Multimedia Information Extraction and Retrieval SoSe 2010
Exercise Sheet 2

Prof. Dr. Ralf Möller, Sebastian Wandelt

28.04.2010

1. 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.

   Solution:

   - *false*
   - *true*
   - *false*
   - *both*

2. Give an example of two english words, which are *mistakenly* conflated by porter's algorithm. Argue why they shouldn't be conflated.

   Solution:

   - *university/universe*

3. 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. **Solution:**

   - *university:univers*
   - *universe:univers*
   - *marketing: market*
   - *markets: market*
   - *pony:poni*

4. 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".

   Solution:

   - *„Such an analysis can reveal features that are not easily visible from the variations in the individual"*

5. 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 $\sqrt{|L|}$

   Solution:

   - *11 comparisons*
   - *skip length* $\sqrt{|L|} = 4$ *... items skipped: 6,10,12,16,18,20,*

6. 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.

   Solution:

   - *„New York is full of interesting sites. ... The York University is located in Toronto.“*

7. 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“

Solution:

- *see exercise, e.g. document 2*
- *see exercise, e.g. document 4*

8. Why is the idf of a term always finite?

   Solution:

   - *by construction of the formula (df is finite and smaller than n)*

9. What is the idf of a term that occurs in every document? Compare this with the use of stop word lists.

   Solution:

   - $log(\frac{n}{n}) = 0$
   - *does not yield any information (could be omitted)*

10. Can the tf-idf weight of a term in a document exceed 1?

    Solution:

    - *by construction of the formula (df is finite and smaller than n)*