**Information Retrieval**

**08 January 2014**

**Ex 1 [ranks 3]** Describe and comment the cost function used in RankNet.

**Ex 2 [ranks 3]** Given the string S=baba, compute its Arithmetic encoding based on empirical frequencies.

**Ex 3 [points 4+3]** Given the following matrix of pair-wise similarity between 5 items



Show the next cluster formed by the agglomerative clustering algorithm based on MIN and MAX similarity-functions, given that we have already formed the following clusters {(I1,I2), (I3,I4), (I5)}

**Ex 4 [points 4+3]** Describe the Permuterm-index of a set of n strings, each of length L, and comment how to index those strings in order to execute efficiently the query P\*Q, where P and Q are any substrings. Also comment on the time and space complexity of your proposed solution given the parameters n, L, and the lengths |P| and |Q|.

**Ex 5 [points 3]** Compute the probability of error for a Bloom Filter that uses a bit-array of m bits, k hash functions and indexes a dictionary of n strings.

**Ex 6 [points 3+4]** Assume that you have 7 items (6,1,3,7,2,4,9) and three servers whose IDs are (1,2,3), and you want to distribute those items among the three servers via *consistent hashing.* How would you do? *(hint: Instead of working on ring [0,1], work on the integers in {0,1,..,10} using a universal hash function h\_a(x)= a\*x mod m)*