ABES Engineering College, Ghaziabad

B. Tech Odd Semester Sessional Test-2

Printed Pages: 3

Session:

2022 - 2023

Course Code: **KDS-501**

Roll No:

Course Name: **INTRODUCTION TO DATA ANALYTICS AND VISUALIZATION** **Date of Exam:**

Maximum Marks: **75** **Time:**

Instructions:

1. **Attempt All sections.**
2. **If require any missing data, then choose suitably.**

Section-A

1. Attempt ALL Parts

a. Compare the stratified and cluster sampling methods.

Solution: Part 1

BASIS FOR COMPARISON	**STRATIFIED SAMPLING**	**CLUSTER SAMPLING**
Meaning	Stratified sampling is one, in which the population is divided into homogeneous segments, and then the sample is randomly taken	Cluster sampling refers to a sampling method wherein the members of the population are selected at random, from naturally

	from the segments.	occurring groups called 'cluster'.
Sample	Randomly selected individuals are taken from all the strata.	All the individuals are taken from randomly selected clusters.
Selection of population elements	Individually	Collectively
Homogeneity	Within group	Between groups
Heterogeneity	Between groups	Within group
Objective	To increase precision and representation.	To reduce cost and improve efficiency.

b. Interpret the Hierarchical clustering approach.

Solution: Part 1

- Hierarchical clustering is another unsupervised machine learning algorithm, which is used to group the unlabeled datasets into a cluster and also known as **hierarchical cluster analysis** or HCA.
- In this algorithm, we develop the hierarchy of clusters in the form of a tree, and this tree-shaped structure is known as the **dendrogram**.
- Sometimes the results of K-means clustering and hierarchical clustering may look similar, but they both differ depending on how they work. As there is no requirement to predetermine the number of clusters as we did in the K-Means algorithm.
- **It is of two types:**

1. **Agglomerative:** Agglomerative is a **bottom-up** approach, in which the algorithm starts with taking all data points as single clusters and merging them until one cluster is left.

2. **Divisive:** Divisive algorithm is the reverse of the agglomerative algorithm as it is a **top-down approach.**

- K-Means clustering always have predetermined number of clusters, and it always tries to create the clusters of the same size. To solve these two challenges, we can opt for the hierarchical clustering algorithm.

c. Explain the antecedent and consequent in Market Basket Analysis.

Solution: Part1

- Market Basket Analysis is modelled on Association rule mining, i.e., the IF {}, THEN {} construct. For example, IF a customer buys bread, THEN he is likely to buy butter as well.
- Association rules are usually represented as: {Bread} -> {Butter}
- Some terminologies to familiarize yourself with Market Basket Analysis are:
- **Antecedent:** Items or 'itemsets' found within the data are antecedents. In simpler words, it's the IF component, written on the left-hand side. In the above example, bread is the antecedent.
- **Consequent:** A consequent is an item or set of items found in combination with the antecedent. It's the THEN component, written on the right-hand side. In the above example, butter is the consequent.

d. Summarize the concept of Human Vision.

Solution: Part1 The human visual system can be regarded as consisting of two parts. The eyes act as image receptors which capture light and convert it into signals which are then transmitted to image processing centres in the brain. These centres process the signals received from the eyes and build an internal “picture” of the scene being viewed. Processing by the brain consists of partly of simple image processing and partly of higher functions which build and manipulate an internal model of the outside world. Although the

division of function between the eyes and the brain is not clear-cut, it is useful to consider each of the components separately.

e. Summarize the techniques used for Data Visualization.

Solution: Part1

Data visualization is a graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data.

Data Visualization Techniques

- Box plots
- Histograms
- Heat maps
- Charts
- Tree maps
- Word Cloud/Network diagram

Section-B

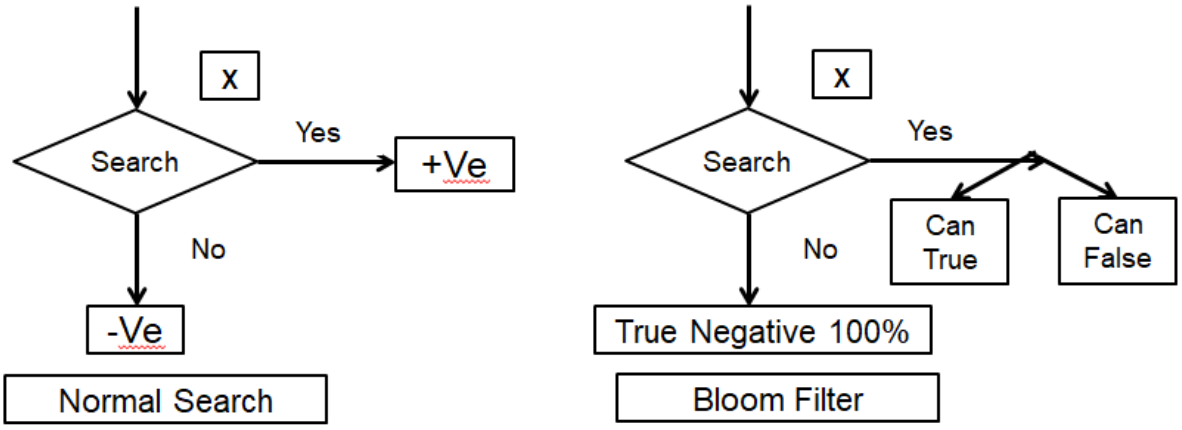
2. Attempt ANY ONE part from the following

a) Introduce the concept of Bloom Stream Filtering. Justify the statement "Bloom filter always produce True Negatives and False Positives" by implementing an example.

