Computer Engineering Department
Bilkent University

CS533: **Information Retrieval Systems**
Assignment No. 4 - Final Form
March 21, 2016; Last Update: March 27, 11:33 pm
Due date/time: April 1, 2016; Friday, by noon time - 11:59 am.

Notes: Handwritten answers are acceptable, a word processor output will be appreciated. Answer the questions in the order given here. You have to solve at least five questions. After class time on April 1 Friday the submissions will be accepted by 10 points penalty by 5:00 pm. After that they will be accepted on April 4 Monday (hardcopy) by 5:00 pm, with a penalty of 20 points. On Tuesday April 5 Tuesday (hardcopy) by class time with 30 points penalty. No acceptance after that point.

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 teh data stream applications?

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

3. 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.

4. Consider a partitioning clustering structure that contains the following clusters. C1= {x, x, x, x, y} C2= {y, y, y, x), C3= {z, z, z, z, x, y}. This presentation means that in C1 there are four 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.

5. Assume that we have 80 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.

6. Consider the following 3 by 3 document by term binary D matrix for m= 3 documents (rows), n= 3 terms (columns).

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

a. Calculate TDVs for all terms using the approximate space density concept. Use the Dice coefficient for similarity calculation.

b. Calculate TDVs for all terms using the approximate cover coefficient concept.

c. When you rank the terms according to their TDV from lowest to highest do the result agree in terms of rank positions of the terms. For this purpose use Kendall's Tau. For obtaining the result you may use a web page. If you do so give its citation.

7. Salton and his co workers define a way of using TDVs for increasing recall and precision in IR. Define their methods. Please see their *CACM* 1975 paper "A vector space model for automatic indexing."

8. 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= {a, b, c, d}
B= {b, a, c, d}
C= {c, d, a, b}
D= {d, c, a, b}

Sort the documents according to the following data fusion methods.

a. Reciprocal rank,

b. Borda count,

c. Condorcet.

d. In your opinion which method is more fair in this case (if any)? Please explain why.

9. Consider the paper entitled "Putting successor variety stemming" by Stein and Potthast. What is the contribution of the paper beyond the idea of successor variety? Explain briefly.

10. Consider the paper Zobel and Moffat survey paper "Inverted files for search engines". Consider the methods "accumulator limiting" and "accumulator thresholding".

a. Briefly explain them.

b. Which method is more effective in memory management, i.e. under what conditions?