Computer Engineering Department
Bilkent University

CS533: **Information Retrieval Systems**
Assignment No. 1
March 1, 2008
Due date: March 13, 2006; Thursday, by noon time (12:00 O'clock) (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 is, respectively, 3 and 4 (for Q2 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.

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. Find the TREC eval package on the Web. Please specify the number of measures listed in the package. Give the definitions of bpref (binary preference) and one more effectiveness measures (other than P@... measures), give their definitions, and calculate them for the given queries. If necessary make some reasonable assumptions and state your assumptions.

Note that bpref is defined in experimental lab environments to measure the performance of an IR approach. In such environments in order to identify the relevant documents the pooling approach is used and documents that are not evaluated by human assessors are assumed to be irrelevant. The bpref measure ignores such unseen/unevaluated documents. In bpref assume that D2 and D4 have not been evaluated before, but they appear in the listing.

Calculate the bpref values one more time, but assume that all of the listed documents have been evaluated by human assessors.

For the definition of bpref you may refer Buckley, C, Voorhees, E. M. Retrieval evaluation with incomplete information. ACM SIGIR 2004 Conference Proceedings, pp. 25-32. You may find easier to understand definitions for it.

2. Consider the following D matrix. In D, rows and columns, respectively, indicate document and term vectors.

$$D=\begin{array}{cccccccc} t_1 & t_2 & t_3 & t_4 & t_5 & t_6 & t_7 & \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & d_1 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & d_2 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & d_3 \\ 0 & 1 & 0 & 1 & 1 & 0 & 0 & d_4 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & d_5 \\ 0 & 1 & 0 & 0 & 0 & 0 & 1 & d_6 \\ 0 & 0 & 1 & 0 & 0 & 1 & 1 & d_7 \end{array}$$

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 explain how did you come up with those numbers.

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

b. Using term inverted indexes.

c. How can we apply the inverted index-based similarity calculation using the Jaccard similarity coefficient?

3. Obtain the similarity matrix S for the above D matrix (you don't need to show your intermediate steps). Use the Jaccard coefficient. Use the S matrix to construct the dendrogram structure corresponding to the single-link and complete link clustering methodologies.

4. Read the article “Another look at automatic text-retrieval systems” by G. Salton, *Communications of the ACM*. Vol. 29, No. 7, 1986, pp. 648-656.

Consider the following formula

$$w_{ij} = tf_{ij} \times idf_i$$

Explain the meanings of its individual components. Can we incorporate the idf component to indexing on the fly (i.e., during query processing time)? Explain your answer.

5. 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, 1> <8, 3> <10, 2> <12, 3> <17, 4> <18, 3>, <22, 2> <24, 2> <33, 4> <38, 5> <43, 5> <55, 3>. The posting list indicates that term-a appears in d1, two times and in d3, only once, etc.

Assume that we have the following posting list for term-b: <25, 2> <57, 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?

6. In this part consider the paper A. K. Jain, M. N. Murty, P. J. Flynn, “Data Clustering: A Review” *ACM Computing Surveys*, Vol. 31, No. 3, September 1999.

a. What is the difference between clustering (unsupervised classification) and discriminant analysis (supervised classification)? What kind of classification would you have in Information Filtering and why?

b. What are the components of a clustering task? Explain each step within the framework of an information retrieval environment.

c. What is the purpose of “cluster tendency” analysis?

d. What is k_Means clustering algorithm? How does it work?

e. Are the single-link and complete-link clustering methods we studied in class “agglomerative” or “divisive?” Explain.

f. What are the applications areas of clustering? List all of them. Explain two areas other than IR with one or two paragraphs. (That is explain how clustering is being used in these areas.)

7. Please read the article Voorhees, E. M. TREC: Continuing information retrieval's tradition of experimentation. Commun. ACM Vol. 50, No. 11, November 2007, 51-54.

a. Give a definition of the Cranfield methodology with your own words.

b. Briefly explain the purpose of the TREC conference.

c. Voorhees claims that SMART results can be considered representative of the IR filed. What is her justification for that? Do you agree or disagree? Please explain.