Savitribai Phule Pune University
Final Year of Information Technology (2019 Course)
(With effect from Academic Year 2022-23)

Semester VII

Course Code	Course Name	Teaching Scheme(Hours/week)			Examination Scheme and Marks						Credit Scheme			
		Lecture	Practical	Tutorial	Mid-Sem	End-Sem	Termwork	Practical	Oral	Total	Lecture	Practical	Tutorial	Total
414441	Information and Storage Retrieval	03	-	-	30	70	-	-	-	100	3	-	-	3
414442	Software Project Management	03	-	-	30	70	-	-	-	100	3	-	-	3
414443	Deep Learning	03	-	-	30	70	-	-	-	100	3	-	-	3
414444	Elective III	03	-	-	30	70	-	-	-	100	3	-	-	3
414445	Elective IV	03	-	-	30	70	-	-	-	100	3	-	-	3
414446	Lab Practice III	-	04	-	-	-	25	-	25	50	-	2	-	2
414447	Lab Practice IV	-	02	-	-	-	25	25	-	50	-	1	-	1
414448	Project Stage-I	-	-	02	-	-	50	-	-	50	-	-	2	2
414449	Audit Course7													
										Total Credit	15	03	02	20
	Total	15	06	02	150	350	100	25	25	650	15	03	02	20

Elective III:

- Mobile Computing
- High Performance Computing
- Multimedia Technology
- Smart Computing

Elective IV:

- Bioinformatics
- Introduction to DevOps
- Computer Vision
- Wireless Communications

Lab Practice-III:

It is based on subjects:

- Information and Storage Retrieval

Lab Practice-IV:

It is based on subjects:

- Deep Learning

B.E. (INFORMATION TECHNOLOGY)

SEMESTER-VII

Time– 2.00 p.m. to 3.00 p.m.

Day& Date	SUBJECT (2019 Course)	SUBJECT CODE
Tuesday 19/08/2025	Information and Storage Retrieval	414441
Wednesday 20/08/2025	Software Project Management	414442
Thursday 21/08/2025	Deep Learning	414443
Friday 22/08/2025	(ELECTIVE-III) Mobile Computing	414444A
	(ELECTIVE-III) High Performance Computing	414444B
	(ELECTIVE-III) Multimedia Technology	414444C
	(ELECTIVE-III) Smart Computing	414444D
Saturday 23/08/2025	(ELECTIVE-IV)Bioinformatics	414445A
	(ELECTIVE-IV) Introduction to DevOps	414445B
	(ELECTIVE-IV) Computer Vision	414445C
	(ELECTIVE-IV) Wireless Communications	414445D

COURSE CONTENTS		
Unit I	**Introduction to Information Retrieval**	**(06 hrs)**
Basic Concepts of IR, Data Retrieval & Information Retrieval, Text mining and IR relation, IR system block diagram, **Automatic Text Analysis:** Luhn's ideas, Conflation Algorithm, Indexing and Index Term Weighting, Probabilistic Indexing, Automatic Classification. Measures of Association, Different Matching Coefficients, Cluster Hypothesis, **Clustering Techniques: Rocchio's** Algorithm, Single pass algorithm, Single Link algorithm.		
Mapping of Course Outcomes for Unit I	**CO1**	
Unit II	**Indexing and Searching Techniques**	**(06 hrs)**
Indexing: Inverted file, Suffix trees & suffix arrays, Signature Files, Scatter storage or hash addressing. **Searching Techniques:** Boolean Search, sequential search, Serial search, cluster-based retrieval, Query languages, Types of queries, Patterns matching, structural queries. **IR Models:** Basic concepts, Boolean Model, Vector Model, Probabilistic Model.		
Mapping of Course Outcomes for Unit II	**CO2**	

Unit I – Introduction to Information Retrieval (06 hrs)

1.Basic Concepts of IR

1. Data Retrieval & Information Retrieval
2. Relation between Text Mining and IR
3. IR system block diagram

2.Automatic Text Analysis

1. Luhn's ideas
2. Conflation Algorithm
3. Indexing and Index Term Weighting
4. Probabilistic Indexing
5. Automatic Classification

