CS267 Final Spring 2021

Name:
StudID:

Instructions:

1. This final is due 11:59pm PST, May 22, 2021.
2. To complete the final, print it out, fill in your answers on the final, and scan it back (or take pictures of the pages and make a pdf) into a file Final.pdf where the total size is less than 10MB.
3. If you don't have a printer, copy and paste each problem into a word processor document. Then after each problem write your solution. Make a less than 10MB Final.pdf file of the result and submit that.
4. Use the same submit mechanism as for the homeworks to submit your completed final.
5. Each problem on this final is worth the same amount (3pts).
6. If you have a question on the interpretation of a problem on the final, you can email me at chris@pollett.org.
7. Due to the coronavirus this is an open book, open internet final.
    a. What that means is that you can consult any static (on the order of static for weeks) source of information related to the final material.
    b. You cannot directly or indirectly ask another person how to do any problem off the final.
    c. To receive credit on problems that make use of your personal information, you need to have correctly filled in that personal information.
    d. When you submit your completed final, you are asserting all of the work in the final is your own.

1. Give a map-reduce algorithm to compute the number of occurrences of the year you were born in a corpus of text documents (1.5pts). Give an example including shuffle and combiner steps of your algorithms working on a small corpus of at least five documents of at least five words (1.5pts).

2. Prove $\lim_{f_{t,d} \to \infty} TF_{BM25}(t, d) = k_1 + 1$ (1.5pts). Briefly explain the Maxscore heuristic as used in document at a time query processing (1.5pts).

3. Explain one method to estimate P1 (1.5pts) and one method to estimate P2 (1.5pts) in the divergence-from-randomness approach to coming up with a relevance measure.

4. Give the nextSolution algorithm from class for positive boolean queries (1.5pts). Show how it would work on a concrete corpus and query involving lines from your favorite movie (1.5pts).

5. Give an example corpus involving three documents and the query your full name where the vector space model would score the documents in reverse order of the proximity score for the same documents (1pt working example, 2pts explanation why works).

6. Suppose we have a corpus of a billion documents each of a 1000 terms of average length 7 bytes. Let n be the amount of memory on your laptop (give me the value). Assume we have enough drive capacity, walk me through how sort based index construction could build an inverted index for this corpus on your laptop, giving me details on how many blocks would be written to/ read from disk, how the merging would happen, etc. (3pts)

7. Briefly describe the formats of files needed to compute trec_eval statistics (2pts). Give an example of such files for topics and queries related to the coldest place you have ever been. (1pt)

8. Explain and give an example using personal information (0.5pt explain - 0.5pt example): (a) how the Porter Stemmer works, (b) chargramming, (c) encoding the codepoint Ԉ, U+0508 as UTF-8 (don't need to go personal on this last one).

9. Sort the first five digits of your student ID in increasing order add i to digit i. For example, if your ID was 009781234 then the first five digits would be 00978, in increasing order one would have 00789, and after adding i one would get 1, 2, 10, 12, 14. Let L be the list you get on your ID. Compute its Δ-list, then encode it using a γ-code.

10. Describe the power method for computing the largest eigenvector of a matrix (1.5pts). Explain the three main matrices involved in computing page rank via the power method (1.5pts).