Solution: Part 1

- Another common process on stream is selection or filtering. We want to accept those tuples in the stream that meet a criteria. Accepted tuples are passed to the another process as a stream, while other tuples are dropped.
- If the selection criteria is the property of the tuple that can be calculated, then the selection is easy to do.
- The problem becomes harder when the criteria involves lookup for membership in a set. It is specially hard when that set is too large to store in main memory.
- Bloom filtering is the way to estimate most of the tuples that do not meet criteria.

- A Bloom filter is a space-efficient probabilistic data structure, conceived by Burton Howard Bloom in 1970.
- A Bloom filter is a data structure designed to tell you, rapidly and memory-efficiently, whether an element is present in a set.
- The price paid for this efficiency is that a Bloom filter is a probabilistic data structure: it tells us that the element either *definitely is not* in the set or *may be* in the set. A bloom filter is very much like a hash table in that it will use a hash function to map a key to a bucket.



b) Demonstrate the Estimating Moment in stream computing.

$$Kth - moment = \sum_{i \in A} (mi)^k$$

Calculate 0^{th} Moment, 1^{st} Moment, and 2^{nd} Moment of given stream: {10,9,9,9,9,9,9,9,9,9,9}.

Solution: Part-1

- Estimating moment is a generalization of counting distinct elements in a stream.
- The problem called "moments" involves the distribution of frequencies of different elements in the stream.
- **Goal:** Computing distribution of frequencies of different elements in stream.
- **Example:**
- **1^{st} Moment:** sum of all mi = count the number of elements.
- **2^{nd} Moment:** Sum of all mi^2 = surprise number of (s)
- A measure of how unseen the distribution is.
- **0^{th} Moment:** number of distinct elements. (A measure of how even the distribution is)

Part-2

- {10, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9}.

2nd moment = $10^2 + 10(9)^2 = 10^2 + 10 \times 81 = 910$.

3. Attempt ANY ONE part from the following.

a) Implement the three main components of Market Basket Analysis, Confidence, Support, and Lift, using a favorable dataset.

Solution: Part1

Jeans	Shirt	Jacket	Shoes
1	**1**	0	**1** ←
0	0	0	**1**
0	**1**	0	**1** ←
0	**1**	1	**1** ←
1	0	1	0
1	0	1	0
1	0	1	0
1	0	0	0
0	**1**	0	**1** ←
0	**1**	0	**1** ←
0	0	1	**1**
0	0	1	0
0	**1**	1	0
1	**1**	1	0
1	0	0	**1**
1	0	0	**1**
1	**1**	0	**1** ←
1	**1**	0	**1** ←
0	**1**	0	**1** ←

Lhs	Rhs	Frequency	Support	Confidence	Lift
Shirt	Shoes	8	0.42	0.8	1.27

Rule

Lhs → Rhs

Shirt → Shoes

Frequency

frq (Lhs , Rhs)

frq (Shirt , Shoes) = 8

Support

frq(Lhs , Rhs) / N

8 / 19 = 0.42

Confidence

frq (Lhs , Rhs) / frq (Lhs)

8 / 10 = 0.8

Lift

$$\frac{Support}{Support(Lhs) \times Support(Rhs)}$$

$$\frac{Support}{Support(Shirt) \times Support(Shoes)}$$

$$\frac{8/19}{10/19 \times 12/19} = 1.27$$

b) Show the steps involved in the Apriori algorithm for Market Basket Analysis.

Solution: Part 1

- The Apriori algorithm was proposed by Agrawal and Srikant in 1994.
- Apriori is designed to operate on databases containing transactions (for example, collections of items bought by customers, or details of a website frequentation).
- Apriori algorithm refers to the algorithm which is used to calculate the association rules between objects. It means how two or more objects are related to one another. In other words, we can say that the apriori algorithm is an association rule leaning that analyzes that people who bought product A also bought product B.
- The primary objective of the apriori algorithm is to create the association rule between different objects. The association rule describes how two or more objects are related to one

another. Apriori algorithm is also called frequent pattern mining. Generally, you operate the Apriori algorithm on a database that consists of a huge number of transactions.

- Step-1: Determine the support of itemsets in the transactional database, and select the minimum support and confidence.
- Step-2: Take all supports in the transaction with higher support value than the minimum or selected support value.
- Step-3: Find all the rules of these subsets that have higher confidence value than the threshold or minimum confidence.
- Step-4: Sort the rules as the decreasing order of lift.

4. Attempt ANY ONE part from the following.

a) Summarize the different tools available for effective Data Visualization.

Solution-Part1

Tableau Desktop –A business intelligence tool which helps you in visualizing and understanding your data.

●**ZohoReports** –ZohoReports is a self-servicebusiness intelligence (BI) and analyticstool that enables you to design intuitive data visualizations.

●**Microsoft Power BI** –Developed by Microsoft, this is a suite of business analytics tools that allows you to transform information into visuals.

●**MATLAB** –A detailed data analysis tool that has an easy-to-use tool interface and graphical design options for visuals.

●**Sisense**–A BI platform that allows you to visualize the information to make better and more informed business decisions.

b) Explain Human Vision and compare the human vision with computer vision.

Solution: Part-1

Perception:– Humans see objects, scenes, patterns, and people as they are, like trees in a landscape, people inside a car, clouds in a sky, or books in a shelf. Humans perceive the things as they are and retain what they recognize, storing it deep in the brain until they come across those things again. No obvious deductions or extra effort is required each object or people. Computer vision, on the other hand, allows computer to sense their surroundings and identify things, similar to how human vision perceive things.

