Computer Engineering Department
Bilkent University

CS533: **Information Retrieval Systems**
Assignment No. 1
February 25, 2015
Due date: March 11, 2015; Wednesday, 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 (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. Find R-Precision (see TREC-6 Appendix A for definition) for Query1 and Query2.

c. Find MAP for these queries.

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

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

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 the term inverted indexes (the 3rd and most efficient method we discussed in the class).

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. Look at the agreement of the similarity matrix implied by these two clustering algorithm with the original similarity matrix obtained from the given D matrix. For agreement calculation use the product moment correlation coefficient (see the appendix below). Please show your steps, but do not exaggerate in terms of details.

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 fifth document, and show the calculation of c_{52} 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 find them.

5. Prove that the number of clusters implied by the C and C' matrices are the same.

6. Obtain the Rand similarity for the clustering structures CS1= { {a, c}, {b, d, e}, {f, g}} and another clustering structure CS2= {{a}, {b, c, d}, {e, f, g}} -where the last cluster of CS2 contains the members e, f, and g-. Optional: you may also obtain the corrected Rand coefficient using these two clustering structure. Show the contingency table that needs to be constructed for the Rand coefficients.

7. 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> <8, 2> <10, 3> <12, 4> <17, 4> <17, 4>, <22, 3> <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: <10, 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?

8. 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 are the typical application areas for clustering? For one of them other than IR explain its use.

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 after some modification.

9. Define an algorithm for maintenance of the single-link structure. Please define clearly how you add new comers to the existing clustering structure obtained. Please consider both hierarchical version and the flat version when we obtain a partitioning structure by cutting the dendrogram at a certain similarity threshold. Use your imagination when needed.

10. 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.

11. Define cluster hypothesis.
Find three real life example that follows the intuition of cluster hypothesis other than the library example. Library example: You look for *Kara Kitap* and go to library find the book and then you find other Orhan Pamuk books next to it.

Find three real life example that would not follow cluster hypothesis. In other words you would expect to observe cluster hypothesis like behavior but it does not follow your expectation.

APPENDIX

A. The definitions of c_{ij} and c_{ij}' are as follows.

$$c_{ij} = \alpha_i . \sum_{k=1}^{n} \left(d_{ik} . \beta_k . d_{jk} \right)$$

$$c_{ij}' = \beta_i . \sum_{k=1}^{m} \left(d_{ki} . \alpha_k . d_{kj} \right)$$

B. The product moment correlation between X and Y is defined as follows.

$$r = r(X,Y) = \frac{\text{cov}(X,Y)}{\left[\text{var}(X).\text{var}(Y)\right]^{\frac{1}{2}}} = \frac{\sum (x_i - x_{avg})(y_i - y_{avg})}{\left[\left[\sum (x_i - x_{avg})^2\right]\left[\sum (y_i - y_{avg})^2\right]\right]^{\frac{1}{2}}}$$