Department of Computer Science, CUSAT

M.Tech. Semester II End-Semester Examinations, April 2023

ALGORITHMS FOR MASSIVE DATASETS

Time: 3 hours marks: 50

1. (a) Given two binary vectors V1 = [0 0 1 1 1 0 0] and V2 = [0101100]. Compute (i) the Cosine distance and (ii) the Jaccard distance between the vectors.

   (b) The minimum hash values returned by a set of documents are compared to suggest a possible duplication in Minhashing. Explain the rationale behind the method.

OR

2. (a) Given are two text documents D1 and D2. From a set of similarity measurement methods (Jaccard, Cosine) which method would you recommend for detecting any potential plagiarism? Justify your views.

   (b) Given a map with 4 documents and presence/absence of 6 tokens in the documents.

token	S1	S2	S3	S4
0	0	1	0	1
1	0	1	0	0
2	1	0	0	1
3	0	0	1	0
4	0	0	1	1
5	1	0	0	0

   Compute the minhash signature for each column if we use the following three hash functions: h1(x) = 2x + 1 mod 6; h2(x) = 3x + 2 mod 6; h3(x) = 5x + 2 mod 6.

3. Explain the steps followed in processing a query (q) to retrieve all near neighbours from the data points using locality sensitive hashing (LSH) method.

OR

4. There are several distance measures used to compute the distance between data points. Conceptually, similarity is defined as (1-distance) in many practical problem solving situations. When does a similarity admit an LSH?

5. Consider the *new user signing up* process in an online application. A bloom filter is used in the application to determine whether a username that is opted by the user was already taken. Will the application ensure that it never tells you that username doesn't exist when it actually exists?

OR

6. (a) Describe how a data structure called *bloom filter* shall be used for allowing only the whitelisted URLs from an incoming stream of URLs.

   (b) Compute the approximate number of bits per entry (e.g. unique URL) and the number of hash functions needed for a false positive rate of 1%.

7. What are the common characteristics of *Probabilistic data structures and algorithms* (PDSA)? Describe the linear probabilistic counter and the associated algorithm used for counting of distinct elements in a massive dataset.

OR

8. Describe the function of the Loglog counter used for cardinality estimation upon massive datasets. What is the underlying motivation to use Harmonic mean for the estimation in the HyperLoglog counters as against the arithmetic mean used in the Loglog counter?

9. Following 12 data points are represented through a pair of feature values

   (2,2), (5,2), (3,4), (9,3), (12,3), (11,4), (10,5), (12,6), (4,8), (6,8), (4,10), (7,10)

   Identify 3 clusters using the BFR Algorithm.

OR

10. Reservoir sampling is a process for streaming through a list of n items and randomly sampling k of them, while using only O(k) memory. Briefly explain how important is the selection of the number k in the algorithm as per the randomization process.

Department of Computer Science, CUSAT

M.Tech. Semester II Second Series Examinations, Apr 2023

21-479-0201: Algorithms for Massive Datasets

Time: 2 hours marks: 20

1. Counting the unique users (IP addresses) who visited the web pages across all the products (about 10 million in the store) in a large popular online e-commerce application on a given day is a challenging problem. Describe the challenges. [5]

2. What is the advantage of using buckets (or *bucketized* sketch) in the Log-Log counting algorithm? [5]

3. In the BFR algorithm for identifying clusters in dataset, the space efficiency is enhanced when more data points are discarded during the execution. Give a method to determine the 'closeness' of a given data point to existing clusters. [5]

4. Discuss the first pass in CURE algorithm for cluster analysis. What is the rationale behind the creation of synthetic representative points using uniform compression technique? [5]

-------/\------

Department of Computer Science, CUSAT

M.Tech. Semester II First Series Examinations, Feb 2023

21-479-0201: Algorithms for Massive Datasets

Time: 2 hours

marks: 20

1. (a) Define the property of pair-wise independence between a set of hash functions upon a collection of random variables. [3]

   (b) How can we generate many instances of universal hash functions for a given universal hash function family of the form H(x) = (**a**x+b) mod p? [2]

2. Let there are two equal sized documents D1 and D2 that are 100% disjoint. Now, contents of the documents are swapped in such a way that half of the number of lines of text in D1 is replicated into D2 and *vice versa,* simultaneously. What is the likely value of Jaccard similarity if we consider MinHash with k hash functions for calculating similarity between D1 and D2 after the swapping? [5]

3. (a) Define the (r,R,p,P)-sensitive hash function family *F* corresponding to a similarity function *S*(). [3]

   (b) In one LSH implementation, in order to find the pair-wise similarity between **m** numbers of documents, it is planned to randomly choose **t** number of bands from the signature matrix and hash them to equal sized hash tables that has **L** buckets.

   (i) What is the optimal size of a hash table? [1]

   (ii) How many hashing operations are to be performed to set the buckets across all hash tables? [1]

4. (a) Let **n** keys are stored in a Bloom filter of size **m** that uses **k** number of hash functions. Arrive at an expression for the *false positive rate* **f**. [2]

   (b) Let the criteria for designing a Bloom filter is that it should store 1000 keys and should filter within a false positive rate of 0.1% (*i.e.* 1 out of 1000 queries would result in a 'may be' answer). If the number of hash functions to be used has been fixed at 4, what should be the minimum size of the Bloom filter B (*i.e.* size of the vector B in number of bits)? [3]

------/\------

2019

Department of Computer Science, CUSAT

IInd Semester M.Tech. Full-time 2nd Series Examinations, March 2019

CSC3201: ALGORITHMS FOR MASSIVE DATASETS

Time: 2 hours

marks: 20

1. (a) Describe one method to determine the k-centroids after reading in the first chunk of data to the memory in BFR algorithm for clustering of data. [2]

   (b) The variance-covariance matrix for a multi-dimensional dataset with 3 dimensions, viz. d1, d2 and d3 is given below.

$$\begin{bmatrix} 0.6 & 0.62 & 0.46 \\ 0.62 & 0.5 & 0.58 \\ 0.46 & 0.58 & 0.66 \end{bmatrix}$$

   Comment on the axis-aligned nature of the given dataset. [2]

2. Discuss the steps performed during the first pass through the dataset in CURE algorithm for clustering of data. [3]

3. Explain why the candidate itemset generation in Apriori algorithm to discover frequent large itemsets is found as expensive. [2]

4. The table below shows a set of transactions on items. Construct the FP-tree corresponding to the set of transactions for a minimum support of 3.

Transactions	List of items
T1	F, A, C, D, G, I, M, P
T2	A, B, C, F, L, M, O
T3	B, F, H, J, O, W
T4	B, C, K, S, P
T5	A, F, C, E, L, P, M, N

[2]

5. (a) Counting distinct elements in a data stream is a challenging problem. Explain the reason. [2]

   (b) Compare Probabilistic counting algorithm to the Loglog counting algorithm in terms of the memory space usage efficiency. [2]

6. (a) Explain how the reservoir sampling algorithm is able to efficiently make a random sample of k elements that are evenly distributed. [3]

   (b) A bloom filter implementation is restricted to the usage of 1MB of main memory space. The estimated maximum size of the data stream is 2,00,000 elements. What is the minimum number of hash functions to be used so that the false positive rate is minimized to 2%? [2]

------^------