Working – Human vision is all about eyes and how they detect light patterns and

coordinate with the brain to translate light into images that we see. Human eye is like a camera which needs light; when light hits the eyes, it forms a particular angle and the image is formed in the retina in the back of the eye, and the image is then inverted. Human vision requires coordination of the eye and the brain to function. Computer vision uses machine learning techniques and algorithms to identify, distinguish and classify objects by size or colour, and to discover and interpret patterns in visual data such as photos and videos. Computer vision simulates human vision by identifying objects in its field of vision.

Object Recognition – One of the key abilities of human vision system is invariant object recognition, meaning humans can instantly and accurately identify objects in different variations. Humans recognize objects effortlessly, and have no problems describing objects in a scene, even if they have never seen these objects before. Recognizing 3D objects from a single 2D image is one of the most challenging problems in computer vision. Computer needs to extract a set of features from the image to produce descriptions of the image different from an array of pixel values.

Section-C

5. Attempt ANY ONE part from the following.

a) Determine the distinct element in the stream using the Flajolet Martin algorithm.

Input stream X: {1,3,2,1,2,3,4,3,1,2,3,1}

Hash function, h(x) = 6x + 1 mod 5.

Solution: Part-1

Flajolet Martin Algorithm.

- Flajolet - Martin Algorithm approximate the number of unique objects in the stream or a database in one pass.
- If the stream consist n elements with m of them unique, this algorithm runs in O(n) time and needs O(log(m)) memory..
- Space consumption logarithmic in the maximal number of posible distinct elements in the stream.

Let us take an example to under F-M algorithm :-

Q:. Determine the distinct element in the stream using FM.

Input stream of integers X =
{1, 3, 2, 1, 2, 3, 4, 3, 1, 2, 3, 1}

Hash function, $h(x) = 6x + 1 \bmod 5$.

Soln.

Step-1:-
Calculate hash function $h(x)$ for the given input stream =
{1, 3, 2, 1, 2, 3, 4, 3, 1, 2, 3, 1}.

$h(x) = 6x + 1 \mod 5$

$h(1) = 2$	$h(4) = 0$
$h(3) = 4$	$h(3) = 4$
$h(2) = 3$	$h(1) = 2$
$h(1) = 2$	$h(2) = 3$
$h(2) = 3$	$h(3) = 4$
$h(3) = 4$	$h(1) = 2$

Step-2:

Evaluate binary bits.

- For every hash function calculated, write the binary equivalent for the same.

$h(1) = 2 = 010$	$h(4) = 0 = 000$
$h(3) = 4 = 100$	$h(3) = 4 = 100$
$h(2) = 3 = 011$	$h(1) = 2 = 010$
$h(1) = 2 = 010$	$h(2) = 3 = 011$
$h(2) = 3 = 011$	$h(3) = 4 = 100$
$h(3) = 4 = 100$	$h(1) = 2 = 010$

Step-3:

Counting the trailing zeroes.

Now we will count the number of trailing zeros in each hash function bit.

Step-3:

Counting the trailing zeroes.

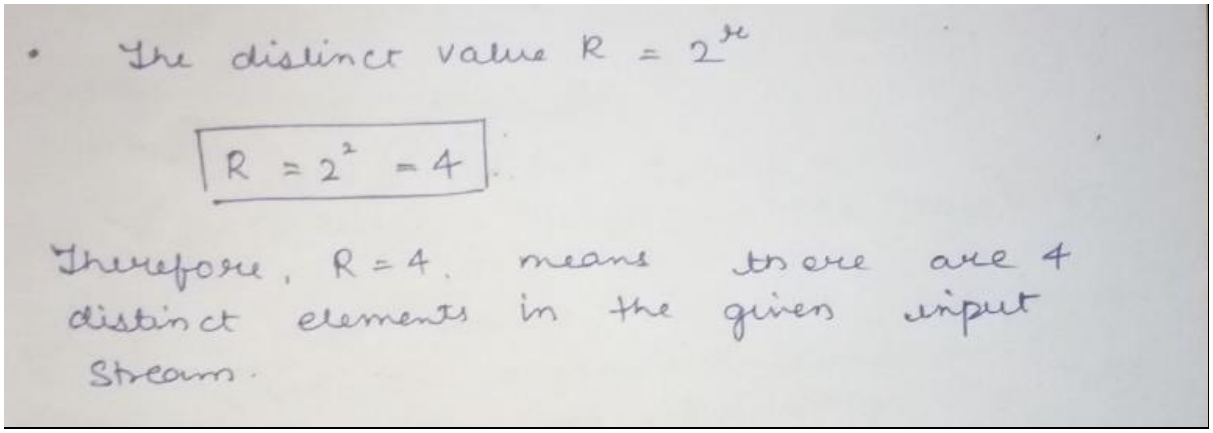
Now we will count the number of trailing zeros in each hash function bit.

$h(1) = 2 = 010 = 1$	$h(4) = 0 = 000 = 0$
$h(3) = 4 = 100 = 2$	$h(3) = 4 = 100 = 2$
$h(2) = 3 = 011 = 0$	$h(1) = 2 = 010 = 1$
$h(1) = 2 = 010 = 1$	$h(2) = 3 = 011 = 0$
$h(2) = 3 = 011 = 0$	$h(3) = 4 = 100 = 2$
$h(3) = 4 = 100 = 2$	$h(1) = 2 = 010 = 1$

Step-4:

Calculate the number / count of distinct elements.

- From the binary equivalent trailing zero values, write the number of (maximum) of trailing zeros.
- The value of $r = 2$.



