**Information Retrieval**

**15 February 2019**

**Name: Surname: Matricola:**

**Ex 1 [points 5+5+5]**

* Define what is a Bloom filter, the operations supported, and show the formula of the error rate (as a function of the space and number of keys) and prove its correctness.
* Define what is min-hashing and how it is used to estimate the Jaccard similarity between two sets A and B.
* Define what is two-level indexing and why it is introduced in the external-memory setting, when the internal-memory has size M and the page size is B.

**Ex 2 [points 4+2+3]** Consider the WAND algorithm over four posting lists by assuming that at some step the scanning algorithm is examining

t1 🡪 (…, 4, 9, 10, 11)

t2 🡪 (…, 2, 3, 4, 5, 7, 8, 11)

t3 🡪 (…, 8, 13, 15)

t4 🡪 (…, 3, 4, 7, 8, 9)

At that time the current threshold equals 3.1, and the upper bounds of the scores in each posting list are: ub\_1 = 1, ub\_2 = 2, ub\_3 = 4, ub\_4 = 2.3.

* Which is the next docID whose full score is computed? *(Motivate your answer)*
* Let us assume that the full score computed for the item selected at the previous point is 3.3, what happens?
* Which is the next docID whose full score is computed? *(Motivate your answer)*

**Ex 3 [points 3+3]** You are given a binary tree T formed by n=5 nodes {a,b,c,d,e}, rooted in “a”, and having the following edges {(a,b), (a,c), (b,d), (c,e) }, where “d” is the left child of “b” and “e” is the left child of “c”.

* Show the succinct encoding of T (recall that it takes 2n+1 bits).
* Describe how to follow the path that starts from the root “a” and then goes right to “c” and finally goes left to “e”.