Information Retrieval

16 November 2020 – time 45 minutes

Question #1 [rank 6]. Given the three adjacency lists compress them via the Web graph algorithm by choosing always the best previous list to differentially encode the current one. (So no bound is set on the backward window to "copy".)

14 → 3, 10, 11, 13, 14, 17, 19, 21, 25
15 → 5, 10, 11, 13, 14
16 → 3, 10, 11, 13, 21, 25, 30

Question #2 [rank 6]. You are given the following five binary vectors:

A = 01100, B = 00000, C = 11111, D = 00010, E = 10010.

Compute the groups of similar vectors using the clustering algorithm based on LSH and connected components, by assuming k=2 and L=2, with projections I1={1,2} and I2={3,4}.

Question #3 [rank 3+4]. Let us given a set of strings S = { bud, budy, turdy, tus }.

- Build a 2-gram index of S
- Given P = "tusdy", show how the index executes the search for 1-edit errors based on the 2-gram index, a filter based on overlap distance, and then computing the dynamic programming matrix.

Question #4 [rank 3+5]. You are given the two files:

F_old = "cane ratto matt", F_new = "cane gatto pane",

Assume a block size B=3 chars, and

- Show the execution of the algorithm *rsync*.
- Show the execution of the algorithm *zsync*.

Question #5 [rank 3]. Compress the string S = bababc via LZ77-gzip by assuming window w=3