**Information Retrieval**

**20 July 2015**

**Ex 1 [ranks 3+3]** Describe the iterative algorithm that binary encodes a fractional number in [0,1), and then comment on its correctness.

**Ex 2 [points 3]** State the Johnson-Linderstrauss lemma and comment on its usefulness in terms of document comparison.

**Ex 3 [points 3+3]**

* State the difference between hard and soft clustering.
* Show a graphical example of points for which a k-means clustering algorithm (with k=2) converges to two different partitionings, depending on the positions of the two starting seeds.

**Ex 4 [points 2+3+3]**

* Define the TF-IDF(w,d) weight for a word w in a document d
* Given the four texts:
	1. T1=”la bella casa”
	2. T2=”le belle cose”
	3. T3=”casa bella bella”
	4. T4=”la casa le cose bella rosa”

Show the inverted list with its dictionary and postings, compute the vectors TF-IDF of the four texts. Indicate the text which is more similar to T3 under the dot-product distance.

* Show how the TF-IDF matrix is space-efficiently stored in the posting lists.

**Ex 5 [points 3+4]**

* Describe the algorithm that solves the top-k in auto-completion search, provided that k is small and fixed at preprocessing time.
* Describe the algorithm that solves the top-k in auto-completion search, provided that k is **not known** at preprocessing time and possibly large (*hint:* you need an additional data structure, which one?).