3.Measures of Association

1. Different Matching Coefficients
2. Cluster Hypothesis

4.Clustering Techniques

1. Rocchio's Algorithm
2. Single Pass Algorithm
3. Single Link Algorithm

Q1) a) Explain the process to generate document representative of a document with proper example using conflation algorithm. **[7]**

b) Differentiate between data retrieval & information retrieval. **[4]**

c) Explain information retrieval process with the help of block diagram. **[4]**

OR

Q2) a) Explain single pass clustering algorithm with a suitable example. **[8]**

b) What are the different measures of association? Explain any three matching coefficients with suitable examples. **[7]**

Q1) a) Analyse the concept behind the single pass algorithm and the single link algorithm for clustering. Compare and contrast these two techniques, highlighting their advantages, disadvantages, and situations where each approach would be more suitable. Provide a scenario for each algorithm to illustrate their application. **[8]**

b) Differentiate between data retrieval and information retrieval. **[4]**

c) Define the following terms. **[3]**

i) Precision

ii) Recall

OR

Q2) a) Explain Conflation algorithm to generate document representative of a document with a proper example. **[8]**

b) What are the different measures of association? Explain any three matching coefficients with suitable examples. **[7]**

Q1) a) Explain information retrieval process with the help of block diagram. **[4]**

b) Explain conflation algorithm in detail with its advantages & disadvantages. **[8]**

c) Why is index term weighting used? **[3]**

OR

Q2) a) Explain single link clustering algorithm with suitable example. **[8]**

b) Discuss difference between data retrieval & information retrieval. **[4]**

c) Explain Rocchio's algorithm in brief. **[3]**

1.Basic Concepts of IR

1. Data Retrieval & Information Retrieval
2. Relation between Text Mining and IR
3. IR system block diagram

b) Differentiate between data retrieval & information retrieval. **[4]**

Explain information retrieval process with the help of block diagram. **[4]**

1. Data Retrieval & Information Retrieval

1. Data Retrieval (DR)

•Definition: Data retrieval ka matlab hai exact, structured data nikalna from a database ya kisi storage system, jahan tum already format aur structure jaante ho.

•Nature: Pre-defined queries ke basis pe exact match nikalta hai.

•Source: Structured data (tables, rows, columns).

•Accuracy: Exact and deterministic result — jo maanga, wahi milega.

Example (Real Life):

•Bank database me tum ek query run karte ho:

"Mujhe customer ID 12345 ka balance dikhao."
Database tumhe bilkul wahi balance amount return karega — na zyada, na kam.

•Jaise **Phone contact list** me kisi ka naam search karna — "Rohan Sharma" likha, to bas wahi entry milegi

2. Information Retrieval (IR)

•Definition: Information retrieval ka matlab hai unstructured ya semi-structured data me se relevant information nikalna, mostly text, audio, video, etc.

•Nature: Keyword-based search, ranking, relevance — perfect match zaroori nahi.

•Source: Unstructured data (documents, web pages, images).

•Accuracy: Approximate / best match results based on relevance score.

Example (Real Life):

•Google pe search karte ho:

