Computer Engineering Department
Bilkent University

CS533: **Information Retrieval Systems**
Assignment No. 3
March 16, 2014
Due date: March 24, 2014; Monday, by class time (hardcopy is required)

Notes: Handwritten answers are not acceptable.

1. Consider the following D matrix.

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

Obtain the corresponding single-link clustering structure (dendrogram). Give the clustering structure approach if the dendrogram is cut at the similarity level 0.45 (note that you will obtain a partitioning structure). For similarity calculation use the Dice coefficient.

2. Consider the D matrix of question no. 1. Obtain the corresponding complete-link clustering structure (dendrogram). Give the clustering structure approach if the dendrogram is cut at the similarity level 0.45. For similarity calculation use the Dice coefficient.

3. Consider the D matrix given in question 1.

a. Construct the corresponding C matrix (can be obtained either by matrix multiplication or the related formula), you may just give the C matrix.
b. Calculate the number of clusters.
c. Find the seed power of all documents.
d. Determine the cluster seeds. Explain your reasoning.
e. Construct IISD (Inverted Index for Seed Documents).
f. Use the IISD data structure to cluster d_5. Show your computations.
g. Construct the clusters.
h. In an efficient implementation of the C^3M how many entries of the C matrix do we have to calculate? Answer this question (1) in general using the symbols such as m, n, n_c, etc.; and (2) for the D matrix of this question.

4. According to the cover coefficient concept for a D matrix of size m by n prove that if

a. All documents are unique then all c_{ii} values are equal to 1, ($1 \leq i \leq m$)
b. All documents are identical then all c_{ii} values are equal to 1/m ($1 \leq i \leq m$)

5. Questions based on the clustering-indexing relationships implied by the cover-coefficient concept.

a. For the D matrix of question number 4, calculate the number of clusters by using the clustering-indexing relationships implied by the cover-coefficient concept (m x n) / t= n_c. Are the two results, the ones obtained by the precise cover coefficient formula and the approximation formula given above, close to each other? Under what kind of conditions we would expect to have values not matching each other? (The related paper may have some hints about this. Be creative and justify your answers.

b. Create a binary D matrix that one can easily see the number of clusters and the number of clusters implied by the (m x n) / t= n_c are inconsistent. Is this mismatch a problem in practical environments?

6. How can we use the concepts of C^3M for cluster maintenance?
Hint: Refer to Can, F. Incremental Clustering for Dynamic Information Processing, *ACM Trans. on Information Systems.* Vol. 11, No. 2 (April 1993), pp. 143-164. A short paragraph is enough.

How can we extend the single-link clustering algorithm for cluster maintenance (addition of new documents).

7. What is the purpose of "cluster tendency" analysis? Please refer to the Data clustering: A review, A. Jain, Murty, Flynn, *ACM Computing Surveys* paper.

Can we use the clustering-indexing relationships revealed by the cover coefficient concept for this purpose? If so how, please explain.

In a group of documents if we do not have enough clustering tendency (lower than our expectation) how can we change the data?

8. Consider the following specifications for a document database:

m (No. of documents)	= 150
n_c (No. of clusters)	= 10
k (No. of relevant documents for a given query)	= 5

Assume that (a) documents are randomly distributed among the clusters; (b) each cluster has the same size. What is the expected number of clusters to be accessed to retrieve all relevant documents of the query? (Use Yao's formula, see the related paper: Yao, S. B., "Approximating block accesses in database organizations." *Communication of the ACM*, Vol. 20, No. 4, 1977, pp. 260-261.

9. Obtain the similarity matrix implied by the dendrogram of question number 1. Calculate the "product moment correlation coefficient" (see Appendix B below) between the corresponding elements of the implied similarity matrix and the original similarity matrix obtained by using the given D matrix.

10. Consider the Rand index and cluster purity concepts used for cluster validation. Can we use the ground truth data constructed for Rand index for the calculation of cluster purity? If your answer is no explain why, if your answer is yes give a simple example for calculating the values of Rand index and cluster purity values.

11. Gather a group of about fifteen words in Turkish and obtain the stems for two of them using the successor variety method. Please make sure that you have a reasonable corpus which is not too difficult to use or too easy to use while applying the method.

12. Compare the successor variety and the n-gram-based stemming methods. Which one is more robust - if any- and explain why. Please also explain what is meant by robustness within the context of this application domain.

13. Consider the binary string 01101110101101001110111110001
Construct the PAT tree for the first 12 sistrings. Show some of the intermediate steps so that one would be able to follow your presentation. Please also give the PAT array.

14. In a bit pattern based PAT tree environment how can we answer a query which requires to have at the most a certain number of bits between two bit patterns?

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.

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