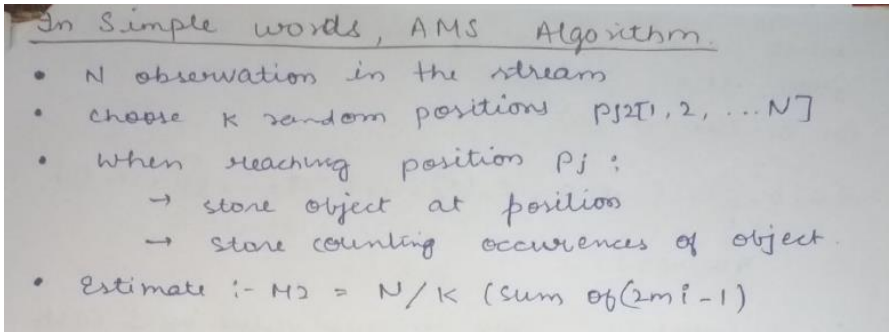
b) Elaborate the process of the Alon, Matias, Szegedy (AMS) algorithm used for limited space consumption by solving the given problem:

Stream: {a, b, c, b, d, a, c, d, a, b, d, c, a, a, b}

Length of Stream: 15

Random 3 Positions: c, d, a

Solution: Part1



Part 2

For example:-

consider this stream

{a, b, c, b, d, a, c, d, a, b, a, c, a, a, b}

By using 2nd moment:-

2nd moment = $a^2 + b^2 + c^2 + d^2$

= $5^2 + 4^2 + 3^2 + 3^2$ = 59

Problem solved by AMS Algorithm.

Given stream = {a, b, c, b, d, a, c, d, a, b, d, c, a, a, b}

Soln:-

length of the stream $n = 15$.

choose 3 random positions with different values say c, d, a

{a, b, (c), b, d, a, c, (d), a, b, d, c, (a), a, b}.

X_1.element X_2.element X_3.element.

Count the number of times of occurances from that positions

for X_1. element (c)

{a, b, (c), b, d, a, c, (d), a, b, d, c, (a), a, b}

X_1.value = 1

Set X.value = 1, now increase value by 1 each time we encounter another occurence of c.

{a, b, (c), b, d, a, c, d, a, d, c, a, a, b}

X_1.value = 1 X_1.value = 2 X_1.value = 3

We we repeat the same process for all 3 elements c, d, a.

{a, b, (c), b, (d), a, (c), (d), a, d, c, (a), a, b}.

X_1.value = 1 X_1.value = 2 X_2.value = 2 X_3.value = 1

X_2.value = 1 X_1.value = 3 X_3.value = 2.

So, we get.

X_1 = (c, 3)

X_2 = (d, 2)

X_3 = (a, 2)

Calculate the estimate

Estimate = (n x (2 * x-value - 1))

X_1 estimate = 15 x (2 * 3 - 1)

= 75

X_2 estimate = 15 x (2 * 2 - 1)

= 45.

X_3 estimate = 15 * (2 * 2 - 1)

= 45.

Calculate the average of X_1, X_2, X_3 :-

Avg = { Sum of estimates } / 3

= (75 + 45 + 45) / 3.