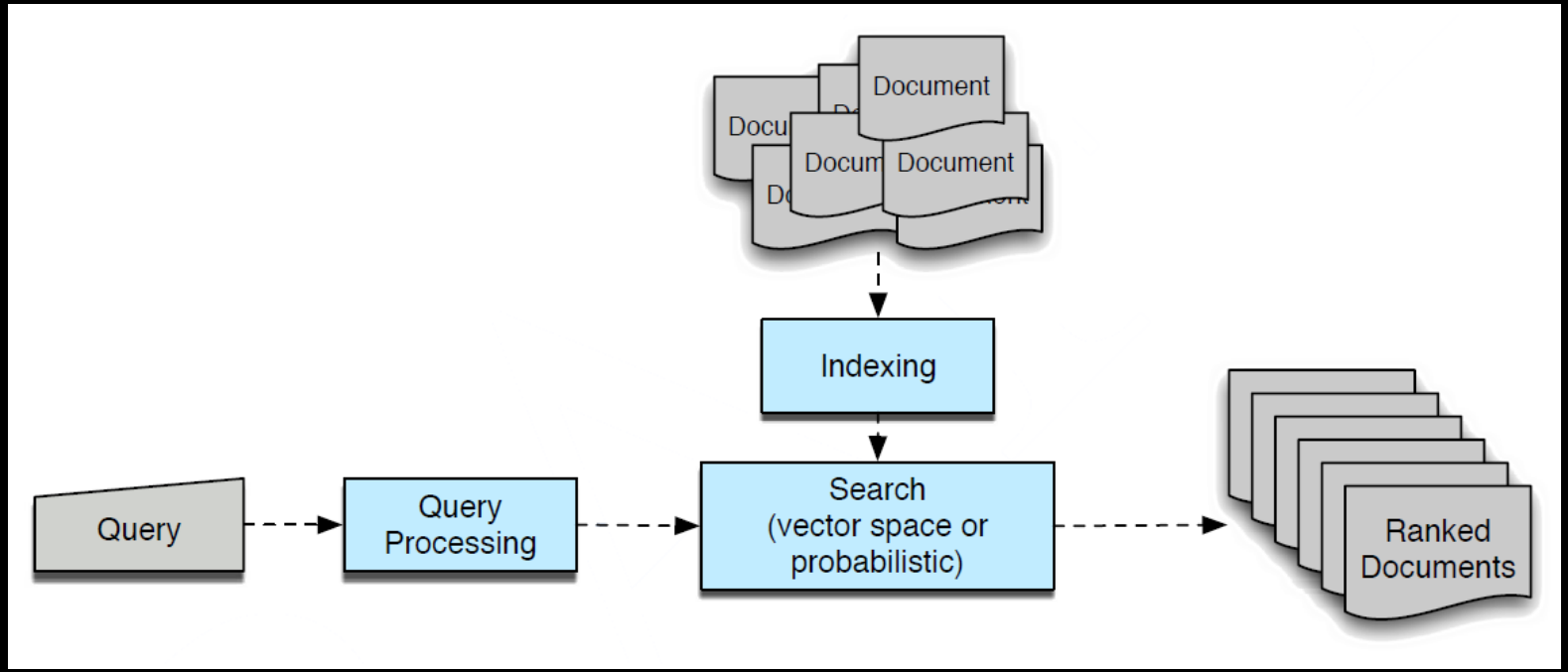
"Best restaurants near me“

Tumne exact data nahi diya, but search engine tumhare query ke context samajh ke multiple restaurants ki list dega, sorted by relevance.

•Jaise **YouTube search** — tum “C programming" likhte ho, to exact file name match nahi, par related aur relevant videos dikhega.

Feature	Data Retrieval (DR)	Information Retrieval (IR)
Data type	Structured	Unstructured / Semi-structured
Match type	Exact match	Relevant / best match
Source	Database, tables	Documents, web, media
Example	Bank account balance query	Google search / YouTube search
Accuracy	100% exact	Relevance-based

Information Retrieval Process



Step-by-step:

1.User Query

1. User apna search request deta hai (keywords, sentence, ya question).
2. Example: "Best coffee shops in Delhi".

2.Query Processing

1. System query ko clean, tokenize, aur optimize karta hai taaki search better ho.
2. Example: Stop words remove karna ("in", "the", etc.).

3.Searching in Index

1. Documents ka ek **index** already bana hota hai (like a book index).
2. System relevant documents ko quickly locate karta hai.

4.Matching & Ranking

1. Search engine match karta hai documents ke content ko query se.
2. Relevant results ko rank karta hai based on importance/relevance.

