LNMIIT/B. Tech./CSE/PE/2018-19/EVEN/CSE3152/MT

LNMIIT
The LNM Institute of Information Technology

The LNM Institute of Information Technology

Department: Computer Science and Engineering

Mining of Massive Datasets (CSE3152)

Exam Type: Mid Term

Time: 90 minutes **Date: 28th Feb 2019** **Max. Marks:** 50

Instruction: *There are five questions and each question carries equal mark. Answer each question on a separate page.*

Q.1. We want to use a single MapReduce pass to perform matrix multiplication P=MN. Write the Map and Reduce functions for this.

Q.2. Write an algorithm in MapReduce framework to compute Minhash signatures assuming the input matrix is stored in chunks that corresponds to some columns. That is, each Map task gets some of the columns and all the hash functions as input. Assume there are 100 hash functions i.e. the number of rows of the signature matrix is 100.

Q.3. Suppose there are a large number of strings with varying length (which cannot fit in main memory) and the goal is to identify all pairs of strings whose Jaccard similarity is at least 0.9. Explain the "prefix indexing" method of finding all such pairs of strings. Also provide example(s) supporting your explanation.

Q.4. Suppose a stream consists of integers 3, 4, 5, 9, 4, 2, 9, 3, 8, 4. Find out the number of distinct elements in the stream based on Flajolet-Martin approach using the hash function $h(x) = 3x + 1 \bmod 32$.

Q.5. Out of the two options given below, which one can differentiate documents better and why? Notice that both options require 4 bytes to represent a shingle.

Option 1: extract 9-shingles and hash them down to four bytes

Option 2: extract 4-shingles