**Information Retrieval**

**16 January 2015**

**Ex 1 [ranks 4+2]** Given the string S=bbba, compute its Arithmetic encoding based on the empirical frequencies of symbols ‘a’ and ‘b’, and compare against the empirical entropy of S.

**Ex 2 [points 3+5]** Given the matrix of pair-wise similarities between five items



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

* MAX
* SECOND MINIMUM, which defines the similarity between a pair of clusters as the **second minimum distance** between all pairs of items of the two clusters

**Ex 3 [points 3]** Discuss the pro and cons of using skip pointers in AND-queries implemented via inverted lists.

**Ex 4 [points 3+3+2]** Given the string S=ababca, construct its suffix array and then show how it is executed the search for a string P=ab.

Indicate also the space in bits used by a Suffix Array built on a string of n characters, and the time complexity needed to search there for a string of p characters.

**Ex 5 [points 3+2]** Given the sequence of keys S = (1, 7, 4, 3, 11, 13, 12) insert them via cuckoo hashing in two tables of size m=5, and hash functions h1(k) = k mod 5, h2(k) = 3\*k mod 5. Comment also on the possibility to insert the key 6.