Multimedia Information Extraction and Retrieval
Summer Term 2012
Exercise Sheet 1

Ralf Möller, Karsten Martiny

Exercise Session:
Thursday, April 19, 2012, 8.00-8.45 , SBS95-D1025

1. Explain why naive matrix-representations for term incidence vectors are not feasible in practice. What is the underlying idea for alternative representations?
2. Explain the notions of *dicitionary files* and *posting files*.
3. Why are postings sorted by document ID?
4. What is the idea of *tokenization*? What are the major issues/problems?
5. Are *stop words* still used nowadays when creating dictionaries?
6. Which approaches for optimizing merge operations do you know?
7. Which approaches for answering *phrase queries* do you know, and what disadvantages do they have?
8. Are the following statements true or false?
   - In a Boolean retrieval system, stemming never lowers precision.
   - In a Boolean retrieval system, stemming never lowers recall.
   - Stemming increases the size of the vocabulary.
   - Stemming should be invoked at indexing time but not while processing a query.
9. Give an example of two english words, which are *mistakenly* conflated by porter's algorithm. Argue why they shouldn't be conflated.
10. Apply porter's algorithm to the terms *university*, *universe*, *marketing*, *markets* and *pony*. If you find any problems, discuss how to adapt the rules of porter's algorithm.
11. Inverse porter's algorithm and guess the original text for the following output of porter's algorithm: „stemmer: such an analysi can reveal featur that ar not easili visibl from the variat in the individu".

12. We have a two-word query. For one term the postings list consists of the following 16 entries:

    [4,6,10,12,14,16,18,20,22,32,47,81,120,122,157,180]

    and for the other it is the one entry postings list:

    [47].

    Work out how many comparisons would be done to intersect the two postings lists with the following two strategies. Briefly justify your answers:

    - Using standard postings lists
    - Using postings lists stored with skip pointers, with a skip length of $\sqrt{|L|}$

13. Assume a biword index. Give an example of a document which will be returned for a query of „New York University“ but is actually a false positive which should not be returned.

14. Given the following positional posting list term : doc1 (pos1, pos2, pos3); doc2 (pos1, pos2, pos3) .... :

    - angels: 2: (36,174,252,651); 4: (12,22,102,432); 7: (17);
    - fools: 2: (1,17,74,222); 4: (8,78,108,458); 7: (3,13,23,193);
    - fear: 2: (87,704,722,901); 4: (13,43,113,433); 7: (18,328,528);
    - in: 2: (3,37,76,444,851); 4: (10,20,110,470,500); 7: (5,15,25,195);
    - rush: 2: (2,66,194,321,702); 4: (9,69,149,429,569); 7: (4,14,404);
    - to: 2: (47,86,234,999); 4: (14,24,774,944); 7: (199,319,599,709);
    - tread: 2: (57,94,333); 4: (15,35,155); 7: (20,320);
    - where: 2: (67,124,393,1001); 4: (11,41,101,421,431); 7: (16,36,736);

    Which document(s), if any, match each of the following queries, where each expression within quotes is a phrase query?

    - „fools rush in“
    - „fools rush in“ AND „angels fear to tread“