**Computer Engineering Department**
**Bilkent University**

CS533: **Information Retrieval Systems**
Assignment No. 1
February 26, 2009
Due date: March 12, 2009; 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, 4 and 5 (Q2 two of the relevant documents are 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.

**2**. Define a binary D matrix of your own choice (different from the ones we used in our class discussions etc.). Your D matrix must have a dimension of 5 by 7 (i.e., 5 documents and 7 terms).

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. Can you define a method which is more efficient than the method that uses the term indexes? If so please describe and explain why it is more efficient.

**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> <9, 2> <10, 3> <12, 4> <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 twice and in d3 once, etc.

Assume that we have the following posting list for term-b: <28, 2> <56, 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.** How can we use clustering in data mining applications? Please explain briefly (max three paragraphs).

**7.** Please read the article Voorhees, E. M. TREC: Continuing information retrieval's tradition of experimentation. *Com. Of the 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.** Define the major tasks of TDT (Topic Definition and Tracking) conference (e.g., first story detection, …). Note that this question is not related to the Voorhees article cited above.

**8.** **a.** Prove that the single-link clustering method is order-independent.

**b.** Prove that the complete-link clustering method is order dependent. For this purpose use a similarity matrix of your choice (try to keep the dimensions of the matrix as small as possible,

**9.** In this part consider the paper X. Qi, B. D. Davison, "Wep Page Classification: Features and Algorithms" *ACM Computing Surveys*, Vol. 41, No. 2, Article 12, February 2009.

**a.** What are the types of page classifications?

**b.** State the applications of Web classifications.

**c.** According to the authors what is the difference between text classification and Web classification?