**Information Retrieval**

**25 June 2013**

**Exercises**

1. **[ranks 5]** Describe the fancy-hits heuristic to return the top-k documents for a conjunctive query of two terms when TF-IDF and Pagerank scores are taken into account.
2. **[rank 3+4]** Describe the principles behind the agglomerative clustering algorithms, their pro&cons. Describe K-means, and show the why it reaches a local minimum.
3. **[ranks 4+3]** Given a text T = abab.
   1. Compute the compressed output produced by Arithmetic coding
   2. Provide other 2 texts that have the same compressed length of T, motivate the answer without compressing the texts themselves.
4. **[ranks 4+4+3]** You are given the texts:
   1. T1=”a beautiful book”
   2. T2=”book after book”
   3. T3=”after a beautiful girl a book”
   4. T4=”girl after girl”

Show the inverted list for these texts by using gamma-coding for the docID gaps.

Compute the TF-IDF vectors of the four texts above (logs are in base two).

Find the most similar text to T3 in the vector space model (no normalization).

**[rank \*]** Discuss how it could be chosen the output of Arithmetic encoder by looking only at the binary representations of bin(L) and bin(L+s), and motivate the answer.