COS 435, Spring 2016 - Problem Set 4

Due 11:59 pm Wednesday March 30, 2016 by DropBox submission
Due at 5:00 pm, Wednesday, March 30, 2016 if submitting handwritten work on paper.

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 (3/31/16).
- Penalized 25% of the earned score if submitted by 4:30 pm Friday (4/1/16).
- Penalized 50% if submitted later than 4:30 pm Friday (4/1/16).

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_S2016/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 may either scan your writing to produce a PDF file for submission through DropBox or turn in your document by 5:00 PM Wed. March 30 in the bin outside Prof. LaPaugh's office.

Problem 1 (from a 2010 exam 2 problem)

Consider a Web crawler that uses F different priority levels for fetching URLs, based on the frequency of change of the URL. The crawler also uses different minimum delays between requests for different hosts. It will contact a host known to have a large capacity for handling requests more frequently than a host that has less capacity. For each host, h, that has been contacted, the earliest next contact time t_h is recorded. Assume the fetching priorities and minimum delays between requests are independent.

Part A: In the Mercator Web crawler, the URL frontier is managed by two sets of first in first out (FIFO) queues. Give an example in which a crawler must wait before fetching a URL even though there is a URL that could be fetched immediately in one of the front queues. Your example should show as much of the state of the front and back queues as necessary to make clear that the state is legal and crawler is waiting unnecessarily.

Part B: Consider replacing each FIFO **front** queue in the Mercator URL frontier with a priority queue that is sorted on earliest next contact time of the host of each URL in the queue. That is, the next element removed from the k^{th} front priority queues is the URL with the earliest next contact time among all the URLs with fetch priority k. Does this eliminate the situation that the crawler must wait before fetching a URL even though there is a URL that could be fetched immediately in one of the front queues? Does using such a priority queue for each of the F front queues cause new problems? Explain all your answers.

Part C: What information is needed to compute the earliest next contact times for all previously seen hosts? Where is this information stored? Be specific.

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?