K. J. Somaiya Institute of Technology, Sion, Mumbai-22
(Autonomous College Affiliated to University of Mumbai)

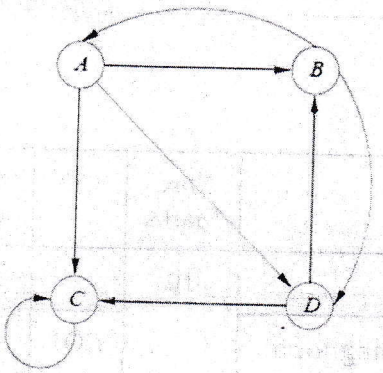
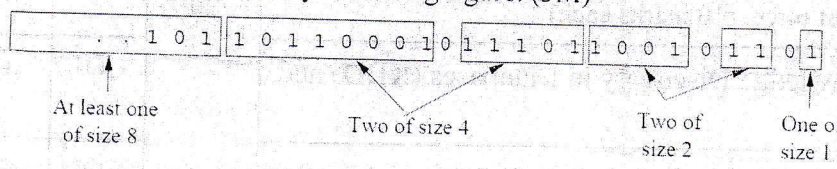
May-June 2025
Program: M. Tech. Scheme: II
Regular Examination: FY Semester: 2
Course Code: PCEC202 and Course Name: Big Data Analytics
Date of Exam: 04/06/2025 Duration: 02.5 Hours Max. Marks: 60

Instructions:
(1) All questions are compulsory.
(2) Draw neat diagrams wherever applicable.
(3) Assume suitable data, if necessary.

Q. No.	Question	Max. Marks	CO	BT level
Q 1	Solve any **two** questions out of three: (05 marks each)	10		
a)	Demonstrate what data preprocessing is required in data mining for a weather forecasting system.		CO1	AP
b)	What are the five Vs of Big Data? Briefly explain each.		CO2	U
c)	List and explain any two goals of Hadoop's architecture.		CO3	U
Q 2	Solve any **two** questions out of three: (05 marks each)	10		
a)	What are the different phases of a MapReduce job? Explain most powerful phase of MapReduce.		CO4	U
b)	Define re-identification and its impact on big data analytics.		CO5	U
c)	Distinguish between static, incremental, and streaming data. Give example.		CO6	AN
Q.3	Solve any **two** questions out of three. (10 marks each)	20		
a)	Explain the process of Knowledge Discovery in Databases (KDD) and its importance.		CO1	U
b)	For the graph given below show the page ranks of all the nodes after running the Page Rank algorithm for two iterations considering the teleportation factor $\beta = 0.8$		CO2	AP

K. J. Somaiya Institute of Technology, Sion, Mumbai-22
(Autonomous College Affiliated to University of Mumbai)

May-June 2025
Program: M. Tech. Scheme: II
Regular Examination: FY Semester: 2
Course Code: PCEC202 and Course Name: Big Data Analytics
Date of Exam: 04/06/2025 Duration: 02.5 Hours Max. Marks: 60

				
c)	How is data moved in and out of Hadoop? Explain the process with examples.		CO3	U
Q.4	Solve any **two** questions out of three. (10 marks each)	20		
a)	What id Hadoop Ecosystem? Discuss role of Pig and Hive in the Hadoop ecosystem and Compare them.		CO4	U
b)	What is the importance of privacy in data science? Discuss privacy-preserving data mining techniques with relevant examples.		CO5	U
c)	i) Using an example bit stream explain the working of the DGIM algorithm to count the number of 1's (Ones) in a data stream. (4M) ii) Estimate the number of 1's in the last k positions, for k = 15. How far off the correct value is your estimate? (3M) iii) Describe what happens to the buckets if three more 1's enter the window represented by following figure. (3M) 		CO6	AP

****