5.Results Display

1. Top ranked results user ko show kiye jaate hain.

Example (Real Life):
Tum Google pe search karte ho: *"cheap hotels in Goa"*
•Google tumhara query process karta hai
•Index me se relevant hotel pages find karta hai
•Ranking ke saath tumhe top results show karta hai

1.Automatic Text Analysis

1. Luhn's ideas
2. Conflation Algorithm
3. Indexing and Index Term Weighting
4. Probabilistic Indexing
5. Automatic Classification

a) Explain the process to generate document representative of a document with proper example using conflation algorithm. [7]

c) Why is index term weighting used?

1. Luhn’s Ideas

•**Concept:** Hans Peter Luhn ne text summarization aur keyword extraction ke liye statistical approach di.

•**Kaise kaam karta hai:** High-frequency words ko important maana jata hai aur document ka summary ya indexing un words pe based hoti hai.

•**Example:** Agar tumhare paas ek article hai “Climate Change and Global Warming”, aur “climate” aur “temperature” words bahut baar aate hain, to system unko key terms maanega.

2.Conflation Algorithm :

Conflation ka matlab hai **different word forms ko ek single root ya stem form me lana**, taaki indexing aur searching me uniformity rahe.

•Ye process **word stemming** ya **lemmatization** ke naam se bhi jana jata hai.

•Iska main goal hai:

- Synonyms ya morphological variations ko ek base form me convert karna.
- Search results ka coverage improve karna.

Process to Generate Document Representative using Conflation Algorithm

1.Document Collection
- •System ke paas ek ya multiple documents hote hain jinka indexing karna hai.

2.Tokenization
- •Text ko chhote units (words/tokens) me break kiya jata hai.
- •Example: "The connections were established" → ["The", "connections", "were", "established"]

3.Stop-word Removal
- •Common words (like "the", "is", "were") jo search me useful nahi hote, remove kiye jate hain.

4.Conflation / Stemming
- •Remaining words ko unke root form me convert kiya jata hai.
- •Example: connections → connect, established → establish

5.Indexing
- •Root words ka index banaya jata hai with frequency count.
- •Example: connect: 5, establish: 3

6.Document Representative Generation
- •Final set of root words with their importance (TF-IDF weighting) ban jata hai jo document ka summary representation hota hai.

Example
Original Document:
"Connections between computers are established using networking cables."

Steps:
1.Tokenization → ["Connections", "between", "computers", "are", "established", "using", "networking", "cables"]
2.Stop-word Removal → ["Connections", "computers", "established", "networking", "cables"]
3.Conflation (Stemming) → ["connect", "computer", "establish", "network", "cable"]
4.Frequency Counting →

- ❑ connect: 1
- ❑ computer: 1
- ❑ establish: 1
- ❑ network: 1
- ❑ cable: 1

5.This set becomes the document representative.

3. Indexing and Index Term Weighting

Why is Index Term Weighting Used?

Definition:

Index term weighting ka matlab hai har term (keyword) ko ek **importance score** dena, taaki pata chale kaunsa term document ke liye kitna relevant hai.

Reasons for Using Index Term Weighting:

1.Identify Important Terms

1. Sab words equal important nahi hote. Weighting se pata lagta hai kaunse words search ke liye useful hain.
2. Example: “machine learning” ka weight high hoga, “the” ka weight low hoga.

2.Improve Search Accuracy

1. High-weight terms ko zyada preference di jati hai, taaki search results relevant aayein.
2. Example: Agar query “AI in healthcare” hai, to “healthcare” ka weight zyada hone se health-related documents pehle aayenge.

3.Handle Common Words

1. Stop words ya bahut common words (like “is”, “and”) ka weight kam hota hai, taaki wo results ko distort na kare.

4.Rank Documents

1. TF-IDF jaise methods se weight calculate karke documents ko rank kiya jata hai.
2. High-weight terms wale documents top results me aate hain.

Example:

Document:

> "Machine learning is a part of artificial intelligence."

Weights (example values):

machine: 0.8
learning: 0.8
artificial: 0.7
intelligence: 0.7
is: 0.1
a: 0.1
part: 0.3
of: 0.1

Yaha "machine" aur "learning" ka weight high hai, isliye search engine in terms ko zyada importance dega.