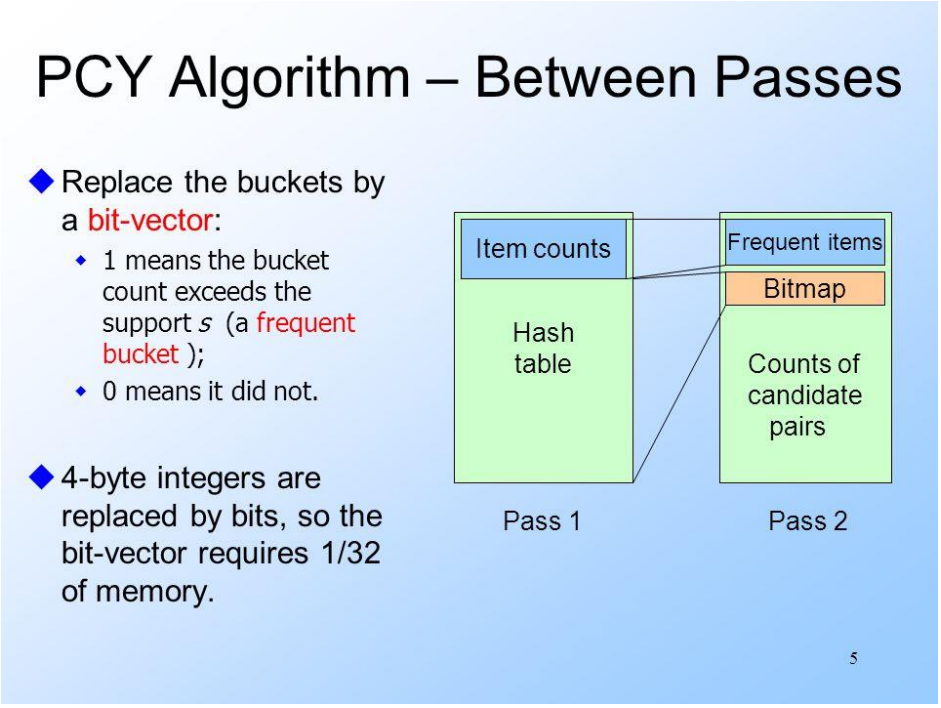
= 55

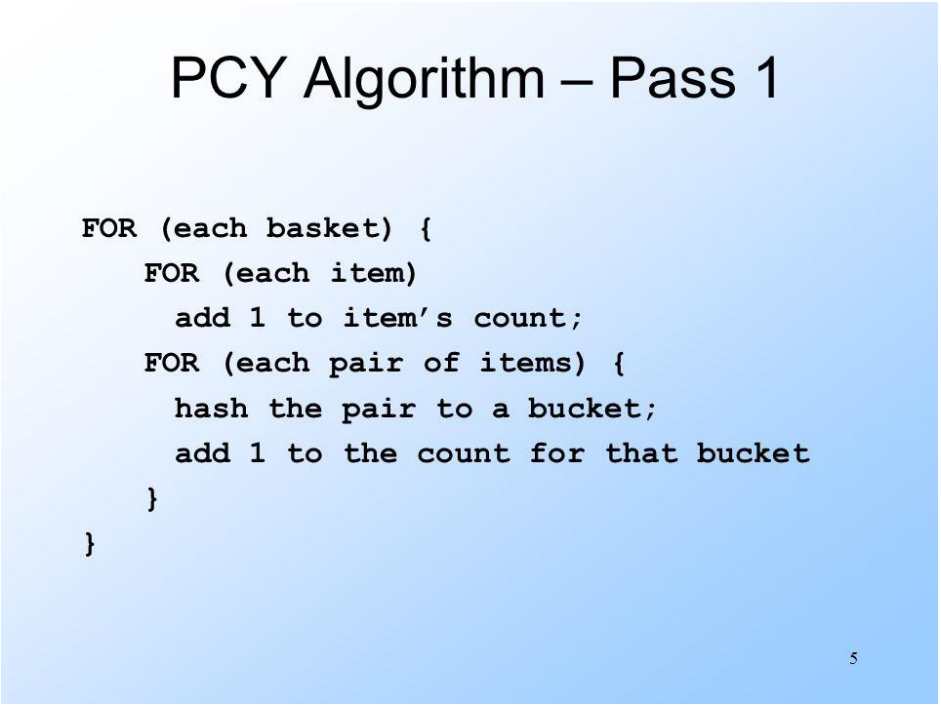
55 is somewhat close to the answer 59.

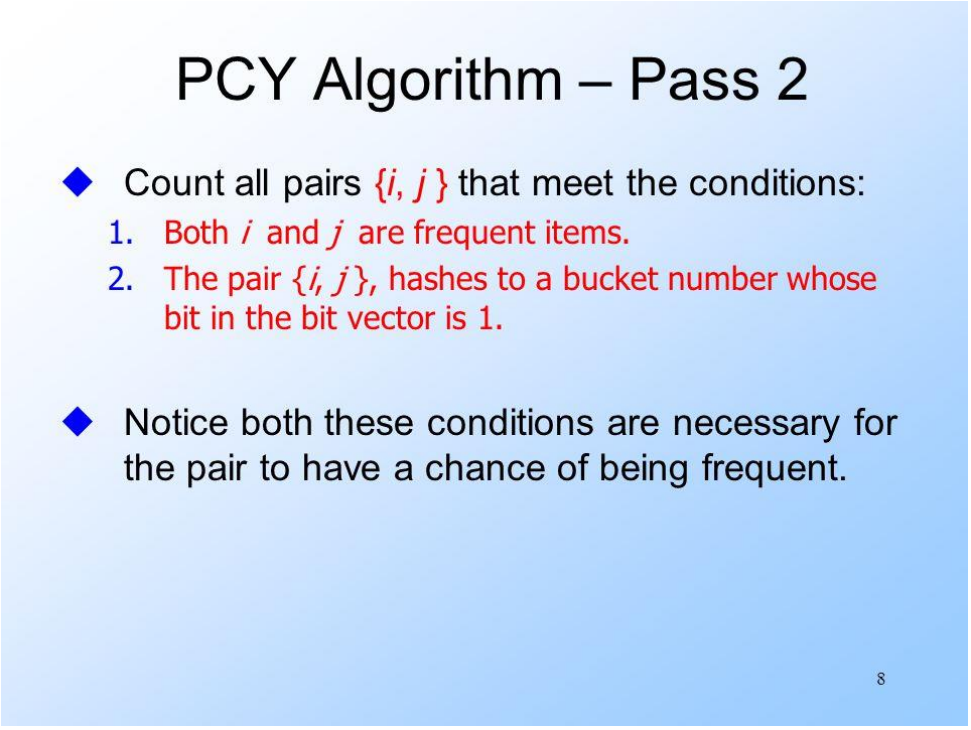
6. Attempt ANY ONE part from the following.

a) Explain the PCY algorithm for handling the extensive data in the main memory by taking suitable data items.

Solution: Part1







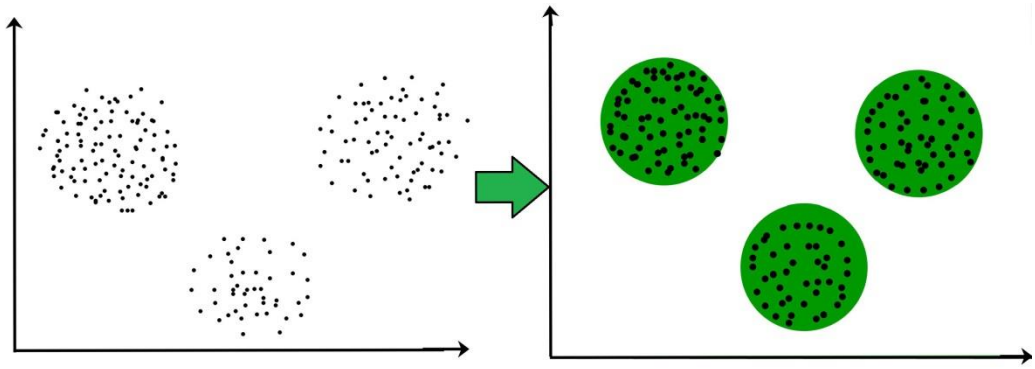
b) Summarize the clustering approach. Explain different kinds of clustering methods to achieve better-segregated groups.

