Information Retrieval – exercises
05 July 2023 – time 60 minutes

Name and Surname: #matricola:

Question #1 [scores 3] Compute the size in MB needed by a Bloom filter for achieving an error probability of e^{-10} on $n = 2^{20}$ objects, assuming an optimal number of hash functions is used.

Question #2 [scores 3+3] Assume you are given the following Elias-Fano encoding of a sequence of integers:

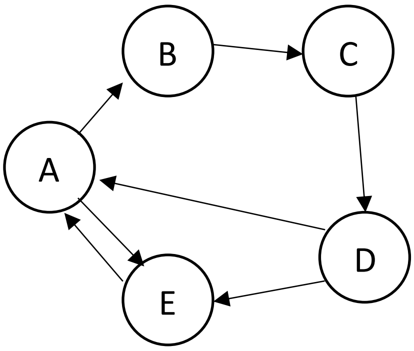
L = 01 10 11 01 10 00 00 11
H = 1110101010000110

- Show and explain how many integers are encoded, and which is the number of bits used by the original encoding for each integer.
- Decompress the 5th integer.

Question #3 [scores 3+3] Given the dictionary of strings D = {babc, bcaa, cab} construct a bigram index (hence k=2) and then search the string Q = "bcab" 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 #4 [scores 3+2] Given the following graph:



- Compute the personalized PageRank for the node E by assuming a starting distribution [1/5, 0, 1/5, 1/5, 2/5] and alpha = 0.5. [WARNING: the starting distribution is not the uniform one.]
- Comment on whether a random walk computed over this graph is converging to a single state which is independent of the starting distribution.

Information Retrieval – theory
05 July 2023 – time 45 minutes

Name and Surname: #matricola:

Question #1 [scores 2+2] State the:

- Zipf’s law
- Heaps’ law

Question #2 [scores 2+2] Describe:

- The LSH-sketch for the Hamming distance between two vectors.
- The LSH-sketch for the cosine distance between two vectors.

Question #3 [scores 2] Write the tf-idf formula, and comment on it.