Information Retrieval

13 June 2022

Name: **Surname:** **Matricola:**

Ex 1[points 3]
Given a dictionary of $n=2^{16}$ strings, compute the error rate of a Bloom Filter which uses an array of 2^{20} bits and an optimal number of hash functions. *[Assume that logs are in base 2]*

Ex 2 [points 4+3+3] Let us given a set of strings S = { dad, atom, momo, oma }.

- Build a 2-gram index over S
- Given pattern P = mom, show how the index executes the search for 1-edit error
- Given pattern P = mom, show how the index executes the search for 2-edit errors

Ex 3 [points 5] Consider the WAND algorithm over four posting lists by assuming that at some step the algorithm is examining the heads of the following lists:

t1 → (…, 5, 6, 7, 8, 11)
t2 → (…, 2, 3, 5, 7, 8, 11)
t3 → (…, 8, 13, 15)
t4 → (…, 4, 5, 8, 9)

At that time the current threshold equals 2.3, and the upper bounds of the scores in each posting list are: ub_1 = 0.4, ub_2 = 2, ub_3 = 4, ub_4 = 0.1.
Which is the next docID whose full score is computed? *(Motivate your answer)*

Ex 4 [points 4+4+4] Show the compressed encoding of:

- The integers 7 and 18 with DELTA-code
- the sequence (1, 3, 2, 4, 2, 3, 1, 1, 2, 9) with PForDelta: base=0 and b=2 bits
- the sequence (1, 4, 6, 10, 12, 15) via Elias-Fano.