BE|IT|SEM-VIII|CBSGS

20 MAY 2019

(Time: 3 Hrs) Marks: 80

N.B. : 1. Question no. 1 is **compulsory**.

2. Solve any **Three** questions out of remaining **Five** questions.

Q.1. (a) Explain Blooms filter for stream data mining. (5)

(b) Find the jaccard distance and cosine distance between the following pairs of set: X=(0,1,2,4,5,3) and Y=(5,6,7,9,10,8). (5)

(c) Explain the steps of the HITS algorithm. (5)

(d) Explain "Shuffle & Sort" phase and "Reducer Phase" in Map Reduce. (5)

Q.2. (a) Write a Map reduce pseudo code to multiply two matrices. Illustrate with an example showing all the steps. (10)

(b) Explain Hadoop Ecosystem with core components. Explain its physical architecture. State the limitations of Hadoop. (10)

Q.3. (a) Suppose a data stream consists of the integers 1,3,2,1,2,3,4,3,1,2,3,1. Let the Hash function being used is h(x) = (6x+1) mod 5; estimate the number of distinct in this stream using Flajolet - Martin algorithm. (10)

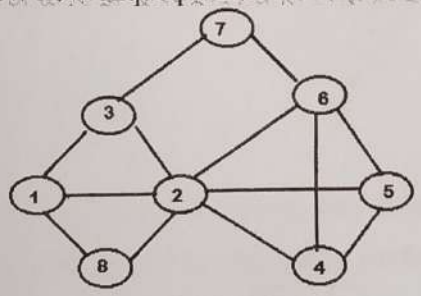
(b) Distinguish the following: (10)

a) PCY, Multistage

b) Document data store and Column family data store

Q.4. (a) Give two applications for counting the number of 1's in a long stream of binary values. Using a stream of binary digits, Illustrate how DGIM will find the number of 1's. (10)

(b) For the given graph show how clique percolation method will find cliques. (10)



Q.5. (a) Consider the web graph given below with six pages (A, B, C, D, E, F) with directed links as follows. (10)

A→ B,C

B→A,D,E,F

C→AF

Assume that the PageRank values for any page m at iteration 0 is PR(m)=1 and teleportation factor for iterations is β=0.85. Perform the page rank algorithm and determine the rank for every page at iteration 2.

(b) Explain clearly how the SON partition based algorithm helps to perform frequent item set mining for large data sets. How does this algorithm avoid false negatives? (10)

Q.6. (a) Explain collaborative filtering system. How is it different from content based system? (10)

(b) Clearly explain how CURE algorithm can be used to cluster big data sets. (10)
