Computer Engineering Department
Bilkent University

CS533: **Information Retrieval Systems**
Assignment No. 5 - Solve any two questions
May 1, 2016
Due date: Final Day - Beginning of the Final Exam, Bring a hardcopy to the exam, No late submission will be accepted

See Spring 2015 HW3 solutions as examples on the same/similar subjects.

Notes: Handwritten answers are acceptable. If your handwriting is not tidy I will assume that you did not submit anything. If you are not sure use a word processor.

1. Consider a document collection containing 400,000 objects. The signature of an object requires 512 bits. What are the signature file sizes using the following signature file organization methods?
a. Sequential Signatures (SS),
b. Bit-sliced Signatures (BS).

2. In the database environment of question 1 consider a query with 5 bit positions equal to one. These bit positions are 1, 2, 4, 5. (The leftmost position of a signature is bit position 1.) For filtering (i.e., for query signature - document signatures matching) how many pages need to be accessed in the case of SS and BS? (Page size is 0.5 K bytes.) Note that in SS we place signatures one after the other and in the case of BS we place bit slices one after the other: Place the first bit slice and then right after that place the second bit slice and if there is room in the page allocated to slice 1 use the remaining space for the second bit slice and carry on like this.

3. Consider the following signatures.
S1: 0100 0110
S2: 1110 0011
S3: 1100 0011
S4: 0000 1111
S5: 1010 0110
S6: 1011 0100
S7: 1100 1010
S8: 0101 0101
S9: 1011 0100
S10: 1001 1010

a. Use the fixed prefix method to partition the above signatures. Take k (key length) as 2. Show the file structure (contents of the pages etc.).

b. Now consider the following queries.
Q1: 1110 0001
Q2: 0110 0011
Q3: 1100 1100
Q4: 0011 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? Please also calculate the average turnaround time both for sequential and parallel processing environments.

4. Partition the signatures of question 3 using the following partitioning methods. (To answer this question consider the paper Lee and Leng 1989 paper in ACM TOIS, “Partitioned Signature Files: Design Issues and Performance Evaluation,” or “Signature Files: An Integrated Access Method for

Formatted and Unformatted Databases” by Aktug & Can.
a. EPP (take z= 2).
b. FKP (take k= 2).
c. To process the queries of question no. 3 which pages need to be accessed and why?

5. Partition the signatures of question 3 this time by using the extendible hashing algorithm (using prefixes). Assume that each data block can contain two signatures.

For Q1 and Q2 of question 3 please specify which pages need to be access and please explain briefly.

6. Partition the signatures of question 3 this time by using the linear hashing algorithm (using suffixes). Assume that each data block can contain three signatures. (Bkfr= 3) and LF= 2/3 as in our in class example.

For Q1 and Q2 of question 3 please specify which pages need to be access and please explain briefly.

For these queries indicate which data pages need to be accessed.

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

P1= a, b, c, d, e, f
P2= a, b, e, 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.

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, *ACM TODS*) and draw the directory and the posting lists for the ranked key method and the tree method.

8. Consider the following document collection containing four documents (rows) defined by four terms (columns).

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

A query is submitted and its vector is defined as follows: Q= [1 0 0 0]

Assume that we want to use the MMR algorithm for selecting the best matching first two documents. After each case what is the cohesiveness (similarity) and diversity among the selected documents and how can we measure it? Does the MMR algorithm provide what it promises. For each case please show your steps explicitly. For similarity calculations use the Dice coefficient.

a. Use λ = 1.00 and indicate the selected documents.

b. Use λ = 0.00 and indicate the selected documents. What is the diversity among the selected documents and how can we measure it? Do we have more diversity with respect to the first case above? Please explain.

c. Use $\lambda = 0.50$ and indicate the selected documents. What is the diversity among the selected documents and how can we measure it? Do we have more diversity with respect to the first case above? Please explain.

9. The search result for a query in ranked order are given in the following table. Different meanings of documents d_1, d_2 ... d_{10} are shown by m_1, m_2 , ... m_6.

Rank	Document	Subtopic
1	d_1	m_4
2	d_2	m_3
3	d_3	m_1. m_2
4	d_4	m_6
5	d_5	m_5, m_6
6	d_6	m_5
7	d_7	m_4
8	d_8	m_3
9	d_9	m_1
10	d_{10}	m_1

a. Find s-recall at rank position 3, 6, and 10.

b. Find precision IA at rank position 3, 6, and 10.