Computer Engineering Department
Bilkent University

CS533: **Information Retrieval Systems**
Assignment No. 2
February 25, 2013
Due date: March 12, 2013; Tuesday, by class time (hardcopy is required)

Notes: Handwritten answers are not acceptable. The next assignment may overlap with this one.

1. Consider the following search results for two queries Q1 and Q2 (the documents are ranked in the given order, the relevant documents are shown in bold).

Q1: **D1**, **D2**, D3, **D4**, D5, **D6**, D7, D8, D9, D10.

Q2: **D1**, D2, **D3**, D4, D5, **D6**, D7, D8, **D9**, and D10.

For Q1 and Q2 the total number of relevant documents are, respectively, 5 and 8 (for example for Q1 one of the relevant documents is not retrieved).

a. Using the TREC interpolation rule, in a table give the precision value for the 11 standard recall levels 0.0, 0.1, 0.2, … 1.0. Please also draw the corresponding recall-precision graph as shown in the first figure of TREC-6 Appendix A (its link is available on the course web site).

Hint. "Interpolated" means that, for example, precision at recall 0.10 (i.e., after 10% of rel docs for a query have been retrieved) is taken to be MAXIMUM of precision at all recall points >= 0.10. Values are averaged over all queries (for each of the 11 recall levels). These values are used for Recall-Precision graphs.

Please do this for each query separately and obtain one table for both queries using the average of two values at each recall point.

b. Find R-Precision (TREC-6 Appendix A for definition) for Query1 and Query2.

c. Find MAP for these queries.

d. Calculate precision and recall values @10 using the concepts of TP, FP, TN, FN: true positive, false positive, true negative, and false negative.

2. Consider the following document by term binary D matrix for m= 6 documents (rows), n= 6 terms (columns).

$$\begin{bmatrix} 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix}$$

Consider the problem of constructing a document by document similarity, S, matrix. How many similarity coefficients will be calculated using the following methods? For each case explain your answer briefly: give exact numbers for each document and briefly explain how you came up with those numbers.

a. Straightforward approach (using document vectors) -the 1st method discussed in the class-.

b. Using term inverted indexes.

c. Calculate the similarity values of all documents using the cosine and Dice coefficients. Calculate the match (correlation) between these two measures by using the Kendall's tau measure. Please briefly show your calculations.

3. In this part consider the paper J. Zobel, A. Moffat, "Inverted files for text search engines." *ACM Computing Surveys*, Vol. 38, No. 2, 2006.

a. Understand the skipping concept as applied to the inverted index construction.

Assume that we have the following posting list for term a: <1, 2> <3, 2> <9, 5> <10, 3> <12, 4> <17, 4> <18, 3>, <22, 2> <24, 4> <33, 4> <38, 5> <43, 4> <55, 3><64, 2> <68, 4> <72, 3> <75, 1> <88, 2>.. The posting list indicates that term-a appears in d1 twice and in d3 once, etc.

Assume that we have the following posting list for term-b: <12, 2> <45, 2> <66, 1>.

Consider the following conjunctive Boolean query: term-a **and** term-b. If no skipping is used how many comparisons do you have to find the intersection of these two lists?

Introduce a skip structure, draw the corresponding figure then give the number of comparisons involved to process the same query.

State the advantages and disadvantages of large and small skips in the posting lists. Note that in the paper it is assumed that compression will be used. The skip idea is applicable in an uncompressed environment too.

b. Give a posting list of of term-a (above it is given in standard sorted by document number order) in the following forms: 1), a) ordered by $f_{d,t}$, b) ordered by frequency information in prefix form. What are the advantages of the approaches a and b? Do they have any practical value?

4. What are the components of an information retrieval test collection? Explain the pooling approach? Please read the paper by Zobel (How Reliable Are the Results of Large-Scale Information Retrieval Experiments?) and give some reflections of his criticism of this approach.

5. Please read the article entitled "Just the facts," by Marina Krakovsky *Communications of the ACM*, Vol. 56 No. 1, Pages 25-27, January 2013.

In that article there are two sides about news aggregation on the web. Please briefly summarize the arguments of each view. In your opinion which side of the aggregation debate is right and why? Please discuss briefly.

6. Please read the article entitled "Technology strategy and management: The Apple-Samsung lawsuits," by Michael A. Cusumano, *Communications of the ACM*, Vol. 56 No. 1, Pages 28-31, January 2013.

After reading this article please state your view on the impact of patents on innovations and social welfare.

7. Please define the words lemma, root, and stem and give examples in English.

8. What is tf.idf? How would you use it in a binary D matrix environment. Please explain by using teh D matrix given in the second question. In this question as a resouce please consider the Zobel-Moffat (2006) cited above.

9. What are the possible versions of caching in information retreival (again consider teh Zobel-Moffat paper). In your opinion which one is the most influential one in terms of efficiency, and in terms of effectiveness?