1·6·2023 (E)

SOMAIYA VIDYAVIHAR UNIVERSITY

Semester: January 2023 –May 2023 **Maximum Marks: 100** **Examination: ESE Examination** **Duration:3 Hrs.**		
Programme code: 5 **Programme: Honours- Data Science and Analytics**	**Class: T Y B.Tech**	**Semester:VI(SVU 2020)**
Name of the Constituent College: **K. J. Somaiya College of Engineering**	**Name of the department: Computer**	
Course Code: 116h54C601	**Name of the Course: Advanced Data Mining**	
Instructions: 1)Draw neat diagrams 2) All questions are compulsory **3) Assume suitable data wherever necessary**		

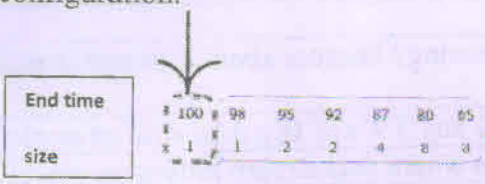
Que. No.	Question	Max. Marks
Q1	Solve any **Four**	**20**
i)	What is descriptive data mining? Discuss about different descriptive data mining tasks.	5
ii)	What is Big Data? Discuss any 3 v's of Big data with example	5
iii)	Explain any 3 applications where data stream mining is useful?	5
iv)	Write the HITS algorithm	5
v)	Give any application of time series data mining. Also explain any 2 time series data mining task	5
vi)	Discuss with an example method to find the purity of clusters	5

Que. No.	Question	Max. Marks
Q2 A	Solve the following	**10**
i)	Discuss the features of Hadoop framework for processing Big Data	5
ii)	Give example using k-means clustering a) Different initial centroid results in different clusters b) Sensitive to outliers	5
	OR	
Q2 A	Explain Distributed Data Mining in detail.	10
Q 2 B	Solve any **One**	**10**
i)	A database has five transactions. Let *min_sup_count* be 2 and *min_conf* be 70%. TID | Items_bought T100 | a,b,c T200 | b,c,d,e T300 | c,d T400 | a,b,d T500 | a,b,c Find all frequent patterns using Apriori algorithm . List any 3 valid association rules with support, confidence, lift	10
ii)	Illustrate Flajolet Martin algorithm with suitable example. Discuss an application of the algorithm.	10

Que. No.	Question	Max. Marks
Q3	Solve any **Two**	**20**
i)	Term frequency matrix for 5 articles (A1 to A5) is shown below. Using Tf-idf score, find the similarity between articles? Identify the two articles that are most similar.	10

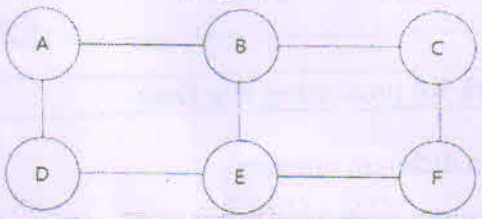
Article/Terms	Trump	JNU	AAP	Corona	Divestiture
A1	14	1	0	6	3
A2	0	21	5	0	0
A3	0	15	18	0	5
A4	5	2	0	12	0
A5	0	0	5	0	10

ii) Employ the DGIM algorithm. Shown below is a data stream with N=40 and current bucket configuration. — 10



Suppose that at times 101 through 105, 1's appear in the stream. Compute the set of buckets that would exist in the system at time 105. Also compute the number of 1's in latest k=12 bits of the window.

iii) Apply Girvan-Newman algorithm and calculate the betweenness centrality for edges on social graph given below — 10



Que. No.	Question	Max. Marks
Q4	Solve any **Two**	**20**
i)	Discuss 4 components of Apache Hadoop framework	10
ii)	What does apriori property state? How is the use of apriori property in apriori algorithm of finding frequent patterns?	10
iii)	Consider the following distance matrix given below. Apply agglomerative clustering using single link and complete link approach to find hierarchy of clustering. Clearly show the steps of construction of dendogram	10

Item	A	B	C	D	E
A	0	1	2	2	3
B	1	0	2	4	3
C	2	2	0	1	5
D	2	4	1	0	3
E	3	3	5	3	0

Que. No.	Question	Max. Marks
Q5	(Write notes / Short question type) on any **four**	**20**
i)	Discuss 5 steps of KDD process	5
ii)	Write short notes naïve bayes text classification approach	5
iii)	Consider the Web graph with three nodes 1, 2, 3. The links are as follows: 1->2, 2->1, 2->3. Compute the page rank with β=0.5	5
iv)	Write the algorithm for sequence pattern discovery.	5
v)	Find the Jaccard coefficient (Jc) for the query q and docs d1 and d2 below. Query: top university (set q) Doc 1: university of California (set d1) Doc 2: best university in USA (set d2) What are the limitations of Jaccard score ?	5
vi)	Write a short notes on architecture of Distributed data mining system	5