WALCHAND COLLEGE OF ENGINEERING

(Government Aided Autonomous Institute)

Vishrambag, Sangli - 416415

Final Year B. TECH. (Information Technology)

MSE , ODD SEMESTER, AY 2022-23

Data Mining (5IT401)

MSE

PRN: _____

Day, Date and Time: Monday, 10/10/2022, 03.00 pm to 04.30 pm

Max Marks: 30

IMP: Verify that you have received question paper with correct course, code, branch etc.

Instructions:
a) All questions are compulsory.
b) Writing question number on answer book is compulsory otherwise answers may not be assessed.
c) Assume suitable data wherever necessary.
d) Figures to the right of question text indicate full marks.
e) Mobile phones and programmable calculators are strictly prohibited.
f) Except PRN anything else writing on question paper is not allowed.
g) Exchange/Sharing of stationery, calculator etc. not allowed.

Text on the right of marks indicates course outcomes (only for faculty use).

		Marks	
Q1 A)	State the methods to fill in the missing values for attributes in data mining process.	3	CO1
Q1 B)	State the possible ways of integrating a data mining system with data warehouse.	3	CO2
Q1 C)	State use of normalization of data. Perform min-max normalization of following data in the range of 0 to 1.	4	CO2

Name	Years of Experience
A	8
B	20
C	10
D	15

Q2 A)	State the basic principles of Attribute-Oriented Induction.	3	CO1
Q2 B)	What is concept hierarchy? State the major types of concept hierarchy with example of each type.	3	CO2
Q2 C)	Find the 5 number summary for the given data set and draw box-plot for it Data : 23, 42, 12, 10, 15, 14, 9.	4	CO3
Q3 A)	Give the classification of association rule mining based on different criteria.	3	CO1
Q3 B)	Draw flowchart for ARCS (Association Rules Clustering System)? Give the limitations of ARCS.	3	CO2
Q3 C)	Find maximal frequent itemset from following transaction data. Assume minimum support of 50%.	4	CO2

TID	A	B	C	D	E	F
T_1	1	0	1	1	0	0
T_2	0	1	0	1	0	0
T_3	1	1	1	0	1	0
T_4	0	1	0	1	0	1

D) When we prefer to use hierarchical clustering? Perform single linkage hierarchical clustering on following data. *CO2* **6**

Distance	A	B	C	D	E
A	0	-	-	-	-
B	0.71	0	-	-	-
C	5.66	4.95	0	-	-
D	3.61	2.92	2.24	0	-
E	4.24	3.54	1.41	1.00	0
F	3.20	2.50	2.50	0.50	1.12

Q6 A) What is Hotspot analysis in spatial data mining? What are its applications? **4** CO2

B) Short note on - Similarity Search in Multimedia Data. **4** CO2

C) Elaborate First Order Markov Model. **5** CO1

End of question paper

WALCHAND COLLEGE OF ENGINEERING

(Government Aided Autonomous Institute)

Vishrambag, Sangli – 416415

Final Year B.Tech. (Information Technology)

ESE , ODD SEMESTER, AY 2022-23

Data Mining (5IT401)

ESE

PRN: _____

Day & Date: Tuesday, 13/12/2022 Time : 3.00 pm to 5.00 pm

Max Marks: **50**

IMP: Verify that you have received question papers with correct course code, branch etc.

Instructions
a) All questions are compulsory.
b) Writing question number on answer book is compulsory otherwise answers may not be assessed.
c) Assume suitable data wherever necessary.
d) Figures to the right of question text indicate full marks.
e) Mobile phones, smart gadgets and programmable calculators are strictly prohibited.
f) Except PRN anything else writing on question paper is not allowed.
g) Exchange/Sharing of stationery, calculator etc. not allowed.

Text on the right of marks indicates course outcomes (Only for faculty use) Marks

Q1 A) State and explain major Tasks in Data Preprocessing. 3 CO1

Q2 A) What is Attribute Relevance Analysis? How it is performed? 3 CO1

Q3 A) Draw FP-tree for following transaction data (Table 1). 4 CO2

Table 1

Transaction	List of items
T1	I1,I2,I3
T2	I2,I3,I4
T3	I4,I5
T4	I1,I2,I4
T5	I1,I2,I3,I5
T6	I1,I2,I3,I4

Q4 A) Which difficulties may arise when a decision tree is constructed? 4 CO2

B) Why Gini index is used? Consider the data given in Table 2, to find Gini Index of - {(Supercar, Heavy),Economy} 5 CO3

