**Information Retrieval**

1. **January 2017**

**Ex 1 [points 4]** Describe the role of the min-heap data structure in Mercator.

**Ex 2 [points 4+4]** You are given the two files: F\_old = “il mio cane bello”, F\_new = “il miao di Nello”, and assume a block size B=3 chars.

* Show the execution of the algorithm rsync. *(comment the various steps)*
* Show the execution of the algorithm zsync. *(comment the various steps)*

**Ex 3 [ranks 4]** State and prove how the hamming distance between two binary vectors v and w of length n is related to the probability of declaring a “match” between their LSH-fingerprints, according to the various parameters involved.

**Ex 4 [points 5]** Consider the WAND algorithm over four posting lists by assuming that at some step the scanning algorithm is examining

t1 🡪 (…, 4, 6, 7, 8, 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 = 0.2, ub\_2 = 2, ub\_3 = 4, ub\_4 = 0.3.

Which is the next docID whose full score is computed? *(Motivate your answer)*

**Ex 5 [points 4+5]** Given the directed graph G consisting of nodes {A, B, C, D} and edges {(A,B), (B,C), (D,C), (B,A)}:

* Compute one step of the PageRank of G’s nodes by assuming that the teleportation step occurs with probability 0.5 and the starting probability distribution is uniform and equal to (¼, ¼, ¼, ¼).
* Comment how the similarity between node B and all the other G’s nodes can be estimated by using Personalized PageRank. Apply your algorithm over G for one step only and with the teleportation step occurring with probability 0.5.

**Ex [lode]** Describe an LSH function for the cosine similarity and motivate its design and correctness.