Solution: Part 1

Clustering is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same

group and dissimilar to the data points in other groups. It is basically a collection of objects on the basis of similarity and dissimilarity between them.

The data points in the graph below clustered together can be classified into one single group. We can distinguish the clusters, and we can identify that there are 3 clusters in the below picture.



Part-2

- **Density-Based Methods:** These methods consider the clusters as the dense region having some similarities and differences from the lower dense region of the space. These methods have good accuracy and the ability to merge two clusters. Example *DBSCAN (Density-Based Spatial Clustering of Applications with Noise), OPTICS (Ordering Points to Identify Clustering Structure)*, etc.
- **Hierarchical Based Methods:** The clusters formed in this method form a tree-type structure based on the hierarchy. New clusters are formed using the previously formed one. It is divided into two category
  - **Agglomerative** (bottom-up *approach*)
  - **Divisive** (top-down *approach*)
- Examples CURE (Clustering Using Representatives), BIRCH (Balanced Iterative Reducing Clustering and using Hierarchies), etc.
- **Partitioning Methods:** These methods partition the objects into k clusters and each partition forms one cluster. This method is used to optimize an objective criterion similarity function such as when the distance is a major parameter example *K-means, CLARANS (Clustering Large Applications based upon Randomized Search)*, etc.

- **Grid-based Methods:** In this method, the data space is formulated into a finite number of cells that form a grid-like structure. All the clustering operations done on these grids are fast and independent of the number of data objects example *STING (Statistical Information Grid), wave cluster, CLIQUE (CLustering In Quest)*, etc.

7. Attempt ANY ONE part from the following.

a) Cluster the following eight points (with (x, y) representing locations) into three clusters: A1(2, 10), A2(2, 5), A3(8, 4), A4(5, 8), A5(7, 5), A6(6, 4), A7(1, 2), A8(4, 9) Initial cluster centers are: A1(2, 10), A4(5, 8) and A7(1, 2).

The distance function between two points a = (x1, y1) and b = (x2, y2) is defined as- P(a, b) = |x2 – x1| + |y2 – y1|

Use K-Means Algorithm to find the three cluster centers after the second iteration.

Solution: Part1

Iteration-01:

- We calculate the distance of each point from each of the center of the three clusters.
- The distance is calculated by using the given distance function.

The following illustration shows the calculation of distance between point A1(2, 10) and each of the center of the three clusters-

Calculating Distance Between A1(2, 10) and C1(2, 10)-

P(A1, C1)

= |x2 – x1| + |y2 – y1|

= |2 – 2| + |10 – 10|

= 0

Calculating Distance Between A1(2, 10) and C2(5, 8)-

P(A1, C2)

= |x2 – x1| + |y2 – y1|

= |5 – 2| + |8 – 10|

= 3 + 2

= 5

Calculating Distance Between A1(2, 10) and C3(1, 2)-

P(A1, C3)

= |x2 – x1| + |y2 – y1|

= |1 – 2| + |2 – 10|

= 1 + 8

= 9

In the similar manner, we calculate the distance of other points from each of the center of the three clusters.

Next,

- We draw a table showing all the results.
- Using the table, we decide which point belongs to which cluster.
- The given point belongs to that cluster whose center is nearest to it.

Given Points	Distance from center (2, 10) of Cluster-01	Distance from center (5, 8) of Cluster-02	Distance from center (1, 2) of Cluster-03	Point belongs to Cluster
A1(2, 10)	0	5	9	C1
A2(2, 5)	5	6	4	C3
A3(8, 4)	12	7	9	C2
A4(5, 8)	5	0	10	C2

A5(7, 5)	10	5	9	C2
A6(6, 4)	10	5	7	C2
A7(1, 2)	9	10	0	C3
A8(4, 9)	3	2	10	C2

From here, New clusters are-

Cluster-01:

First cluster contains points-

- A1(2, 10)

Cluster-02:

Second cluster contains points-

- A3(8, 4)
- A4(5, 8)
- A5(7, 5)
- A6(6, 4)
- A8(4, 9)

Cluster-03:

Third cluster contains points-

- A2(2, 5)
- A7(1, 2)

Now,

- We re-compute the new cluster clusters.
- The new cluster center is computed by taking mean of all the points contained in that cluster.

For Cluster-01:

We have only one point A1(2, 10) in Cluster-01.

- So, cluster center remains the same.

For Cluster-02:

Center of Cluster-02

= ((8 + 5 + 7 + 6 + 4)/5, (4 + 8 + 5 + 4 + 9)/5)

= (6, 6)

For Cluster-03:

Center of Cluster-03

= ((2 + 1)/2, (5 + 2)/2)

= (1.5, 3.5)

This is completion of Iteration-01.

Iteration-02:

We calculate the distance of each point from each of the center of the three clusters.

- The distance is calculated by using the given distance function.

The following illustration shows the calculation of distance between point A1(2, 10) and each of the center of the three clusters-

Calculating Distance Between A1(2, 10) and C1(2, 10)-

P(A1, C1)

= |x2 – x1| + |y2 – y1|

= |2 – 2| + |10 – 10|

= 0

Calculating Distance Between A1(2, 10) and C2(6, 6)-

P(A1, C2)

= |x2 – x1| + |y2 – y1|

= |6 – 2| + |6 – 10|

= 4 + 4

= 8

Calculating Distance Between A1(2, 10) and C3(1.5, 3.5)-

P(A1, C3)

= |x2 – x1| + |y2 – y1|

= |1.5 – 2| + |3.5 – 10|

= 0.5 + 6.5

= 7

In the similar manner, we calculate the distance of other points from each of the center of the three clusters.

