Information Retrieval – exercises
7 September 2023 – time 60 minutes

Name and Surname: #matricola:

Question #1 [scores 6] Given the sorted sequence of integers S = (3, 6, 10, 12, 17, 27)

- 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 its numbers with base=3, and then taking b = 2 to encode the resulting gaps.

Question #2 [rank 5]. Given the set V = {00000, 00100, 01001, 01101, 10000, 10111}, and the projections I1 = {1,2}, 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+graph_clustering.

Question #3 [rank 6]. Given the dictionary of strings D = {bbcc, bcb, bbbb} construct a bigram index (hence k=2) and then search the string Q = "bbcb" 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 seen in class that works just for e=1 errors to perform the query for Q = "bacb"

Question #4 [rank 3]. Describe rsync, with a block size B=3 chars, running on the following two files: F_old = "il cane bello", F_new = "il pane bello".

Information Retrieval – theory
7 September 2023 – time 60 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 4] Define formally what the Permuterm index is, and comment on the type of queries it solves.

Question #3 [rank 4] Define the measures: precision, recall, F1, and DCG.