Information Retrieval – exercises
17 January 2023 – time 60 minutes

Name and Surname: #matricola:

Question #1 [scores 6] Given the sorted sequence of integers S = (4, 6, 10, 12, 17, 24)

- Show how to compress the gaps between consecutive S's integers via the gamma-code
- Show how to compress S via Elias-Fano code.
- Show how to compress S via PForDelta code by first shifting them via base=4, and then taking b = 2 to encode the resulting gaps.

Question #2 [rank 4]. Given the set V = {00000, 00100, 01001, 01101, 10000, 10111}, and the projections I1 = {1,2}, I2 = {2,3}, and I3 = {4,5}, where index positions are counted from 1, find the most similar vectors according to the Hamming distance and the use of LSH+graph_clustering.

Question #3 [rank 6]. Given the dictionary of strings D = {aacc, acb, abab} construct a bigram index (hence k=2) and then search the string Q = "aacb" by assuming an edit-distance error e=1. More precisely,

- 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 string and Q.
- Show what happens if you use the efficient solution that works just for e=1 errors to perform the query for Q = "aacb"

Question #4 [rank 4]. Consider the WAND algorithm for examining the head of the following four posting lists:

t1 → (5, 6, 7, 8, 11)
t2 → (2, 3, 5, 7, 8, 11)
t3 → (1, 4, 6, 7, 13, 15)
t4 → (6, 7, 8, 11)

The current threshold is 2.2, and the upper bounds of the scores in each posting list are:
ub_1 = 0.4, ub_2 = 1, ub_3 = 0.6, and ub_4 = 0.5.
Which is the next docID whose full score is computed? (Motivate your answer)

Information Retrieval – theory
17 January 2023 – time 45 minutes

Name and Surname: #matricola:

Question #1 [scores 4] State the formulas underlying the PageRank algorithm and the HITS algorithm, and then comment on their differences.

Question #2 [rank 3] Define formally what is the Permuterm index, and comment on the type of queries it solves.

Question #3 [rank 3] Define the measures: precision, recall, F1, and DCG.