Information Retrieval

15 January 2021 – time 45 minutes

Question #1 [rank 5]. Given the two posting lists:

T1 → 1, 2, 3, 5, 6, 7, 8, 10, 15

T2 → 9, 10, 11, 16, 18, 19, 20, 23, 25

Assume that they are divided into logical blocks of three items each, and they have skip pointers to the beginning of blocks. Assume that the last block of each list has a skip pointer of value +infty.
Show the pairs of elements which are compared by the algorithm that computes the intersection of the two lists and exploits the skip pointers. Stop as soon as a first match is found (namely at 10).

Question #2 [rank 5]. You are given the sets A={1, 4, 6, 8}, B = {2, 4, 6, 7, 8}, C = {4, 6, 11}, apply MinHashing technique based on a sketch of size 2 (integers – minima), and the following permutations p1(x) = 2*x mod 13 and p2(y) = 3*y mod 13. Estimate the Jaccard similarity among the three sets.

Question #3 [rank 5+5]. Let us given the dictionary of strings D = {bas, box, bus, cus}**.**

- Construct a data structure that efficiently solves the **1-error search**
- Apply it to search for the pattern P=rox.

Question #4 [rank 5]. Given the four files { aba, aaba, aabb, baaa }, apply the Z-delta algorithm to compute the best compression of this group of files.

Question #5 [rank 3]. Assume that a client is executing the zsync algorithm and stores the file f_old = rane_matte, and receives from the server the hashes h1, h2, h3, h4 that it finds as h1 = att, h2=ran, h4=ane.
Thus the client answers to the server with the bit string 1101 and then it gets <5,3,EOF>.
Show which is the f_new file reconstructed by the client.

Question #6 [rank 2]. Assume that you are given n users, each with a set of movies U_i that (s)he has seen, drawn from a universe U of movies. Design a solution that, given a set of movies S, finds the users which share at least k movies with S. Comment the efficiency and efficacy of the proposed solution.