**Information Retrieval**

**29 October 2018**

**Name: Surname: Matricola:**

**Ex 1 [points 5+3]**

Given a set S of strings {bio, bionic, bit, bitly, buzz, car, caso, cast, zoo},

* show the 2-level indexing of S with Front Coding, by assuming a page size on disk able to contain 3 strings.
* show how it is executed the search for the string P=bis.

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

Given the three strings s1 = “abaco”, s2 = “dada”, s3 = “coaba”, show the compression

* by gzip with **window size w=3 chars** when concatenated as s1 s2 s3,
* by Zdelta applied to the group of those strings using gzip as basic compressor and deriving their concatenation order via the approach based on a weighted graph. *(Hint: refer to “compression of a group of files” by Zdelta)*

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

Answer to the following questions:

1. Describe the role of the Priority Queue in Mercator, and how it is populated.
2. Describe the search-engine indexing based on term-partitioning and the one based on document-partitioning, and then discuss their pro/cons with respect to simplicity and query efficiency.
3. Describe how LSH is applied to find binary vectors which are close according to the hamming distance, and then compute the probability that two vectors with hamming distance H are declared “similar” by LSH.
4. 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.