Computer Engineering Department
Bilkent University

CS533: **Information Retrieval Systems**
Assignment No. 2
October 20, 2016
Due date: November 3, 2016; Thursday, by class time (hardcopy is required)

Notes: Handwritten answers are not acceptable. I expect that you will solve all of the problems and submit a complete solution; however, if you solve half of the questions it is acceptable.

1. Consider the following search results for the query 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**, D10

The total number of relevant documents is 5 both for Q1 and Q2.

a. Find R-Precision (TREC-6 Appendix A for definition) for Q1 and Q2.

b. Find MAP for these queries.

c. 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 F measure express it as a function of TP, TN, FP, and FN. Provide the derivation.

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

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

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. Obtain the S matrix by using the Dice coefficients.

4. 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. 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 d9 5 times, etc.

Assume that we have the following posting list for
term-b: <12, 7> <22, 7>.<66, 3>.<45, 3> <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 such data we have a pointer to next 5th entry (it will also have the lowest document number of the following data chunk) , for example for term-a at the beginning of the inverted index we have a pointer that indicates that the next document number at the beginning of the next data chunk as 17. Similarly at the beginning of 17 there will be a skip pointer that indicates that next data chunk starts with document number 38 and we will also have a pointer to that data chunk of course. Draw the corresponding figure then give the number of comparisons involved to process the same query using this skipping structure.

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

b. Can we take advantage of the skipping structure for disjunctive queries? Please explain.

c. Give a posting list 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? Please again give these pros and cons in a tabular form. Do they have any practical value?

5. 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.

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

b. 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 with some modification.

7. Apply the complete link algorithm to the D matrix given in question 3 using the Dice coefficient. Draw the corresponding dendrogram.

8. Is it possible to obtain a partitioning clustering structure with two cluster from the dendrogram that you obtain in the above question? Please explain how? If not possible explain why?

9. In C^3M prove that the number of clusters implied by C and C' are the same.

10. Obtain the clusters according to C3M. Please show your work.