**Information Retrieval**

**10 September 2015**

**Ex 1 [rank 2]**  Define Precision, Recall and F1.

**Ex 2 [ranks 4]** Given the binary sequence B=110000101 representing the Arithmetic encoding of a string S with p(a)=p(b)=0.2, p(c)=0.6, decode its first 3 characters.

**Ex 3 [rank 3+3]**  Let us given a Bloom Filter. Specify the function which relates error probability of a search with k-hashes, the size m of the bloom filter and the number of hashed keys n. Then indicate the optimal number of hash functions k, for a given n and m, and comment how this is computed.

**Ex 4 [points 4]** 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 MAX similarity function.

**Ex 5 [points 3+2+3]** Consider the E-edit distance search problem over a dictionary D of strings of variable length.

1. Show the algorithm that uses a k-gram index for restricting the E-approximate search over a subset of candidate strings within D.
2. Show the 3-gram index for D={barca, bit, borda, cara, dazio, loda, zobe} by padding the beginning of every string with $, so that every string of length L has exactly L 3-grams.
3. Solve the 2-edit distance search for P=basta over D using the previous 3-gram index and the algorithm of question 4.1.

**Ex 6 [ranks 3+3+\*\*]** Given the string S=brabrarab,

* compute the BWT;
* describe how to augment BWT(S) in order to decompress any substring of S, by assuming to store 3 more integers.
* **[Lode]** Follow the algorithm of the previous item and show how to decompress the substring S[4,5].