Table No-2

Car type	Class
Economy	High
Supercar	High
Supercar	High
Economy	Low
Heavy	Low
Economy	High

1B094BDD3D2FA4D47187E121A1627B6C

C) The crosstab of t-weight and d-weight is given below. Fill the missing data in it. Also write quantitative description rule for target class 'TV'. CO2

Class\Region	State 1			State 2			Both States		
	Count	t-weight (%)	d-weight (%)	Count	t-weight (%)	d-weight (%)	Count	t-weight (%)	d-weight (%)
TV	80	25	40	240		30	320	100	32
PC	120		60	560	82.35	70	680	100	
Both class	200	20	100		80	100	1000	100	100

3

Q4 A) Compare Apriori and FP-growth algorithms for mining frequent patterns in large datasets. 3 CO1

B) State methods to improve Apriori efficiency. 3 CO3

C) Consider given transaction data and - CO3

i. How many association rules generated using Brute-Force approach?
ii. Find support of Itemset : {Milk, Bread, Egg}
iii. Find confidence of following association rules

R1: Bread → {Milk, Egg}

R2: {Milk, Egg} → Bread

Trans_ID	Items purchased
1	Milk, Bread, Egg
2	Milk, Juice
3	Juice, Butter
4	Milk, Bread, Egg
5	Coffee, Egg
6	Coffee
7	Coffee, Juice
8	Milk, Bread, Cookies, Egg
9	Cookies, Butter
10	Milk, Bread

3

····· End of question paper ·····

WALCHAND COLLEGE OF ENGINEERING
(Government Aided Autonomous Institute)
Vishrambag, Sangli – 416415
Final Year B.Tech. Information Technology
MSE, ODD SEMESTER, AY 2023-24
Data Mining (5IT401)

MSE

PRN: _____

Day & Date: Thursday, 21/09/2023 Time : 3.00 pm to 4.30 pm Max Marks: **30**

IMP: Verify that you have received question papers with correct course code, branch etc.

Instructions
a) All questions are compulsory.
b) Writing question number on answer book is compulsory otherwise answers may not be assessed.
c) Assume suitable data wherever necessary.
d) Figures to the right of question text indicate full marks.
e) Mobile phones, smart gadgets and programmable calculators are strictly prohibited.
f) Except PRN anything else writing on question paper is not allowed.
g) Exchange/Sharing of stationery, calculator etc. not allowed.

Text on the right of marks indicates course outcomes (Only for faculty use)

			Marks	
Q1	A)	Answer in short (two/three) sentences.		CO2
		i. What is role of IT engineer/expert in Data mining field?	3	
		ii. How data mining helps in business?		
		iii. Give scope of data mining?		
Q2	B)	What do you mean by preprocessing of data in data mining? Give data preprocessing stages.	3	CO1
	C)	Write 3-4-5 rule? Why it is used?	3	CO3
	D)	Why is Normalization in Data Mining? Find normalized values for following data in the range [0, 1] using min-max normalization method.		CO2

Employee Name	Years of Experience
ABC	8
JKL	20
MNO	10
PQR	15

3

Q3	A)	Define concept hierarchy. State the major types with example of each.	3	CO1
	B)	State the possible ways of integrating a data mining system with data warehouse.	3	CO1

-OR-

Why Info Gain is used? A training data (Table 3) is given as follows, find the Info Gain for Outlook

Table 3

Sr. No.	Outlook	Temperature	Humidity	Windy	Class
1	overcast	hot	high	false	Play
2	overcast	mild	high	true	Play
3	overcast	hot	normal	false	Play
4	rain	mild	high	false	Play
5	rain	cool	normal	false	Play
6	rain	cool	normal	true	NoPlay
7	rain	mild	normal	false	Play
8	rain	mild	high	true	NoPlay
9	sunny	hot	high	false	NoPlay
10	sunny	hot	high	true	NoPlay
11	sunny	mild	high	false	NoPlay
12	sunny	cool	normal	false	Play
13	sunny	mild	normal	true	Play

C) Consider Table 3 as training data and Use Bayes Classifier to predict class of following condition. 5

{Outlook=Sunny, Temperature= cool, Humidity= high, Windy= false}

Q5 A) State the categories of constraints in constraints-based clustering. 4

B) Explain Grid-Based clustering Method. Give example. 5

C) Apply agglomerative hierarchical clustering algorithm to form clusters for following data (Table 4) using single linkage approach. Draw dendrogram for resulting clusters. 5

