**Information Retrieval**

**10 January 2013**

**Exercises**

1. **[ranks 5]** Given the string S=abaa, compute its Arithmetic encoding based on empirical frequencies. (*hint*: you can work with dyadic fractions.)
2. **[ranks 5]** Indicate the formula describing how many bits have to be emitted by Arithmetic coding and prove why this number ensures the correctness of the algorithm.
3. **[ranks 4]** Given a dictionary of 216 strings, compute the error rate of a Bloom Filter which uses an array of 220 bits and an optimal number of hash functions. *[Assume that logs are in base 2]*
4. **[ranks 4+4]** Describe LSI and show how to project a query-vector q in the new topic space, motivating the formula.
5. **[ranks 4]** Design an index that solves the following query over a set of strings S of length greater than 4: return all strings P\*Q where |Q|=3 and P can be of variable length, possibly 0. Comment on the space occupancy and the time complexity of solving this query.
6. **[ranks 4]** Assume you are given K bit vectors of size N each. Design a solution that preprocesses the vectors making use of the min-wise permutation scheme in a way that, given a new binary vector Q at query time, finds the vectors in the input set which are highly-similar to Q.

Compute space and time complexity of the query operation.