**Information Retrieval**

**12 February 2013**

**Exercises**

1. **[ranks 3]** Describe the two forms of caching adopted by search engines, and comment when they are advantageous.
2. **[ranks 4]** Indicate the Rocchio’s formula and comment on its meaning.
3. **[ranks 4]** Describe a semi-external memory algorithm to perform a BFS-visit of a huge graph, by assuming that the internal memory size M > n, where n = #vertices of the graph.
4. **[ranks 2+4]** Define what is a Strongly Connected Component of a directed graph, and then design an algorithm to compute all of them by deploying the BFS procedure of item 3. Compute also its I/O-complexity.
5. **[ranks 2+3]** Define the notions of precision, recall and F1; then plot the precision/recall curve for the following 10 results, incrementally considering each of them (on the left it is shown if a result is relevant or not):

REL res1

REL res2

Not res3

Not res4

REL res5

Not res6

REL res7

Not res8

Not res9

REL res10

1. **[ranks 2+3+3]** You are given the texts:
   1. T1=”a beautiful day”
   2. T2=”day after day”
   3. T3=”after a beautiful girl a day”
   4. T4=”girl day 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).