Table 4

Data/Distance	A	B	C	D	E	F
A	0	-	-	-	-	-
B	662	0	-	-	-	-
C	877	295	0	-	-	-
D	255	468	754	0	-	-
E	412	268	564	219	0	-
F	996	400	138	869	669	0

Q6 A) State and elaborate in short- Classification of Web Mining Techniques 4

B) Elaborate in short with example and diagram- Spatial trend analysis 4

C) How data mining can be employed on Digital Images? State any difficulties in Image mining. 4

End of question paper

1B094BDD3D2FA4D47187CF21A1627B6C

Q2 D) Define support and Confidence for association rule. 3
Consider following data and given two association rules "milk → juice" and "bread →juice"; find support and Confidence for each rule.

Tid	time	Items
101	6:35	milk, bread cookies, juice
102	7:38	milk, juice
103	8:05	milk, eggs
104	8:40	bread, cookies, coffee

Q3 A) What are the difficulties may arise when a decision tree is constructed? 3
Q3 B) State the advantages and disadvantages of Decision Tree classification. 3
Q3 C) A training data is given as follows. Identify the attribute for best splitting at first step of the data set by using info gain and entropy. 4

Outlook	Temperature	Humidity	Windy	Class
sunny	hot	high	false	NoPlay
sunny	hot	high	true	NoPlay
overcast	hot	high	false	Play
rain	mild	high	false	Play
rain	cool	normal	false	Play
rain	cool	normal	true	NoPlay
overcast	cool	normal	true	Play
sunny	mild	high	false	NoPlay
sunny	cool	normal	false	Play

Q3 D) From given data, identify the class of following test case by using Naïve Bayes Classifier 4
Test case - X: (Refund=No; Marital status =Married; Income=120K)

Marital status	Income (K)	Home loan refund	Defaulter
Single	125	Yes	No
Married	100	No	No
Single	70	No	No
Married	120	Yes	No
Divorced	95	No	Yes
Married	60	No	No
Divorced	220	Yes	No
Single	85	No	Yes
Married	75	No	No
Single	90	No	Yes

Q4 A) State the typical requirements of clustering in data mining. 3
Q4 B) State the categories of constraints in constraints-based clustering. 3
Q4 C) State and define in short; different methods for hierarchical clustering based on distance computations. 3

C) Consider following transaction data. CO3

Find (i) Support (Soap); (ii) Confidence (Perfume → Soap)

Transaction ID	Items
1	Jam, Perfume, Snacks, Soap
2	Egg, Bread, Jam, Snacks
3	Onion, Potato
4	Bread, Chocolate, Coffee, Jam
5	Egg, Coffee, Perfume, Snacks, Soap
6	Bread, Chocolate, Soap, Jam
7	Onion, Perfume
8	Bread, Chocolate, Snacks
9	Chocolate, Perfume, Soap
10	Egg, Bread, Jam, Snacks

3

D) Explain multidimensional association rules with example. 3 CO3

E) Which constrains are used in constraint based association rule mining? 3 CO2

Q4 A) State the criteria for comparing and evaluating classification and prediction methods. 4 CO1

B) What is Bayesian belief network (BBN)? Give an example. 4 CO1

C) A training data is given as follows. Identify the attribute for best splitting at first step of the data set by using info gain and entropy. CO3

Outlook	Temperature	Humidity	Windy	Class
sunny	hot	high	false	NoPlay
sunny	hot	high	true	NoPlay
overcast	hot	high	false	Play
rain	mild	high	false	Play
rain	cool	normal	false	Play
rain	cool	normal	true	NoPlay
overcast	cool	normal	true	Play
sunny	mild	high	false	NoPlay
sunny	cool	normal	false	Play
rain	mild	normal	false	Play
sunny	mild	normal	true	Play
overcast	mild	high	true	Play
overcast	hot	normal	false	Play
rain	mild	high	true	NoPlay

5

WALCHAND COLLEGE OF ENGINEERING

(Government Aided Autonomous Institute)

Vishrambag, Sangli - 416415

Final Year B. TECH. (Information Technology)

MSE , ODD SEMESTER, AY 2022-23

Data Mining (5IT401)

MSE

PRN _____

Day, Date and Time: Monday, 10/10/2022, 03.00 pm to 04.30 pm

Max Marks: 30

IMP: Verify that you have received question paper with correct course, code, branch etc.

