Computer Engineering Department
Bilkent University

CS533: **Information Retrieval Systems**
Assignment No. 4
November 21, 2016
Due date/time: December 5, 2016; Monday, class time

Notes: Answer the questions in the order given here. You have to answer at least half of the questions.

1. Consider the incremental version of C^3M: C^2ICM, Cover Coefficient-based Incremental Clustering Methodology, described in Can F, Incremental clustering for dynamic information processing, ACM TOIS, 1993).

a. Briefly explain the algorithm (one paragraph).

b. In the paper there is the concept of clustering similarity, explain its purpose within the context of C^2ICM.

c. How can we use C^2ICM in a data stream environment like news article such as news articles?

Can we use it in an online fashion one document at a time or should we use it in terms of data batches?

What should we do for old documents? What is the definition of old in the data stream applications?

A possible good reference for data stream clustering and classification: H-L Nguyen, et al. A survey on data stream clustering and classification, Knowledge and Information Systems, December 2015, Volume 45, Issue 3, pp 535–569

2. Consider a dynamically changing document environments as in data stream environments objects are coming and leaving, i.e. we have addition of new objects and deletion of old objects and we want to cluster the objects that appear in the latest time window (like news articles of most recent n days). How can we change the following algorithms so that we can use them for this purpose. (I am not expecting a perfect answer, and probably there is none. Just be realistically creative.

a. Single-link.
b. K- means.

3. Find Rand similarity of the clustering structures CS1= { {a, d}, {b, c, e}, {f, g}} and CS2= {{a, b}, {c, g}, {d, e, f}} -where the last cluster of CS2 contains the members e, f, and g-.

4. For the clusters of question 2 assume that CS1 is the ground truth, under this assumption calculate recall, precision and F measure values.

If we change the roles and assume that CS2 is the ground truth and obtain recall, precision and F do we obtain the same values? Please briefly explain.

5. Consider a partitioning clustering structure that contains the following clusters. C1= {x, x, x, y} C2= {y, y, y, x), C3= {z, z, z, z, x, y}. This presentation means that in C1 there are three items of type x and one item of type y and we have similar interpretations for the contents of the other clusters. Calculate the cluster purity value for the above clustering structure.

6. Assume that we have 100 pages and each page contains 20 records. We have a query with 4 relevant records. What is the minimum, maximum, and expected number of pages to be accessed to retrieve all of the relevant records? When appropriate use Yao's formula.

7. There is a paper by Cardenas which is cited in S. B. Yao's Comm. of the ACM 1977 article. If we answer question 5 under the light of the Cardenas article what will be your answer?

8. Suppose that we have the information retrieval system A, B, C, D and they provide the following rankings for the documents a, b, c...

A= (a, d, b, c)
B= (a, b, e, d)
C= (c, a, e, f)
D= (b, e, g, f)

Combine the results of these search engines and rank all documents by using the reciprocal rank, Borda count, a nd Condorcet methods.

The reference for this question is R. Nuray, F. Can:Automatic ranking of information retrieval systems using data fusion. Inf. Process. Manage. 42(3): 595-614 (2006)