COS 435, Spring 2017 - Problem Set 4

Due 11:59 pm Wednesday April 5, 2017 by DropBox submission

Collaboration and Reference Policy

You may discuss the general methods of solving the problems with other students in the class. However, each student must work out the details and write up his or her own solution to each problem independently. For each problem, list the students with whom you discussed general methods of solving the problem.

Some problems have been used in previous offerings of COS 435. You are NOT allowed to use any solutions posted for previous offerings of COS 435 or any solutions produced by anyone else for the assigned problems. You may use other reference materials; you must give citations to all reference materials that you use.

Lateness Policy

A late penalty will be applied, unless there are extraordinary circumstances and/or prior arrangements:

- Penalized 10% of the earned score if submitted by 10am Thursday (4/6/17).
- Penalized 25% of the earned score if submitted by 4:30 pm Friday (4/7/17).
- Penalized 50% if submitted later than 4:30 pm Friday (4/7/17).

Submission

Submit your solutions as a PDF file using the Computer Science Department DropBox submission system for COS435 at
https://dropbox.cs.princeton.edu/COS435_S2017/HW4 Name your file HW3.pdf. If you have not used this facility before, consult the instructions at
https://csguide.cs.princeton.edu/academic/csdropbox - student
Note that you are automatically enrolled in CS DropBox using the registrar's COS435 enrollment list.

You may hand write your solutions as long as they are legible. In this case, you must scan your writing to produce a PDF file for submission through DropBox.

Problem 1 (similar to a 2011 exam 2 problem)

Recall that skip pointers can be used to speed up query evaluation by allowing the algorithm that executes the intersection of postings lists for the different terms of the query to skip sections of a postings list when then next document on one of the other postings list has a much higher docID. This question asks you to estimate the savings in space if the skip pointer representation is combined with the compressed representation of docIDs using gaps.

Assume that a list containing L postings uses floor(sqrt(L))-1 skip pointers that are approximately evenly spaced, starting at the first posting, so that each skip bridges about sqrt(L) postings.

For a collection of one hundred billion documents, and postings that are pairs (DocID, term frequency), let the representation of one posting in a postings list, using ***no*** compression, be one of the following two forms:

form of posting when there is a skip pointer:
| docID | | skip pointer | | term frequency |
5 bytes 3 bytes 2 bytes

form when there is no skip pointer:
| doc ID | |term frequency |
5 bytes 2 bytes

Part A: Suppose we compress each postings list by representing each destination document of a skip pointer by the difference between its docID and the docID of the origin of the skip pointer (i.e. gap between docIDs). Also represent successive docIDs lying between two skip pointers by their successive gaps in docID (see the illustration of skip pointers in the Compression Summary, Part 2 posted under 3/13/17). All gaps should be represented using ***variable byte encoding***. Also use ***variable byte encoding*** to represent the skip pointer. Do not compress the term frequency. Estimate the space in bytes required for a postings list with this compression. Your estimate should be in terms of L. Your answer should be an estimate of the space used, but it will be graded on the ***quality and correctness of the estimate***, i.e. expect deductions for very coarse estimates.

Part B: For a list of one million postings, how much compression is being achieved with the representation of Part A in comparison to the representation without compression presented at the beginning of this problem?

Problem 2:

Part a: Let D denote a document that is 500 words long and contains each of the words "philanthrepist", "pendantic" and "androgenous" exactly once each, with "philanthrepist" occurring in word position 100, "pendantic" in position 205, and "androgenous" in position 320. Each of these words is misspelled. Let D_{cor} be the document with these spelling errors corrected ("philanthropist", "pedantic" and "androgynous"). What is the value of the resemblance $r(D, D_{cor})$ for a 5-shingling of each document if, for each document, 25% of all possible shingles are repeated shingles?

Part b: Let E denote a document that is 500 words long and contains each of the words "philanthrepist" , "pendantic" and "androgenous" exactly once each but as the phrase "pendantic androgenous philanthrepist" starting at word position 200. Let E_{cor} be the document with the spelling errors in this phrase corrected ("pedantic androgynous philanthropist"). What is the value of the resemblance $r(E, E_{cor})$ for a 5-shingling of each document if, for each document, 25% of all possible shingles are repeated shingles?

Part c: For what threshold or thresholds would one of the pairs (D, D_{cor}) and (E, E_{cor}) be considered near-duplicates and the other not? Which is which?