Instructions: a) All questions are compulsory.
b) Writing question number on answer book is compulsory otherwise answers may not be assessed.
c) Assume suitable data wherever necessary.
d) Figures to the right of question text indicate full marks.
e) Mobile phones and programmable calculators are strictly prohibited.
f) Except PRN anything else writing on question paper is not allowed.
g) Exchange/Sharing of stationery, calculator etc. not allowed.

Text on the right of marks indicates course outcomes (only for faculty use)

		Marks	
Q1 A)	State the methods to fill in the missing values for attributes in data mining process.	3	CO1
Q1 B)	State the possible ways of integrating a data mining system with data warehouse.	3	CO2
Q1 C)	State use of normalization of data. Perform min-max normalization of following data in the range of 0 to 1.	4	CO2

Name	Years of Experience
A	8
B	20
C	10
D	15

Q2 A)	State the basic principles of Attribute-Oriented Induction.	3	CO1
Q2 B)	What is concept hierarchy? State the major types of concept hierarchy with example of each type.	3	CO2
Q2 C)	Find the 5 number summary for the given data set and draw box-plot for it Data : 23, 42, 12, 10, 15, 14, 9.	4	CO3
Q3 A)	Give the classification of association rule mining based on different criteria.	3	CO1
Q3 B)	Draw flowchart for ARCS (Association Rules Clustering System)? Give the limitations of ARCS.	3	CO2
Q3 C)	Find maximal frequent itemset from following transaction data. Assume minimum support of 50%.	4	CO2

TID	A	B	C	D	E	F
T_1	1	0	1	1	0	0
T_2	0	1	0	1	0	0
T_3	1	1	1	0	1	0
T_4	0	1	0	1	0	1

For following data; predict the class of given condition (X) using Bayes Classifier. CO3

X= {Outlook=Rain, Temperature= cool, Humidity= high, Windy= True}

Sr. No.	Outlook	Temperature	Humidity	Windy	Class
1	overcast	hot	high	false	Play
2	overcast	mild	high	true	Play
3	overcast	hot	normal	false	Play
4	rain	mild	high	false	Play
5	rain	cool	normal	false	Play
6	rain	cool	normal	true	NoPlay
7	rain	mild	normal	false	Play
8	rain	mild	high	true	NoPlay
9	sunny	hot	high	false	NoPlay
10	sunny	hot	high	true	NoPlay
11	sunny	mild	high	false	NoPlay
12	sunny	cool	normal	false	Play
13	sunny	mild	normal	true	Play

5

State major clustering approaches. 4 CO3

With reference to DBSCAN algorithm define following terms 4 CO2

minPts, eps, Core point, Border point

Explain Single Linkage and complete linkage method for hierarchical clustering. 4 CO1

Perform hierarchical clustering on following distance data. Draw resulting dendogram. CO3

	A	B	C	D	E	F
B	9.06					
C	5.83	5.66				
D	3.16	8.25	7.21			
E	5.39	12.53	7.28	14.42		
F	5.83	14.56	10.00	16.16	3.61	
G	3.61	6.71	2.24	8.60	5.83	8.06

5

Short Note on - Similarity Search in Time-Series Analysis. 4 CO2

Explain - Precision and recall for Text Retrieval. 4 CO2

Define – Trend, Cycle, Seasonal w.r.t time series data. 4 CO1

Differentiate between data mining and web mining. 4 CO3

End of question paper

WALCHAND COLLEGE OF ENGINEERING

(Government Aided Autonomous Institute)
Visharambag, Sangli – 416415

Final Year B.Tech. Information Technology
Re-Exam, Odd and Even Semester AY 2022-23
Data Mining (5IT401)

Re-Exam

PRN: _____

Day & Date: Monday, 10/07/2023 Time : 02.00 pm to 05.00 pm

Max Marks: **100**

IMP: Verify that you have received question papers with correct course code, branch etc.

Instructions
a) All questions are compulsory.
b) Writing question number on answer book is compulsory otherwise answers may not be assessed.
c) Assume suitable data wherever necessary.
d) Figures to the right of question text indicate full marks.
e) Mobile phones, smart gadgets and programmable calculators are strictly prohibited.
f) Except PRN anything else writing on question paper is not allowed.
g) Exchange/Sharing of stationery, calculator etc. not allowed.

