**Information Retrieval**

**27 July 2017**

**Name: Surname: Matricola:**

**Ex 1 [points 4+3+3]** Let us given a set of strings S = { dad, atom, mamo, oma, zoo }.

* 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 2 [points 4]** Compute the Permuterm index for the two strings {BIS, POS} and show how it is possible to search for P\*S into it.

**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 5+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”.

**Ex 5 [points 3]** Show the formula of CoSimRank and comment on its use and significance.

**[LAB TEST]** Let us assume that we have built a Lucene index, with a Whitespace Analyzer, over the following three documents:

* d1 = "The sea Mediterraneo is a very well known sea close to Italy."
* d2 = "Mediterraneo is a sea in front of Italy."
* d3 = "the name mediterraneo is for a sea!"

and then assume that you execute the following query with a Whitespace Analyzer *(hint: keep attention to the lower/upper cases, spaces,…):*

* q1 = "sea Mediterraneo”

Show which documents will be returned and in which order, commenting on your assumptions about the term frequencies.