**Information Retrieval**

**30 October 2017**

**Name: Surname: Matricola:**

**Ex 1 [points 4+3]** Let us given a set of strings S = { pitom, dad, daddy, zoom }.

* Build a 2-gram index over S
* Given pattern P = atom, show how your index executes the 1-edit error search in S.

**Ex 2 [points 4+4]** You are given the files: F\_old = “cane gatto orso”, F\_new = “matto cane ratto”, 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 [points 5]** Given the two adjacency lists of a Web graph for the nodes 14 and 15, namely

14 🡪 3, 10, 11, 13, 14, 17, 19, 21, 25

15 🡪 5, 10, 11, 12, 14, 17, 19, 20, 21, 24, 33

show how the list of “15” can be compressed given the list of “14” by means of the algorithm adopted in WebGraph via copy-lists and copy-blocks.

**Ex 4 [points 5]** You are given the sets A={1, 4, 6, 8, 9}, B = {2, 3, 4, 6, 7, 8}, C = {2, 5, 11, 12}, show how MinHashing technique (in the context of Locality Sensity Hashing) can be used to estimate the Jaccard similarity among them, based on a sketch of size 2 (integers – minima), and assuming that sets A, B, C are drawn from the universe {0, 1, …, 11, 12}.

**Ex 5 [points 5]** Describe the data structure and query algorithm that solves the 1-error match in a dictionary of strings.

**Per i fuori corso sostituire l’esercizio 1 con il seguente**

Given the directed graph G consisting of nodes {A, B, C, D} and edges {(B,A), (A,C), (D,C), (A,B)}:

* 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.
* 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.