Computer Engineering Department
Bilkent University

CS533: **Information Retrieval Systems**
Assignment No. 4
December 16, 2019 (December 23, 2019 Question 10 is added)
Due date/time: December 31, 2019; Tuesday, Class time

Notes: A word processor generated submission is required, handwritten submissions are unacceptable. If you see something missing in a question please make reasonable assumptions and explain.

1. For the following document by term D matrix calculate the TDV by using the cover coefficient concepts (you may 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 model for automatic indexing" by Salton et al., *Comm. of the ACM*, 18(11): 613-620, Nov. 1975.

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

a. Use fixed prefix method to partition the above signatures. Let key length k equal to 3. D.L. Lee, and C.-W. Leng,, "Partitioned signature files: Design issues and performance evaluation," *ACM Trans Information Systems,* 7(2): 158-180, Apr. 1989.

b. Consider the following queries.
Q1: 0001 1101
Q2: 0011 0110
Q3: 1100 1100

i. 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 classroom, e.g., the processing of one page signature requires 1 time unit, etc.).

ii. What is the speed up ratio for the parallel environment (defined as ratio (sequential processing time for all queries / parallel processing time for all queries)?

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. The second one is available on the web.

a. EPP (take z= 2).

b. FKP (take k= 2).

c. To process the following queries which pages need to be accessed? Answer separately for EPP and FKP.

Q1: 1110 0001
Q2: 0110 0011
Q3: 1100 1100
Q4: 0011 1100

d. What is the intuition behind EPP and FKP methods? Explain briefly.

5. Consider the signatures of question 3.

a. Use Extendible Hashing method to partition the signatures. Take block size as 2. Show intermediate steps as you insert the *signatures.* Fagin, R.; Nievergelt, J.; Pippenger, N.; Strong, H. R. (September 1979), "Extendible Hashing - A Fast Access Method for Dynamic Files", *ACM Transactions on Database Systems*, 4 (3): 315–344.

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

Q: 1101 1001

6. 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 -desired load factor level- as 2/3. Show intermediate steps as you insert the signatures. P. Zezula , F. Rabitti , P. Tiberio (October, 1991) Dynamic partitioning of signature files, *ACM Transactions on Information Systems*, 9(4): 336-367.

b. For the following query which pages do we need to access?
Q: 1101 1001

7. Study the BitFunnel paper and explain the generation and use of bit block signature for searching. Bob Goodwin, et al., BitFunnel: Revisiting Signatures for Search, Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2017.

8. Consider the following information filtering profiles used in a Boolean environment.
P1= a, b, c
P2= a, d, f
P3= b, c, f
P4= c, d
P5= a, c, d

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 T. W. Yan and H. Garcia-Molina (June 1994) Index structures for selective dissemination of information under the Boolean model, *ACM Transactions on Database Systems*, 19(2): 332-364. 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. Suggest an algorithm to automatically update user profiles in information filtering for better better performance? Why would you expect that you algorithm would improve performance.

9. You are familiar with data stream processing due to our previous assignment on data stream classification. Consider the paper J. Lu et al. (December 2019) "Learning under concept drift: A review" *IEEE Transactions on Knowledge and Data Engineering*, 31(12): 2346-2360.

a. Define virtual concept drift and real concept drift. Explain their differences.

b. Give two real-life examples for each.

c. In your opinion which one is more common in real life, explain why.

10. Consider the following symmetric similarity matrix for a document collection with four documents. The similarity between d1 and d2 is 0.67, etc.

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

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

(d_1, 0.70), (d_2, 0.30), (d_3, 0.50), (d_4, 0.75)

Use the MMR algorithm for selecting the best matching first three documents. Consider the following λ values and after each case give the diversity among the selected documents: defined as (1-average similarity among selected documents).

a. Use $\lambda = 1.00$.

b. Use $\lambda = 0.00$.

c. Use $\lambda = 0.50$.

See: Carbonell, J. G., Goldstein, J. (1998) The use of MMR, diversity-based reranking for reordering documents and producing summaries. SIGIR Conf., 335-336