Computer Engineering Department
Bilkent University

CS533: **Information Retrieval Systems**
Assignment No. 5
December 23, 2017
Due date: December 28, 2017; Thursday 5 pm sharp

Notes: Solve all questions. Handwritten answers are acceptable, make sure that you solve the questions in the order they are given. If your handwriting is not tidy I will assume that you did not submit anything; if you are not sure use a word processor. Leave your hardcopy submission in my mailbox. You may also submit it as a pdf document by email by 5 pm on December 28. The subject of your email must be "CS533 HW5" followed by your name I will not accept emails that come after 5 pm..

1. Consider the following symmetric similarity matrix for a document collection with four documents. The similarity between d1 and d2 is 0.67, etc. The lower portion of the matrix is not shown.

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

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

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

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).

Does the MMR algorithm provide what it promises? For each case please show your steps concisely. (Ref. Carbonell Goldstein SIGIR 1998 paper.)

a. Use λ = 1.00.

b. Use λ = 0.00.

c. Use λ = 0.50.

d. Please repeat the steps a to c three documents.

2. 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_3
2	d_2	m_4
3	d_3	m_1, m_2
4	d_4	m_5, m_6
5	d_5	m_6
6	d_6	m_5
7	d_7	m_4
8	d_8	m_3
9	d_9	m_2
10	d_{10}	m_1

a. Find s-recall at rank position 5 and 10.

b. Find precision IA at rank position 5 and 10.

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

Rank the documents according to the following data fusion methods. (Ref. Nuray Can IPM 2006 paper.)

a. Reciprocal rank,

b. Borda count,

c. Condorcet.

4. Consider a document collection containing 512,000 objects. The signature of an object requires 1024 bits. What are the signature file sizes using the following signature file organization methods?

a. Sequential Signatures (SS),

b. Bit-sliced Signatures (BS).

5. In the environment of the above question consider a query with 5 bit positions equal to one. These bit positions are 1, 2, 3, and 4. The leftmost most significant 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 given as 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.

6. Consider the following signatures.
S1: 0110 1010
S2: 0100 0110
S3: 1110 0011
S4: 1100 0011
S5: 0011 1010
S6: 1010 0101
S7: 1011 0010
S8: 0000 1111
S9: 1010 0110
S10: 1011 0100

a. Use the fixed prefix, FP, 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

i. Give the partition activation ratio and signature activation ratio for the queries in the FP environment.

ii. Use the partitions of section-a to calculate the average turnaround time to process the queries in sequential and parallel FP environments. Use the assumptions that we used in the class, e.g., the processing of one page signature requires 1 time unit, etc. What is the speed up ratio for the parallel environment?

7. Partition the signatures of question 6 using the following partitioning methods.

a. EPP (take z= 2).

b. FKP (take k= 2).

c. To process the queries of the above question which pages need to be accessed and why?

Ref. 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 on the web (Can Aktug explains it with a simple figure.

8. Partition the signatures of question 6 with 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.

Ref. Zezula et al ACM TOIS 1991 Dynamic partitioning of signature files.

9. 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, f
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.