Explain conflation algorithm in detail with its advantages & disadvantages. [8]

Advantages

1. Improves search coverage – Different word forms (connect, connected, connection) ko ek root form me laane se relevant documents miss nahi hote.

2. Reduces index size – Word variations merge ho jaati hain, isliye indexing me kam space lagta hai.

3. Faster retrieval – Kam unique terms hone se search engine ko query match karne me kam time lagta hai.

4. Simplifies matching – Query aur document ke beech matching easy ho jati hai, even if words ke forms alag hain.

Disadvantages

1. Over-stemming issue – Kabhi kabhi unrelated words bhi same root me aa jaate hain, jisse irrelevant results milte hain.

2. Loss of meaning – Word ka exact sense ya grammatical form kho sakta hai (e.g., “organ” aur “organization” dono “organ” ban jate hain).

3. Synonyms not handled – Algorithm synonyms ko same nahi maanta unless extra semantic processing ho.

4. Language dependency – Har language ke liye alag stemming rules banane padte hain.

4. Probabilistic Indexing

•**Concept:** Probability ka use karke decide kiya jata hai ki koi document query ke liye kitna relevant hai.

•**Example:** Agar query hai “renewable energy” aur ek document me 80% chance hai ki wo relevant hoga, to wo top results me aayega.

5. Automatic Classification

•**Concept:** Documents ko automatically categories me daalna, machine learning models ka use karke.

•**Example:** News articles ko categories me divide karna — “Sports”, “Politics”, “Technology” — without manual tagging.

1.Measures of Association

1. Different Matching Coefficients
2. Cluster Hypothesis

Q2) a) Explain single pass clustering algorithm with a suitable example. **[8]**

b) What are the different measures of association? Explain any three matching coefficients with suitable examples. **[7]**

Measures of Association

Definition:
Measures of Association ka use documents ya terms ke beech similarity (kitne related hain) ya dissimilarity (kitne alag hain) measure karne ke liye hota hai.

Ye **Information Retrieval** me documents ko group karne, search results rank karne aur clustering me use hota hai.

Types:
1.Different Matching Coefficients – Numerical values jo batate hain do sets (terms/documents) kitne similar hain.

2.Cluster Hypothesis – Agar do documents ek relevant document ke similar hain, toh woh apas me bhi similar honge.

Three Matching Coefficients (Examples ke saath)

1. Jaccard Coefficient

- Formula:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

- **Example**:

  Document1 terms = {apple, mango, orange}

  Document2 terms = {apple, banana, mango}

$$J = \frac{2}{4} = 0.5$$

(Common terms = apple, mango → 2 terms; Total unique terms = 4)

2. Cosine Similarity

- Formula:

$$CosSim(A, B) = \frac{A \cdot B}{||A|| \times ||B||}$$

- **Example**:

  A = (1,1,0,0)

  B = (1,0,1,0)

$$CosSim = \frac{1}{\sqrt{2} \times \sqrt{2}} = 0.5$$

(Measures angle between vectors)

3. Dice Coefficient

- Formula:

$$Dice(A, B) = \frac{2|A \cap B|}{|A| + |B|}$$

- **Example**:

  Document1 = {apple, mango, orange}

  Document2 = {apple, banana, mango}

$$Dice = \frac{2 \times 2}{3 + 3} = \frac{4}{6} \approx 0.67$$

Coefficient	Formula	Range	Example Value
Jaccard	(	A ∩ B	/
Cosine	((A·B)/		A
Dice	(2	A ∩ B	/ (

2. Cluster Hypothesis

•Meaning: Agar ek document relevant hai, toh uske nearby cluster me maujood documents bhi likely relevant honge.

•Purpose: Documents ko clusters me group karke retrieval performance improve karna.

•Working:

- Similar documents ko group karo (clustering).
- Agar ek group me ek document relevant mila, toh us group ke baaki documents ko bhi recommend karo.

Example:

Tum ek movie recommendation system use kar rahe ho. Tumhe "Inception" pasand aayi (cluster = sci-fi thrillers), toh system tumhe "Interstellar" aur "The Matrix" bhi suggest karega jo same cluster me hain.

