Information Retrieval
13 November 2023

Name and Surname: #matricola:

Question #1 [rank 4]. Show how it is compressed by the algorithm WebGraph the posting list of the node 16, with respect to the "best one" of previous posting lists (commenting the choice):

14 -> 2, 3, 5, 16, 19, 22, 24, 26, 28, 44
15 -> 1, 3, 5, 6, 7, 8, 10, 16, 17, 18, 22, 24, 44
16 -> 5, 6, 7, 8, 9, 10, 16, 17, 20, 21, 22, 24, 30

Question #2 [scores 4+5] Given the two files

F_old = "what is good", F_new = "what is so good",

and a block size B=3 chars (*hint:* if the length is not a multiple of B, add NULL chars).

- Describe rsync running on them;
- Describe zsync running on them.

Question #3 [rank 4]. Given 4 strings S = { abaco, basco, raco, vasto }, describe how Z-delta compresses these files via a properly constructed weighted directed graph.

Question #4 [rank 2+3]. Given the dictionary of strings D = {abba, abc, babb} construct a bigram index (hence k=2). Then given the string Q = "abcc" use the overlap distance to filter a set of strings from D that are potential candidate for an edit distance e=1.

Question #5 [rank 2+2] Describe the Front queue and the Back queue in the Mercator crawler, and state/comment their goals.

Question #6 [rank 2+2]

- Describe the algorithm that computes the LSH-sketch of a binary vector for the case of hamming similarity, and show how it is used to declare that two vectors are "similar".
- State and prove what is the probability that the above algorithm declares that two vectors are "similar" provided that their real similarity is s.