**Information Retrieval**

**16 July 2013**

**Ex 1 [ranks 3]** Discuss the advantages (if any) of cuckoo hashing with respect to the classic hashing with chaining in terms of space, update time and search time.

**Ex 2 [rank 3+2+3]** Assume you are given 6 strings (aa, ab, bb, bc, ca, db) and you wish to construct a bloom filter of size 11 bits, with two hash functions defined as follows. Assume that rank(x)=1,2,3,4 for the characters x=a,b,c,d respectively. Given a string x’x’’ of two characters, we let the two random functions as h1(x’ x’’) = rank(x’) \* rank(x’’) mod 11 and h2(x’ x’’) = rank(x’) + rank(x’’) mod 11.

1. Show the resulting bloom filter
2. Compute the error probability given its structure: k=2, m=11, n=6.
3. Comment whether or not it may incur in error when checking the strings (ac, ad, dc).

**Ex 3 [ranks 2+3]** Show the output of LZ77 compressor over text T=abbacab,

* when the shifting windows is unbounded
* when you can copy from at most 3 characters back.

**Ex 4 [ranks 3]** Given the string S=abba, compute its Arithmetic encoding based on empirical frequencies. (*hint*: you can work with dyadic fractions.)

**Ex 5 [ranks 4]** Prove the correctness of the formula used to truncate the output of the Arithmetic coder.

**Ex 6 [points 4+3]** Let us given the following matrix of 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 the following clusters formed {(I1,I2), (I3), (I4,I5)}

**[rank \*]** Prove that the probability to have a path of length L between two nodes in a cuckoo-hash table of size if r > 2cn is c^L/r.