			Marks	
		Text on the right of marks indicates course outcomes (Only for faculty use)		
Q1	**A)**	State and explain in short - the stages in data mining	4	CO1
	B)	What is Dimensionality Reduction? State basic Heuristic methods of attribute subset selection.	4	CO1
	C)	Apply min- max and Z-Score Normalization for following data. Data = {1000, 2000, 3000, 5000, 9000}	4	CO3
	D)	What you mean by binning? State different binning methods and smooth following data by these method Data- 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34	5	CO3
Q2	**A)**	Define and give example of- Schema Hierarchies and Set-grouping hierarchies.	6	CO1
	B)	When Analytical Characterization is performed? How it is performed?	4	CO2
	C)	Classify data mining Primitives with the help of real word examples.	4	CO1
	D)	What is Cube? Elaborate Roll-up and Drill-down operation.	4	CO3
Q3	**A)**	Write a short note on Brute-force approach for mining association rule.	4	CO1
	B)	Give the advantages of FP-tree for association rule mining.	4	CO3
	C)	Find out the type of correlation analysis for following data.		CO2

A	B
20	8
12	34
9	4

4

	D)	For the following given transaction data-set generate rules using Apriori algorithm.	5	CO2

0143143A582C1DBAADC2A5C568A5605D

WALCHAND COLLEGE OF ENGINEERING

(Government Aided Autonomous Institute)
Visharambag, Sangli – 416415

Final Year B.Tech. Information Technology
Supplementary ODD SEMESTER, AY 2022-23
Data Mining (5IT401)

Supplementary

PRN: _____

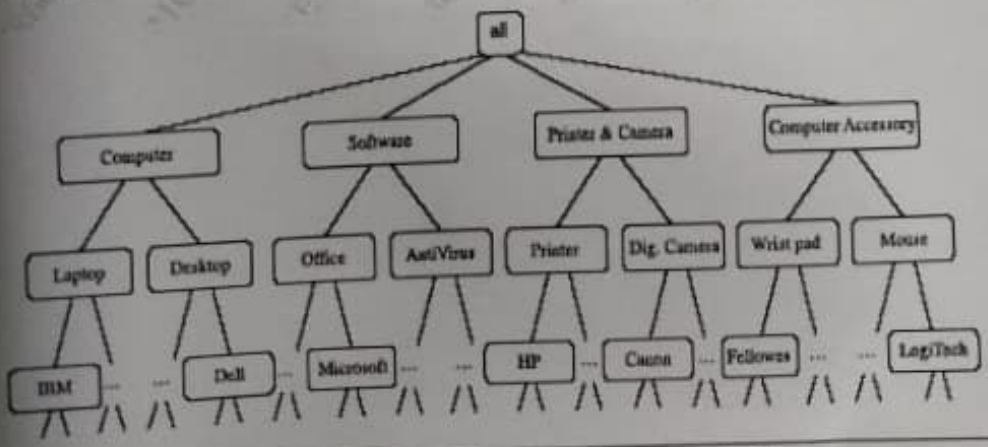
ay & Date: Monday, 30/10/2023 Time : 02.00 pm to 05.00 pm

Max Marks: **100**

IMP: Verify that you have received question papers with correct course code, branch etc.

structions
a) All questions are compulsory.
b) Writing question number on answer book is compulsory otherwise answers may not be assessed.
c) Assume suitable data wherever necessary.
d) Figures to the right of question text indicate full marks.
e) Mobile phones, smart gadgets and programmable calculators are strictly prohibited.
f) Except PRN anything else writing on question paper is not allowed.
g) Exchange/Sharing of stationery, calculator etc. not allowed.

		Marks	
xt on the right of marks indicates course outcomes (Only for faculty use)			
1 A)	Explain the problems/issues in Data Mining	4	CO1
B)	What you mean by binning? Smooth following data by Equal width and Equal frequency method. Data- 0, 4, 12, 16, 16, 18, 24, 26, 28	4	CO2
C)	Elaborate different normalization and techniques. Why normalization is used?	4	CO2
D)	Give Advantages and disadvantages of Data Mining.	4	CO3
2 A)	State data mining task primitives.	4	CO1
B)	State the basic principles of Attribute-Oriented Induction.	4	CO2
C)	What is a box plot? Mention the two conditions that represent the outliers. Find first and third quartile for following data 0, 1, 2, 2, 4, 5, 6, 9, 9, 10, 12, 14, 15, 15, 27.	4	CO3
D)	Define Quantitative Discriminant Rule. Assume a collage where 90 post-graduate students and 210 undergraduate students got placed. Find d_weight.	4	CO3
3 A)	Explain various approaches for mining multilevel association rules with reduced minimum support at lower levels. Elaborate one approach for following example.	4	CO2