Next,

- We draw a table showing all the results.
- Using the table, we decide which point belongs to which cluster.
- The given point belongs to that cluster whose center is nearest to it.

Given Points	Distance from center (2, 10) of Cluster-01	Distance from center (6, 6) of Cluster-02	Distance from center (1.5, 3.5) of Cluster-03	Point belongs to Cluster
A1(2, 10)	0	8	7	C1
A2(2, 5)	5	5	2	C3
A3(8, 4)	12	4	7	C2
A4(5, 8)	5	3	8	C2
A5(7, 5)	10	2	7	C2

A6(6, 4)	10	2	5	C2
A7(1, 2)	9	9	2	C3
A8(4, 9)	3	5	8	C1

From here, New clusters are-

Cluster-01:

First cluster contains points-

- A1(2, 10)
- A8(4, 9)

Cluster-02:

Second cluster contains points-

- A3(8, 4)
- A4(5, 8)
- A5(7, 5)
- A6(6, 4)

Cluster-03:

Third cluster contains points-

- A2(2, 5)
- A7(1, 2)

Now,

- We re-compute the new cluster clusters.
- The new cluster center is computed by taking mean of all the points contained in that cluster.

For Cluster-01:

Center of Cluster-01

= ((2 + 4)/2, (10 + 9)/2)

= (3, 9.5)

For Cluster-02:

Center of Cluster-02

= ((8 + 5 + 7 + 6)/4, (4 + 8 + 5 + 4)/4)

= (6.5, 5.25)

For Cluster-03:

Center of Cluster-03

= ((2 + 1)/2, (5 + 2)/2)

= (1.5, 3.5)

This is completion of Iteration-02.

After second iteration, the center of the three clusters are-

- C1(3, 9.5)
- C2(6.5, 5.25)
- C3(1.5, 3.5)

b) For the following given Transaction Data-set, Generate Rules using Apriori Algorithm. Consider the values as Support=50% and Confidence=75%

Transaction ID	Items Purchased
1	Bread, Cheese, Egg, Juice
2	Bread, Cheese, Juice
3	Bread, Milk, Yogurt
4	Bread, Juice, Milk
5	Cheese, Juice, Milk

Solution: Part-1

Step 1) Find Frequent Item Set and their support

Item	Frequency	Support (in %)
Bread	4	4/5=80%
Cheese	3	3/5=60%
Egg	1	1/5=20%
Juice	4	4/5=80%
Milk	3	3/5=60%
Yogurt	1	1/5=20%

Support (item) = Frequency of item/Number of transactions

Step 2) Remove all the items whose support is below given minimum support.

Item	Frequency	Support (in %)

Bread	4	4/5=80%
Cheese	3	3/5=60%
Juice	4	4/5=80%
Milk	3	3/5=60%

Step 3) Now form the two items candidate set and write their frequencies.

Items Pair	Frequency	Support (in %)
Bread, Cheese	2	2/5=40%
Bread, Juice	3	3/5=60%
Bread, Milk	2	2/5=40%
Cheese, Juice	3	3/5=60%
Cheese, Milk	1	1/5=20%
Juice, Milk	2	2/5=40%

Step 4) Remove all the items whose support is below given minimum support.

Items Pair	Frequency	Support (in %)
Bread, Juice	3	3/5=60%
Cheese, Juice	3	3/5=60%

Step 5) Generate rules

For Rules we consider item pairs:

a) (Bread, Juice)
Bread->Juice and Juice->Bread
b) (Cheese, Juice)
Cheese->Juice and Juice->Cheese

Confidence (A->B) = support (AUB)/support (A)

Therefore,

1. Confidence (Bread->Juice) = support (Bread U Juice)/support (Bread)
   = 3/5 * 5/4=3/4= 75%
2. Confidence (Juice->Bread) = support (Juice U Bread)/support (Juice)

= 3/5*5/4=3/4=75%

3. Confidence (Cheese->Juice) = support (Cheese U Juice)/support (Cheese)

   =3/5*5/3=1=100%

4. Confidence (Juice->Cheese) = support (Juice U Cheese)/support (Juice)

   = 3/5*5/4=3/4=75%

All the above rules are good because the confidence of each rule is greater than or equal to the minimum confidence given in the problem.

8. Attempt ANY ONE part from the following.

a) Visualization is needed to present the facts available in the unstructured datasets. Explain, in brief, the most challenging issues that occur during effective data visualization in real life.

Solution: Part 1

Data visualization is a graphical representation of any data or information. Visual elements such as charts, graphs, and maps are the few data visualization tools that provide the viewers with an easy and accessible way of understanding the represented information. In this world governed by Big Data, data visualization enables you or decision-makers of any enterprise or industry to look into analytical reports and understand concepts that might otherwise be difficult to grasp.

Part-2: Challenging Issues

1. Usability

- The usability issue is critical to everyone, especially in light of successful commercialization stories Although the overall growth of information visualization is accelerating, the growth of usability studies and empirical evaluations has been relatively slow. Furthermore, usability issues still tend to be addressed in an ad hoc manner and limited to the particular systems at hand.

2. Understanding elementary perceptual–cognitive tasks

- Understanding elementary and secondary perceptual– cognitive tasks is a fundamental step toward engineering information visualization systems. The general understanding of elementary perceptual–cognitive tasks must be substantially revised and updated in the context of information visualization.

3. Prior knowledge

- This seemingly philosophical problem has many practical implications. As a vehicle for communicating abstract information, information visualization and its users must have a common ground. This is consistent with the user-centered design tradition in human–computer interaction (HCI).

