**Information Retrieval**

**15 January 2018**

**Name: Surname: Matricola:**

**Ex 1 [points 5+4]**

Given the sequence of integers S = (2, 5, 6, 11, 14, 16, 20, 38),

* show how to encode them based on Elias-Fano by detailing the formulas you use to compute the parameters *z* and *w*.
* design a binary-search algorithm over the compressed S that uses Select\_1 queries on H and simple accesses on L.

**Ex 2 [points 5+4]**

Describe the shingling technique and how it can be used, in combination with min-hashing, to compute the near-duplicate similarity of documents, by commenting also the advantages.

Apply your description to the following documents d1 = “the box is red and large”, d2 = “the large box is red”, d3 = “the fox is red”. Assume that you use shingles of 3 words, hash functions generate random values in {0,…,10}, and repeat min-hashing twice for estimating the Jaccard coefficient of two sets.

**Ex 3 [points 5+2]**

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

* Compute two steps of PageRank by assuming uniform starting probability and parameter ¼ for the teleportation step (*hint:* try to operate only on fractions).
* Assume now that every node has some textual content. Propose an algorithm that extends your previous approach taking now into account also the text content of nodes.

**Ex 4 [points 5]**

Describe the technique of Champion Lists to compute fast the top-k documents answering an AND query between two terms.