B)	State Advantages and Disadvantages of apriori algorithm and FP-Tree algorithm.	4	CO2

Q5	A)	What are Packet Filter Firewall and Application Level Gateways?	8
	B)	Using a neat diagram, describe the working of dual home and single homed bastion host firewall configuration.	8
Q6	A)	Write short notes on: (Any Three) i) Hill Cipher ii) Steganography iii) Malwares iv) Password Management v) Active and Passive attacks	18

..... End of question paper .....

WALCHAND COLLEGE OF ENGINEERING, SANGLI.
(An Autonomous Institute)

ESE

Final Year B.Tech. (Information Technology)

END SEMESTER EXAMINATION SEM.- I NOVEMBER/DECEMBER - 2019

DATA MINING (3IT402)

Day, Date and Time: Monday, 02/12/2019, 10.00am to 12.00Noon

Exam Seat Number: _____

Max Marks: 50

IMP: Verify that you have received question paper with correct course, code, branch etc.

Instructions: i) All questions are compulsory. Writing question number is compulsory. Assume suitable data wherever necessary.

ii) Figures to the right of question text indicate full marks.

iii) Mobile phones and programmable calculators are strictly prohibited.

iv) Except Exam Seat Number writing anything on question paper is not allowed. Exchange/Sharing of stationery, calculator etc. not allowed.

Text on the right of marks indicates course outcomes (only for faculty use).

		Marks	
Q1	Fill in the blanks 1. Classification is _____ a. A subdivision of a set of examples into a number of classes b. A measure of the accuracy, of the classification of a concept that is given by a certain theory c. The task of assigning a classification to a set of examples d. None of these 2. Binary attribute are _____ a. This takes only two values. In general, these values will be 0 and 1 and .they can be coded as one bit b. The natural environment of a certain species c. Systems that can be used without knowledge of internal operations d. None of these 3. Multi-dimensional knowledge is a. A class of learning algorithms that try to derive a Prolog program from examples b. A table with n independent attributes can be seen as an n-dimensional space c. A prediction made using an extremely simple method, such as always predicting the same output. d. None of these 4. OLAP is used to explore the _____ knowledge a. Shallow b. Depp c. Hidden d. Multidimensional 5. Some telecommunication company wants to segment their customers into distinct groups in order to send appropriate subscription offers, this is an example of a. Supervised learning b. Data extraction c. Unsupervised learning d. None of all above 6. You are given data about seismic activity in Asia, and you want to predict a magnitude of the next earthquake, this is in an example of a. Supervised learning b. Unsupervised learning c. Dimensionality reduction d. None of all above	3	CO1
Q2 A)	State major Tasks in Data Preprocessing.	2	CO1
Q2 B)	Definee various Interesting measures required in the discovery of patterns.	2	CO2
Q2 C)	In following data, calculate information gain for "Department" attribute.	3	CO3

Gender	Department	Grade	Count
M	IT	B	16
F	IT	A	22
M	CSE	A	18
F	IT	A	25
M	IT	A	21
F	CSE	A	18

Target class: Campus Recruited

Gender	Department	Grade	Count
M	IT	B	16
F	ELN	C	22
M	ELN	C	18
F	IT	C	25
M	CSE	B	21
F	CSE	A	18

Contrasting class: Not Recruited

(P.T.O.)

Q4	D)	The following table shows distances between cities. Apply hierarchical clustering with single-linkage to form clusters. Terminate the solution after merging 6 cities.	5	CO3

	BOS	NY	DC	MIA	CHI	SEA	SF	LA	DEN
BOS	0								
NY	206	0							
DC	429	233	0						
MIA	1504	1308	1075	0					
CHI	963	802	671	1329	0				
SEA	2976	2815	2684	3273	2013	0			
SF	3095	2934	2799	3053	2142	808	0		
LA	2979	2786	2631	2687	2054	1131	379	0	
DEN	1949	1771	1616	2037	996	1307	1235	1059	0

Q5	A)	Write short note on - Sequence and Subsequence Matching.	3	CO3
Q5	B)	Write short note on - Hotspot analysis	3	CO2
Q5	C)	Write short note on - Mining the Web's Link Structures	3	CO3