1.Clustering Techniques

1. Rocchio's Algorithm
2. Single Pass Algorithm
3. Single Link Algorithm

a) Explain single pass clustering algorithm with a suitable example. **[8]**

Q1) a) Analyse the concept behind the single pass algorithm and the single link algorithm for clustering. Compare and contrast these two techniques, highlighting their advantages, disadvantages, and situations where each approach would be more suitable. Provide a scenario for each algorithm to illustrate their application. **[8]**

Explain Rocchio's algorithm in brief.

1. Rocchio’s Algorithm

Explain Rocchio’s algorithm in brief.

Rocchio’s Algorithm (Brief)

Definition
Rocchio’s algorithm ek **relevance feedback** method hai jo **Vector Space Model** me query ko modify karta hai.

Iska aim hai query vector ko **relevant documents ke closer** aur **irrelevant documents se away** shift karna.

Formula

$$Q_{new} = \alpha Q_{original} + \frac{\beta}{|D_r|} \sum_{d \in D_r} d - \frac{\gamma}{|D_{nr}|} \sum_{d \in D_{nr}} d$$

- **Q_original** = initial query vector
- **D_r** = relevant documents ka set
- **D_nr** = non-relevant documents ka set
- **α, β, γ** = tuning constants (importance weight)

Steps

1. Initial query ka vector banao.
2. Relevant docs ke centroid calculate karo.
3. Irrelevant docs ke centroid calculate karo.
4. New query vector calculate karo using above formula.
5. New query se search fir se run karo for better results.

Example

- **Initial Query**: "Machine Learning" → Q = (0.5, 0.4, 0.1)
- **Relevant Docs**:

  D1 = (0.6, 0.5, 0.1)

  D2 = (0.7, 0.4, 0.2)
- **Non-Relevant Docs**:

  D3 = (0.1, 0.3, 0.6)
- **α = 1, β = 0.75, γ = 0.15**

Calculation:

1. Centroid of relevant docs = (0.65, 0.45, 0.15)
2. Centroid of non-relevant docs = (0.1, 0.3, 0.6)
3.

$$Q_{new} = (0.5, 0.4, 0.1) + 0.75 \times (0.65, 0.45, 0.15) - 0.15 \times (0.1, 0.3, 0.6)$$

$$Q_{new} = (0.9875, 0.7375, 0.1725)$$

Result: New query vector is more biased toward relevant terms.

Advantages

- Improves retrieval accuracy with feedback.
- Adapts query to user needs.
- Works well in vector space model.

Disadvantages

- Needs user feedback (relevant/irrelevant marking).
- Extra computation.
- Not suitable if feedback data is small.

a) Explain single pass clustering algorithm with a suitable example. [8]

Single Pass Clustering Algorithm

Definition
Single Pass Clustering ek incremental clustering technique hai jisme har document ko sequentially (ek ek karke) dekha jata hai aur decide kiya jata hai ki woh existing cluster me add hoga ya naya cluster banayega.
Isme data ek hi baar process hota hai (single pass), isliye fast hota hai.

Steps / Working:

1.Start: Pehla document lo, ek naya cluster banao.

2.For each next document:

1. **Document ka similarity sab existing clusters ke centroid se calculate karo.**
2. **Agar similarity ≥ threshold → document ko us cluster me daal do.**
3. **Agar similarity < threshold for all clusters → naya cluster banao.**

3.Repeat jab tak saare documents process na ho jaye.

4.End: Sab clusters ready.

Example

Documents:

- D1 = {apple, mango, banana}
- D2 = {apple, mango}
- D3 = {car, bus, train}
- D4 = {bus, truck}

Threshold = 0.5 (similarity)

Process:

1. D1 → Cluster 1 bana.
2. D2 → Similarity with C1 high (0.8) → C1 me add.
3. D3 → Similarity with C1 low (0.0) → Cluster 2 bana.
4. D4 → Similarity with C2 high (0.6) → C2 me add.

