Computer Engineering Department
Bilkent University

CS533: **Information Retrieval Systems**
Assignment No. 1
October 17, 2019
Due date: November 1, 2019; ~~Tuesday~~ Friday, 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. Check the literature there are two different ways of calculating MAP either way is fine but clearly explain which one you use. Give a reference for your approach.

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

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

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: <2, 2> <3, 2> <9, 2> <10, 4> <12, 4> <19, 4> <20, 3>, <23, 3> <25, 4> <33, 4> <42, 5> <43, 4> <55, 3><64, 2> <68, 4> <72, 3> <75, 1> <88, 2>.
The posting list indicates that term-a appears in d2 twice and in d10 four times, etc.

Assume that we have the following posting list for
term-b: <10, 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 19. Similarly at the beginning of 19 there will be a skip pointer that indicates that next data chunk starts with document number 42 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. Can we take advantage of the skipping structure for "I want to have this term but not that term" type queries? (Example: clustering but not classification.) Please explain.

5. 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. If we want to rank all of the document of a collection do any of these two approaches provide an advantage? Please explain.

6. Consider the SIGIR 1998 paper by Zobel (How Reliable Are the Results of Large-Scale Information Retrieval Experiments?). In the paper he defines run depth (RD) and pool depth (PD). Define these two terms. According to Zobel which of the following is meaningful in measuring the system effectiveness in what way? a) RD = PD, b) RD > PD, and RD < PD. If he provides no discussion on that possibility provide your own judgment.

7. Consider interactive IR environments.

a. How can we evaluate an interactive information retrieval system? Do some literature research and explain with citations. Consider the use of Mechanical Turk (MT) approach for this purpose. Search the web and try to find some studies based on MT and explain the main concerns and benefits of this approach.

b. What would be your own suggestion? Briefly explain.

8. 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? How can you incorporate clustering tendency measurement to the indexing process? Please try to be creative. For this purpose you may do a literature search and borrow some ideas and use them with some modification.

9. Can we use single-link clustering method in a dynamic document environment? Try to define an algorithm and give an intuitive justification for your approach.

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 hierarchical clustering structure..

Propose a method that would prevent this non uniqueness problem.

11. According to the cover coefficient concept the number of clusters implied by documents and terms are identical ($n_c = n_c'$). Prove this equality..

12. Consider the D matrix given above.

a. Calculate the c_{32} 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. How can we incorporate the indexing clustering relationship implied by the cover coefficient concept to the indexing process? Try to find a creative approach and explain its use.

14. Consider the paper "Anomaly detection: A surveys" paper by Chandola et al. *ACM Computing Surveys*, 2009.

a. Briefly explain the use of anomaly detection in practical applications (itemize),

b. Develop and algorithm based on single-link clustering method for anomaly detection. How can you increase the recall level of your algorithm? How can you increase precision of your algorithm?

c. Develop and algorithm based on C3M for anomaly detection. How can you increase the recall level of your algorithm? How can you increase precision of your algorithm?