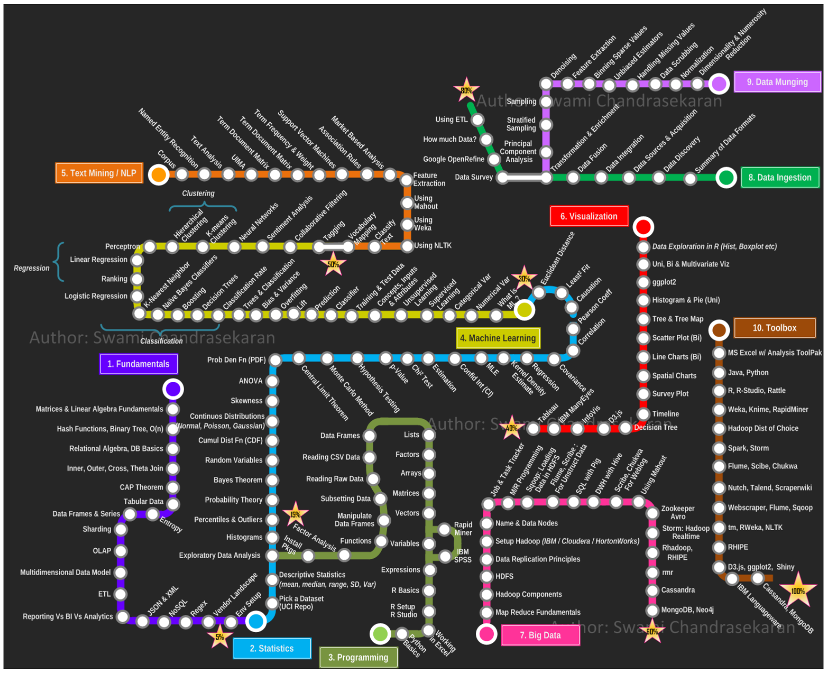
Q1 (0 points)

Please write the following, 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.

Q2 (0.5*10 = 5 points)

Below is a ‘subway map’ of data science - a clever visualization of various “lines” (aspects of data science), with “stops/stations” (topics) along the way - you can search online for ‘RoadToDataScientist1.png’ to find a bigger version, if you like.



Note the ten headings in particular (Fundamentals, Statistics.... Data Munging, Toolbox). **For each heading, looking at our CS585 schedule/syllabus, indicate which of our 15 weeks of topics applies** [there can be just one topic, more than one, or none]; simply put in the week #(s) (1 .. 15) for the relevant topic(s), using a two-column table like so:

Heading	Week#
Fundamentals	...
Statistics	...
Programming	...
...	
...	
Data Munging	...
Toolbox	...

Q3 (6 points)

For database query optimization techniques, one type of classification is based on the information used for the optimizing: statistically-based, versus rule-based. **What does this resemble (remind you of)?** In other words, what major field is based on a similar dichotomy? Explain, in a few sentences.

Q4 (2*3 = 6 points)

In data-warehousing, we start with large amounts of aggregated transactional data (that reside in ‘fact tables’), build a data warehouse (using ETL) , and mine it for ‘BI’ (do analytics). Eg. imagine your fact table consists of these columns of sales data:
Item Type Category Price Discount Date Time StoreID City State Region Referrer
How can a NoSQL column (aka column family) DB be utilized in creating a data warehouse? Answer the following:
a. describe how you might organize the data (be sure to include a simple diagram)
b. a sample query you might perform (no code necessary!)
c. explain how a column DB is suitable for warehousing data

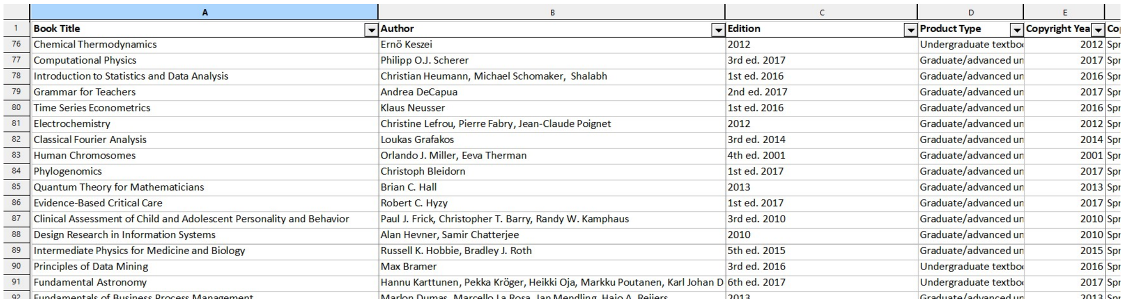
Q5 (6 points)

For your HW3, you collected (long,lat) values for 15 locations in 3 categories, then queried their convex hull and 4 nearest neighbors for a location.

What are three other geospatial analyses/queries you could do, on your collection of data? Explain each, in a couple of sentences (and use diagrams in you like).

