Information Retrieval

7 February 2022 – time 60 minutes

Question #1 [rank 5+4]. Given the sorted sequence of integers S = (2, 4, 6, 10, 12)

- show how to compress it via Elias-Fano code.
- Show how to compress it via PForDelta by taking b = 2 and choosing a proper transformation of the numbers and a proper "base".

Question #2 [rank 3+3]. Given the keys S = {1, 2}, and two hash functions h1(k) = 3*k mod 5, h2 = 2*k mod 5.

- Construct a Spectral Bloom Filter of possibly two levels, each consisting of an array of size 5;
- Discuss what happens at the insertion of the key 3. Is the second level of the SBF being constructed? Motivate the answer.

Question #3 [rank 5]. Given a set of binary vectors V = {00011, 00100, 01010, 10011}, and projections I1 = {1,2} and I2 = {2,3}, where index positions are counted from 1. Find the most similar vectors according to the Hamming distance and the use of LSH.

Question #4 [rank 3+3]. Given the dictionary of strings D = {aabb, abc, acac} construct a bigram index (hence k=2) and then search the string Q = "aabc" by assuming an edit-distance error e=1.

- Use the overlap distance to filter a set of candidates for the parameters k=2 and e=1, relative to Q and S's strings.
- Then compute via dynamic programming the edit distance between the shortest candidate and Q.

Question #5 [rank 4]. You are given the following posting lists for three terms:

T1 -> 2, 4, 5
T2 -> 3, 4
T3 -> 2, 3, 4

Show the triples of docIDs compared by the algorithm that counts how many terms are included in each of these documents (i.e. {2, 3, 4, 5}).