**Information Retrieval**

**4 April 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.
* Show how the Jaccard similarity between two sets A and B can be estimated via min-hashing, as a function of the number of extracted minima.
* Define Link Probability and Commonness as they are used by entity annotators, like TagMe.

**Ex 2 [points 4+3]**

Given the three strings s1 = “asta”, s2 = “balla”, and s3 = “basta”, show the compression

* by gzip with **window size w=5 chars** when concatenated as s1 s2 s3,
* by Zdelta applied to the set of those strings deriving their concatenation order via the approach based on a weighted graph where the cost of edges is estimated via gzip. *(Hint: refer to “compression of a group of files” by Zdelta)*

**Ex 3 [points 4+4]**

Given the graph



* Execute one step of the PageRank algorithm, assuming uniform teleportation step, alpha = ½ and uniform starting probability.
* Execute one step of the Personalized PageRank algorithm with respect to the node 1, assuming alpha = ½ and uniform starting probability.