Computer Engineering Department
Bilkent University

CS533: **Information Retrieval Systems**
Assignment No. 1
October 12, 2018
Due date: October 25, 2018; Thursday, by class time (hardcopy is required)

Notes: Handwritten answers are not acceptable. Use latex or word. You must solve at least half of the questions, always take ceiling (no. of questions / 2) in all assignments. I expect to see several submissions with complete solutions. If you see an incomplete explanation or a problem in wording introduce your solution, explain, and proceed.

I may overlap the next assignment with this one with a later due date.

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 for Q1 and Q2 are, respectively, 5 and 4.

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, P@10 and R@10, using the concepts of TP, FP, TN, FN: true positive, false positive, true negative, and false negative.

2. For the queries given above draw the recall precision graph using the TREC interpolated approach (See TREC 6 Appendix A). Explain the purpose of interpolation. Find related articles that may explain it and provide it/their citations.

3. Precision at 10 (P@10) vs. R-Precision which measure would you prefer to measure the effectiveness of a system? Please explain briefly.

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

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

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.

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. Assume that we have the following posting list for
term-a: <1, 2> <3, 2> <9, 2> <10, 3> <12, 4> <18, 4> <20, 3>, <23, 3> <25, 4> <33, 4> <40, 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 d10 three times, etc.

Assume that we have the following posting list for
term-b: <12, 7> <66, 3> <75, 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?

b. 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 18. Similarly at the beginning of 18 there will be a skip pointer that indicates that next data chunk starts with document number 40 and we will also have a pointer to that data chunk. Draw the corresponding figure then give the number of comparisons involved to process the same query using this skipping structure.

c. 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.

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

6. 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?

7. 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.

8. Consider interactive IR environments.

a. How can we evaluate an information retrieval system? Do some literature research and explain with citations.

b. What would be your own suggestion explain.

9. 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.

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, other than the ones that we may discuss in the classroom? Please try to be creative. For this purpose you may do a literature search and borrow some ideas and use them with some modification.

10. Give a small similarity matrix of your choice with 5 documents and show that the complete link algorithm may produce different, non unique, clustering structures. Draw the corresponding dendrogram.

Propose a method that would prevent this non uniqueness problem.

11. For a weighted D matrix according to the cover coefficient concept prove that diagonal entries of the C matrix can be smaller than the off diagonal entries, c_{ii} can be smaller than c_{ij} where i ǂ j. Proof by example is acceptable, but I prefer a general proof.

12. Consider the D matrix given above.

a. Calculate the c_{12} entry of the C matrix using the double stage experiment and by drawing the tree like structure for this entry of C.

b. Obtain the clusters according to C3M for the D matrix given above.

How many entries of the C matrix do you need to calculate?

Draw the IISD, inverted index for seed documents.

Please show your work for clustering.

13. Consider the double stage probability experiment of the cover coefficient experiment. The D matrix size is given as m (no. of rows) by n (no. of columns). Consider the construction of the C, document by document, matrix. Explain your answer with a simple figure.

a. What is the possible maximum number of active branches (branches that exist, i.e. with non-zero values)?

b. What is the minimum number of active branches?

14. How can we use the indexing clustering relationships implied by the cover coefficient concept: What can be its practical uses?