Computer Engineering Department
Bilkent University

CS533: **Information Retrieval Systems**
Assignment No. 5
January 2, 2016
Due date/time: January 5, 2016; Thursday, Final exam time

Notes: Answer the questions in the order given here. You have to answer at least half of the questions. A word processor generated submission is preferred, handwritten submissions are acceptable. Please bring you hard copy submission to the exam room. If you see something missing in a question please make reasonable assumptions and explain.

1. For the following D matrix calculate the TDV by using the cover coefficient concepts (use the approximate method).

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

2. Salton and his co-workers define a way of using TDVs for increasing recall and precision in IR. Define their methods. Do you agree or disagree please explain? See `A vector space of automatic indexing" by Salton et al., *Comm. of the ACM*, 1975.

3. Consider the following signatures.
S1: 1000 1001
S2: 1100 0010
S3: 0011 1100
S4: 0000 1111
S5: 1011 0100
S6: 0100 1010
S7: 1100 0101
S8: 1000 1110
S9: 1010 0111

a. Use fixed prefix method to partition the above signatures. Let key length k equal to 3.

b. Consider the following queries.
Q1: 1101 0001
Q2: 0110 0011
Q3: 1100 1100
Use the partitions of section-a to calculate the time needed (turnaround time) to process the queries in sequential and parallel environments. (Use the assumptions that we used in the class room, e.g., the processing of one page signature requires 1 time unit, etc.). What is the speed up ratio for the parallel environment (defined as ratio (parallel processing time for all queries / sequential processing time for all queries)?

4. Consider the signatures of question 3.

a. Use LHSS (Linear Hashing with Superimposed Signatures) method to partition the signatures. Take block size as 3 and LF to be maintained as 2/3. Show intermediate steps as you insert the signatures.

b. For the following query which pages do we need to access?
Q: 1101 0001

5. Consider the signatures of question 3.

a. Use extendible Hashing Hashing method to partition the signatures. Take block size as 2. Show intermediate steps as you insert the signatures.

b. For the following query which pages do we need to access?
Q: 1001 1001

6. Consider the following information filtering profiles used in a Boolean environment.

P1= a, b, c, d
P2= a, f
P3= b, c, f
P4= b, d
P5= a, c, f

Assume that when the terms are sorted in frequency order according to their number of occurrences in documents term a is the least frequently used term in the documents and is also the most frequently used term in the user profiles. The sorted term list continues as b, c … f.

a. Consider the ranked key method explained in the paper by Yan and Garcia-Molina (Index structures for selective dissemination of information under the Boolean model) and draw the directory and the posting lists for the ranked key method.

b. What is the intuition behind the ranked key method: how does it improve the filtering efficiency?

c. Do you agree with the following statement: As time passes automatic update of user profiles may provide higher user satisfaction. Please explain your answer. If your answer is yes suggest an algorithm for this purpose.