Final Clusters:

- C1 = {D1, D2}
- C2 = {D3, D4}

Advantage	Explanation
Fast	Sirf ek baar data pass hota hai
Simple	Easy to implement
Incremental	Naye data ko turant add kar sakte ho

Disadvantage	Explanation
Order Dependent	Documents ka order badalne se result change ho sakta hai
Not Optimal	Best cluster guarantee nahi deta
Threshold Sensitive	Threshold galat set hua to poor clustering hogi

Single Link Clustering Algorithm
Definition
Single Link Clustering ek **hierarchical agglomerative clustering** method hai jisme **dono clusters ke sabse nearest points ke distance** ko use karke merge kiya jata hai.
Isme “link” ka matlab hai **minimum distance** between two clusters.

Steps
1.Har data point ko ek separate cluster maan lo.
2.Sabhi clusters ke beech ka minimum pairwise distance nikal lo.
3.Do clusters jinka minimum distance sabse kam hai, unko merge kar do.
4.Distance matrix update karo:
1. **New cluster ka distance kisi bhi cluster se = minimum distance of any point in new cluster with that cluster.**

5.Step 2–4 repeat karo jab tak ek single cluster na ban jaye (ya desired number of clusters mil jaye).

Distance Formula (Single Link)

$$d(C_i, C_j) = \min\{d(x, y) \mid x \in C_i, y \in C_j\}$$

Example

Data points: **A, B, C, D**

Distance matrix:

	A	B	C	D
A	0	2	6	10
B	2	0	5	9
C	6	5	0	4
D	10	9	4	0

Step-by-step:

1. Nearest pair: **A-B** (distance = 2) → Merge into Cluster {A,B}.
2. New distances:
   - {A,B} to C = min(6, 5) = 5
   - {A,B} to D = min(10, 9) = 9
   - C to D = 4
3. Nearest pair: **C-D** (distance = 4) → Merge into {C,D}.
4. New distances:
   - {A,B} to {C,D} = min(5, 9, 6, 4) = 4
5. Merge → Final cluster {A,B,C,D}.

Advantages

- Works well for **non-spherical shapes** of clusters.
- Captures chaining effect (good for continuous patterns).

Disadvantages

- Sensitive to **noise & outliers** (chaining effect may merge unrelated points).
- May produce long, thin clusters.

Feature	Single Pass Algorithm	Single Link Algorithm
Clustering Type	Incremental, non-hierarchical	Hierarchical (agglomerative)
Speed	Very fast (O(n))	Slower ($O(n^2)$)
Memory	Low	Higher
Flexibility	Needs pre-defined threshold	No threshold needed
Output	Final clusters directly	Dendrogram + possible different cluster levels
Best For	Large streaming datasets	Small to medium datasets needing hierarchical analysis

4. Situations Where Suitable

- **Single Pass**:
  Large-scale real-time customer data grouping in e-commerce (e.g., grouping products by user behavior instantly).
- **Single Link**:
  Biological taxonomy (e.g., grouping species based on genetic similarity), where hierarchy and gradual merging are important.

---

5. Example Scenarios

- **Single Pass Example**:
  An online retail system groups users into segments (e.g., “frequent buyer”, “window shopper”) as new transactions arrive, without reprocessing old data.
- **Single Link Example**:
  Clustering different cities based on shortest road distances between them to build a connected travel map.

Q1) a) Analyse the concept behind the single pass algorithm and the single link algorithm for clustering. Compare and contrast these two techniques, highlighting their advantages, disadvantages, and situations where each approach would be more suitable. Provide a scenario for each algorithm to illustrate their application. **[8]**

Metric	Definition	Formula	Meaning
Precision	Measures how many of the predicted positive results are actually correct.	$\text{Precision} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Positives (FP)}}$	"Of all items the system labeled as relevant, how many were truly relevant?"
Recall	Measures how many of the actual positive results the system correctly predicted.	$\text{Recall} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Negatives (FN)}}$	"Of all truly relevant items, how many did the system find?"