**Information Retrieval**

1. **September 2016**

**Ex 1 [points 5]** Describe the main modules which constitute a crawler and detail their pseudo-code

**Ex 2 [points 5+5]** Given Z-delta’s approach to compressing pairs of files,

* Describe how Z-delta works by assuming that f\_known = babbo and f\_new = nababbi.
* Let us assume that you are given 5 strings S = { abaco, baxco, taco, zaxo }, describe how Z-delta compresses these files via a properly constructed weighted directed graph.

**Ex 3 [ranks 4]** Given the keys S = {2, 3, 6, 9, 1, 5}, construct a cuckoo hash table consisting of two arrays of size 5 each, and hash functions h1(k) = 3\*k mod 5, h2 = k+2 mod 5. Discuss whether the insertion of S’s keys is safe or some blocking condition occurs.

**Ex 4 [rank 4+4+3]** Given a set of binary vectors,

* describe how it is computed an LSH-fingerprint of a binary vector v to fast to approximate the hamming distance between vector-pairs with high probability. (*hint: state clearly the parameters involved in the setting.)*
* State and prove how the hamming distance between two vectors v and w is related to the probability of declaring a “match” between LSH-fingerprints of v and w, according to the various parameters involved.
* Apply your description on the set of vectors V = { 00000, 00100, 01010 }, and use it to estimate Hamming(00000,01010) and Hamming(00000,00100)

**Ex [lode]** Specify the Mercator’s approach (back queue and front queue management).