Q6 (6 points)

On Piazza, I posted a note about 65 free ML books, and pointed you to this GitHub page that contains a scraper script and a data input spreadsheet for it:
https://github.com/chris-hamberg/springer_books/



	A	B	C	D	E	
1	Book Title	Author	Edition	Product Type	Copyright Yea	Co
76	Chemical Thermodynamics	Ernö Keszei	2012	Undergraduate textbo	2012	Spr
77	Computational Physics	Philipp O.J. Scherer	3rd ed. 2017	Graduate/advanced un	2017	Spr
78	Introduction to Statistics and Data Analysis	Christian Heumann, Michael Schomaker, Shalabh	1st ed. 2016	Graduate/advanced un	2016	Spr
79	Grammar for Teachers	Andrea DeCapua	2nd ed. 2017	Graduate/advanced un	2017	Spr
80	Time Series Econometrics	Klaus Neusser	1st ed. 2016	Graduate/advanced un	2016	Spr
81	Electrochemistry	Christine Lefrou, Pierre Fabry, Jean-Claude Poignet	2012	Graduate/advanced un	2012	Spr
82	Classical Fourier Analysis	Loukas Grafakos	3rd ed. 2014	Graduate/advanced un	2014	Spr
83	Human Chromosomes	Orlando J. Miller, Eeva Therman	4th ed. 2001	Graduate/advanced un	2001	Spr
84	Phylogenomics	Christoph Bleidorn	1st ed. 2017	Graduate/advanced un	2017	Spr
85	Quantum Theory for Mathematicians	Brian C. Hall	2013	Graduate/advanced un	2013	Spr
86	Evidence-Based Critical Care	Robert C. Hyzy	1st ed. 2017	Graduate/advanced un	2017	Spr
87	Clinical Assessment of Child and Adolescent Personality and Behavior	Paul J. Frick, Christopher T. Barry, Randy W. Kamphaus	3rd ed. 2010	Graduate/advanced un	2010	Spr
88	Design Research in Information Systems	Alan Hevner, Samir Chatterjee	2010	Graduate/advanced un	2010	Spr
89	Intermediate Physics for Medicine and Biology	Russell K. Hobbie, Bradley J. Roth	5th ed. 2015	Graduate/advanced un	2015	Spr
90	Principles of Data Mining	Max Bramer	3rd ed. 2016	Undergraduate textbo	2016	Spr
91	Fundamental Astronomy	Hannu Karttunen, Pekka Kröger, Heikki Oja, Markku Poutanen, Karl Johan D	6th ed. 2017	Undergraduate textbo	2017	Spr
92	Fundamentals of Business Process Management	Marlon Dumas, Marcello La Rosa, Jan Mendling, Hajo A. Reijers	2013	Graduate/advanced un	2013	Spr

The spreadsheet, 'Free+English+textbooks.xlsx', contains columns like so:

How would you express the data using JSON, XML, and another format of your choice (an existing one, or one you make up)? Assume we don't need all the columns in Chris' spreadsheet, instead we want just these:
* BookTitle
* Author
* Edition
* ISBN
* Subject Classification
* OpenURL
You don't need to specify the entire data, of course - just indicate the column names, and use '...' as stand-in data - in other words, simply specify the overall format.

Q7 (2+4= 6 points)

Unlike the other three types of NoSQL data storage types (k-v, column family, document), a graph DB cannot be parallel processed using a Hadoop (ie. MapReduce) platform in a straightforward way.

a. why not?

b. what would be needed to make it work?

Q8 (6 points)

How would you rapidly search an inverted index created from a large volume of data? Eg. the data could relate to COVID-19... Explain, using a simple diagram. Hint: @1177 :)

Q9 (1.5*4 = 6 points)

Explain how ‘iteration’ is a useful (algorithmic) design principle, for processing data in four DM/ML techniques - feel free to draw diagrams to illustrate.

Q10 (2+2+2 = 6 points)

In HW5, you used a ‘batch size’ of 16 (FYI we call this, ‘minibatch’), to train your NN using image data for cats and dogs - this means, the errors (losses) from 16 images were aggregated into a single value, and used in back propagation during one epoch. **Why do we do this (use a minibatch)? What if the batch size was set to 2000 (the number of training samples), and what if it was set to 1?** You don’t need to use accurate ML terminology - just answer using your own words, that is sufficient.

Q11 (1*6 = 6 points)

Given the severity of COVID-19 (that has caused an ‘unprecedented’, WORLDWIDE shutdown, the damage from which is going to take years to recover), there is a plethora of data regarding it: virus-related (genome, antibodies, vaccines, cures), disease-related (number of people infected, dead, recovered, tests), economic, stock market, supply chain... **From our ‘catalog’ of visualization techniques (from the lecture), pick any 6, and indicate for each, what COVID-19 data you would use it on (eg. a bar chart, showing ‘top 20’ countries’ infection numbers).**

Q12 (2*3 = 6 points)

COVID-19 data presents issues related to governance, privacy, security - how? **Explain in a few sentences, each of these three aspects.**