4. Education and training

- The education problem is the fourth user-centered challenge. We are facing the challenge internally and externally. The internal aspect of the challenge refers to the need for researchers and practitioners within the field of information visualization to learn and share various principles and skills of visual communication and semiotics.

5. Intrinsic quality measures

- It’s vital for the information visualization field to establish intrinsic quality metrics. Until recently, the lack of quantifiable quality measures has not been much of a concern. In part, this is because of the traditional priority of original and innovative work in this community. The lack of quantifiable measures of quality and benchmarks, however, will undermine information visualization advances, especially their evaluation and selection.

6. Scalability

- The scalability problem is a long-lasting challenge for information visualization. Unlike the field of scientific visualization, supercomputers have not been the primary source of data suppliers for information visualization. Parallel computing and other high-performance computing techniques have not been used in the field of information visualization as much as in scientific visualization and a few other fields. In addition to the traditional approach of developing increasingly clever ways to scale up sequential computing algorithms, the scalability issue should be studied at different levels—such as the hardware and the high-performance computing levels— as well as that of individual users.

7. Aesthetics

- The purpose of information visualization is the insights into data that it provides, not just pretty pictures. But what makes a picture pretty? What can we learn from making a pretty picture and enhancing the representation of insights? It's important, therefore, to understand how insights and aesthetics interact, and how these two goals could sustain insightful and visually appealing information visualization.

b) Although data visualization is a popular mechanism to get insight from data, it has various limitations. Discuss.

Solution: Part-1

- It gives assessment not exactness –
- While the information is exact in foreseeing the circumstances, the perception of similar just gives the assessment. It without a doubt is anything but difficult to change over the robust and protracted information into simple pictorial configuration yet such a portrayal of data may prompt theoretical ends now and then.
- **One-sided** –
- The essential arrangement of information representation occurs with the human interface, which means the information that turns out to be the base of perception

can be one-sided. The individual bringing the information for the equivalent may just think about the significant part of the information or the information that requirements center and may reject the remainder of the information which may prompt one-sided results.

- **Absence of help –**

  One of the downsides of information perception is that it can't help, which means an alternate gathering of the crowd may decipher it in an unexpected way.

- **Inappropriate plan issue –**

  On the off chance that information perception is viewed as such a correspondence. At that point, it must be certifiable in clarifying the reason. In the event that the plan isn't legitimate, at that point, this can prompt disarray in correspondence.

- **Wrong engaged individuals can skip center messages –**

  One of the issues with information perception is however it could be logical its clearness in clarification is totally subject to the focal point of its crowd.

9. Attempt ANY ONE part from the following.

a) Interpret the Design Exploration of Complex Information Space to provide a foundation for making design decisions.

Solution: Part1

More recent studies on actual information practices paint a different picture. There has been an increase in studies of information practices in everyday life uncovering complex information activities and strategies. Information researchers have more closely investigated the role of serendipitous discoveries for information seeking. Furthermore, pleasure and positive emotions have been identified as neglected aspects of many information seekers' experiences.

- These developments around every day, serendipitous, and positive information practices point to a perspective on information seeking that informs the research presented in this dissertation.
- Information spaces and their interfaces are not inevitable technical solutions, but cultural artefacts that need to be open for reflection, critique, and reinvention. Since growing information spaces form the backdrop of many human activities today they have a dual complexity, concerning both their technical realization and social adoption. On the one hand, growing information spaces raise technological challenges around scale, heterogeneity, and dynamics that are driving innovations in computer science and other areas of research.
- On the other hand, growing information spaces imply a social complexity with regard to communities and their representation, which is typically addressed in the humanities and social sciences. This dual nature between technical challenges and social implications is seldom considered in concert. For example, search can be seen as an engineering challenge to optimize precision and recall, yet it is important to realize that result rankings can also have embedded values with social or political ramifications.
- Technological trends on theWeb enable the design of novel interfaces for exploring growing information spaces. TheWeb, arguably the most significant information space today, is undergoing considerable transformations that enable entirely new ways of information seeking. In particular, developments around Web-based semantics, graphics, and interaction allow new ways for exploring information.

b) Explain the Space Perception and Data in Space in human vision.

Solution: Part-1

There are several striking similarities between growing cities of the 19th century and growing information spaces of today, especially with regard to the relation between the individual and the whole. As cities have become the cultural backdrops of daily activities for the majority of people in the world, digital information spaces increasingly assume a similar role. In the following, we briefly highlight growth, significance, and conflict as important commonalities.

- **Growth.** The city of the flaneur and today's information spaces continuously grow. In both cases, there is a significant discrepancy between the individual and the isproportionately large—urban or digital—environment.
- **Significance.** Like the late-modern city can be seen as a grandiose cultural artefact, information spaces arguably form the culturally significant phenomenon of our times. They are becoming an important context for our daily activities as part of work, play, and community.
- **Conflict.** Cities and information spaces are also contexts for social struggle and negotiation. Urban issues such as acceleration and alienation do not remain uncontested. Similarly, information spaces pose issues such as copyright, network neutrality, and information poverty.

Considering these parallels between cities and information spaces, we see the flaneur as a lens through which to investigate new perspectives on information seeking. We are particularly interested in his exploratory mindset towards the city. In order to experience the city, the flaneur does not methodically navigate streets, checking each edifice like a building inspector in search of code violations. Nor does the flaneur hastily interrogate each city-dweller, like a police officer in search of a thief. Because the flaneur does not accurately scrutinize everything that crosses his path, he is able to sense what city life is about on a higher level. It can be argued that the flaneur is the embodiment of exploration and serendipity, while the police officer and building inspector personify traditional search and browsing.