Computer Engineering Department
Bilkent University

CS533: **Information Retrieval Systems**
Assignment No. 1
October 10, 2017
Due date: October 23, 2017; Monday, class time (hardcopy is required) You have to answer at least ceiling of half of the questions. All is more appreciated.

Note: Handwritten answers are not acceptable.

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**.

For Q1 and Q2 the total number of relevant documents is, respectively, 5 and 3 documents.

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).

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. What is the intuition behind interpolation?

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

d. Find MAP for these queries.

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

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

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 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.

3. Obtain the similarity matrix S for the above D matrix (you don't need to show your intermediate steps). Use the Dice similarity coefficient.

Use the S matrix to construct the dendrogram (cluster tree) structure corresponding to the single-link and complete-link clustering methodologies.

Explain how to use the dendrogram structure to obtain a partitioning clustering structures.

4. Give an example for the complete-link algorithm to prove that it can generate different clustering structures for a given S matrix.

4. Consider the above D matrix. Cluster the documents using the cover coefficient-based clustering methodology (C^3M). Please a) Show the double-stage probability experiment tree for the second document, and show the calculation of c_{24} of the corresponding C matrix, b) obtain the C matrix (you do not need to show the intermediate steps), c) find the number of clusters implied by the C matrix – explain how-, d) find the cluster seeds, e) obtain the IISD (inverted index for seed documents), f) obtain the clusters and explain how you them.

5. Anomaly or outlier detection is a data mining problem. See the related 2009 paper in the *ACM Computing Surveys* with the title Anomaly Detection: A Survey. Skim the paper to understand the problem. Read the section on cluster-based anomaly detection. How can we use a clustering algorith to detect outliers within the context of a document collection. Define a method based on teh clustering algorithms we studied in the classroom. Assume that you have a static document collection.

6. 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, 5> <10, 3> <12, 4> <17, 4> <18, 3>, <22, 2> <24, 2> <33, 4> <38, 5> <43, 5> <55, 3><64, 2> <68, 4> <72, 5> <75, 5> <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: <1, 2> <12, 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?

7. 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. Please explain the stages of clustering as defined in this paper.

b. Consider fuzzy clustering and introduce and idea that we can use fuzzy clustering approach in connection with C^3M.

c. In connection with simulated annealing the authors mention "tabu search." What does it mean? Explain its use within the context of simulated annealing-based clustering.

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

e. In connection with the above question (section d) please also explain what is meant by clustering tendency? Does it make sense to use clustering tendency in some stage(s) of clustering? What would you propose to use for identifying clustering tendency? Please try to be creative. For this purpose you may do a literature search and borrow some ideas and use them after some modification.

8. Is the complete-link clustering method order-independent? Explain/prove your claim. (You may see a related formal proof for the single-link method on our course web site).

9. 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.

10. What is query result caching? What are the advantages and disadvantages of query result caching?