Computer Engineering Department
Bilkent University

CS533: **Information Retrieval Systems**
Assignment No. 2
November 7, 2018
Due date: November 20, 2018; Class Time

Notes: Handwritten answers are not acceptable. Solve all problems, answer all questions. Please remember ACM digital library can be accessed from a Bilkent IP.

1. Consider a partitioning clustering structure with four clusters of sizes 2, 4, 8, and 16: the first cluster contains two documents, the second contains 4 documents etc. For a query with 3 relevant documents what are the minimum, maximum, and expected number of target clusters when documents are randomly distributed among the clusters. A target cluster is a cluster that contains one or more relevant documents for a query.

See: Yao, S. B. (1977) Approximating block accesses in database organizations, *Com. of the ACM*, 20(4): 260-261.

2. Dynamic clustering is important in dynamically changing environments such as data streams. In data stream environments the data items, documents, have temporal importance they come stay in the system for some time then disappear and new ones come. However, some of them can be persistent and stay in the system. As in the news or stock information.

The purpose of dynamic or incremental clustering is updating the clustering structure due to addition and deletion of objects (in our case documents) in an efficient way without doing reclustering, i.e., using the existing structure as much as possible.

a. How can we use the concepts of the C^3M in data stream environments?

See: Can, F. (1993). "Incremental clustering for dynamic information processing". *ACM Trans on Information Systems*. 11 (2): 143–164.

b. How can we modify the single link clustering approach for data streams? Is it possible or not? Please explain.

3. Cluster labeling, i.e., assigning natural language descriptions for the contents of clusters is an important summarization type problem.

a. Is this problem similar to document snippet generation by search engines for their search results?

b. Find two papers from literature on cluster label generation. Explain each one separately with about two sentences.

c. Suggest a method of yours for cluster labeling briefly. Give a step by step explanation. Explain your intuition and explain why would you expect that it would work.

d. Can we do user-oriented cluster labeling? Under what condition(s) such a thing would make sense?

See: You decide.

4. In this question consider the supervised and unsupervised ways of grouping data items.

a. Which problem you think may give the users of machine learning (ML) a false happiness? I mean ML users falsely expect that the algorithm will work fine.

b. Do we have problems similar to overfitting and underfitting in clustering? Please explain. If you use a resource for your explanation please cite that work(s).

See: Pablos Domingos (2012) in his paper titled "A few useful things to know about machine learning" *Com. of the ACM* . 55(10): 78-87. This paper may or may not have enough information to answer this question.

5. Consider the following symmetric similarity matrix for a document collection with four documents. The similarity between d1 and d2 is 0.67, etc.

$$S = \begin{bmatrix} 1.00 & 0.67 & 0.50 & 0.20 \\ - & 1.00 & 0.80 & 0.10 \\ - & - & 1.00 & 0.00 \\ - & - & - & 1.00 \end{bmatrix}$$

Consider the following respective similarities of these documents to a given query:

(d_1, 0.70), (d_2, 0.40), (d_3, 0.60), (d_4, 0.80)

Use the MMR algorithm for selecting the best matching first two documents. Consider the following λ values. After each case give the diversity among the selected documents; where diversity is defined as (1-average similarity among selected documents).

a. Use λ = 1.00.

b. Use λ = 0.00.

c. Use λ = 0.50.

d. Use the MMR algorithm for selecting the best matching first three documents with Use λ = 0.50.

e. A question not related to above numeric MMR questions:
How can we use an approach like MMR for a task other than summarization and ranking. In your answer you may replace relevance and diversity with some other concepts.

See: Carbonell, J. G., Goldstein, J. (1998) The use of MMR, diversity-based reranking for reordering documents and producing summaries. SIGIR Conf., 335-336

6. Consider the following search engines A, B, C, and D and ranking provided by them for the documents a, b, c, d, e, and f.

A= {b, a, c, d}
B= {b, d, a, f}
C= {b, d, c, a}
D= {a, c, d, e}

Rank the documents according to the following data fusion methods.

a. Reciprocal rank,

b. Borda count,

c. Condorcet.

See: Nuray, N., Can, F. (2006) Automatic ranking of information retrieval systems using data fusion. *Information Processing and Management.* 42(3): 595-614.