University of Central Missouri

Department of Computer Science & Cybersecurity

CS5710 Machine Learning

Fall 2025

Home Assignment 4.

Student name: SANTHOSH REDDY KISTIPATI

Submission Requirements:

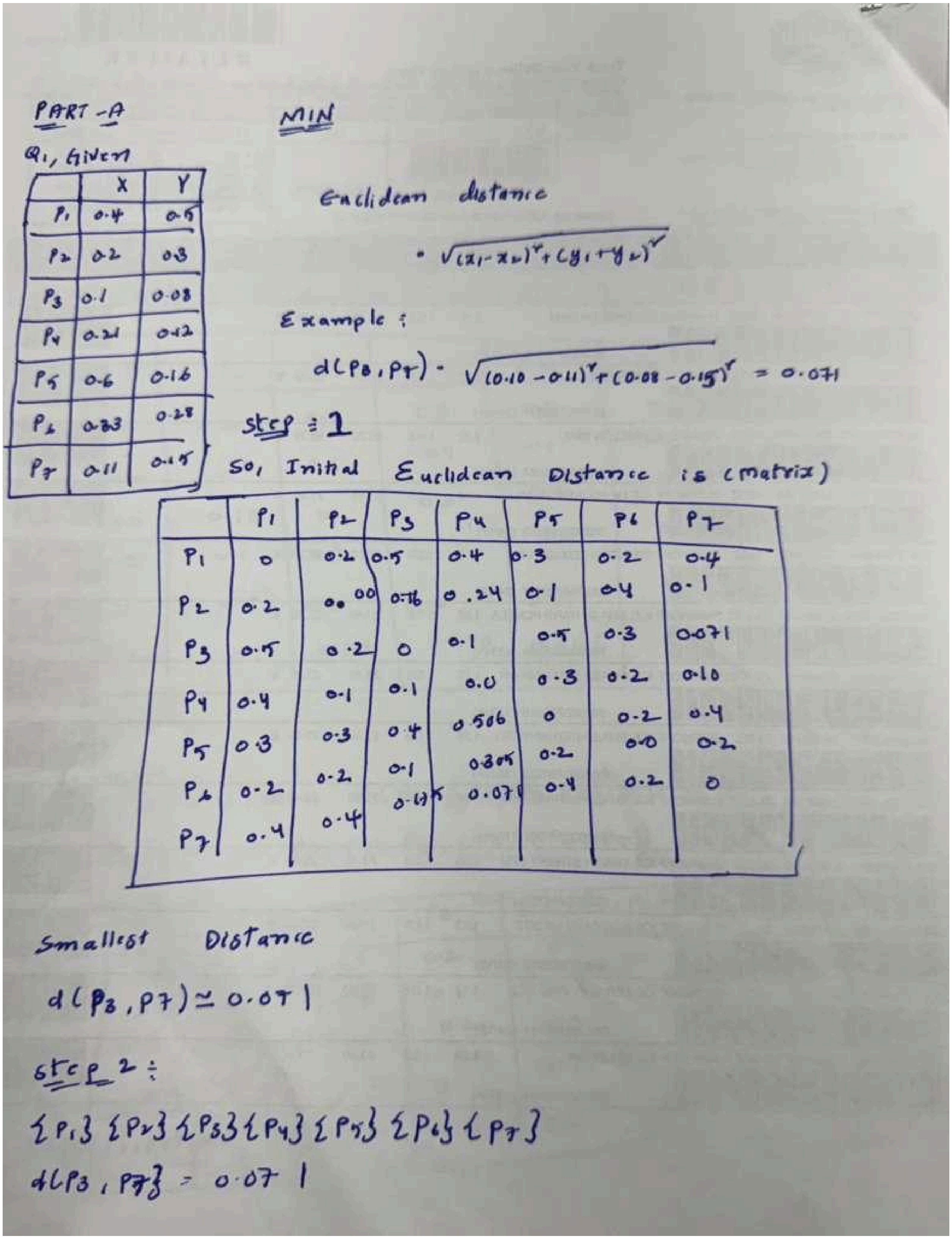
- Once finished your assignment push your source code to your repo (GitHub) and explain the work through the ReadMe file properly. Make sure you add your student info in the ReadMe file.
- Comment your code appropriately ***IMPORTANT.***
- Any submission after provided deadline is considered as a late submission.

Part A: Calculation

Q1. Find the cluster using the Average and MIN technique. Use Euclidean distance to build the complete distance matrix, update the distance matrix to the final step and draw the dendrogram for each.

	X	Y
P1	0.4	0.5
P2	0.2	0.3
P3	0.1	0.08
P4	0.21	0.12
P5	0.6	0.16
P6	0.33	0.28
P7	0.11	0.15

Ans :



PART - A

MIN

Q1, Given

	X	Y
P_1	0.4	0.5
P_2	0.2	0.3
P_3	0.1	0.08
P_4	0.21	0.12
P_5	0.6	0.16
P_6	0.23	0.28
P_7	0.11	0.15

Euclidean distance

$$= \sqrt{(x_1-x_2)^2+(y_1+y_2)^2}$$

Example :

$$d(P_3,P_7) = \sqrt{(0.10-0.11)^2+(0.08-0.15)^2} = 0.071$$

step : 1

So, Initial Euclidean Distance is (matrix)

	P_1	P_2	P_3	P_4	P_5	P_6	P_7
P_1	0	0.2	0.5	0.4	0.3	0.2	0.4
P_2	0.2	0.00	0.76	0.24	0.1	0.4	0.1
P_3	0.5	0.2	0	0.1	0.5	0.3	0.071
P_4	0.4	0.1	0.1	0.0	0.3	0.2	0.10
P_5	0.3	0.3	0.4	0.506	0	0.2	0.4
P_6	0.2	0.2	0.1	0.305	0.2	0.0	0.2
P_7	0.4	0.4	0.175	0.071	0.4	0.2	0

Smallest Distance

$d(P_3, P_7) \simeq 0.071$

step 2 :

$\{P_1\}\ \{P_2\}\ \{P_3\}\ \{P_4\}\ \{P_5\}\ \{P_6\}\ \{P_7\}$

$d(P_3, P_7\} = 0.071$

update matrix

$d(P_i, (8, P_7)) = mad(d(P_7, P_3), d(P_i, P_7))$

Ex :

$d(P_4, (P_3, P_7)) = min(d(P_4, P_3) = 0.117$

$d(P_2, (P_3, P_7)) = min(0.242, 0.175) = 0.175$

New cluster ÷ $P_1, P_2, P_4, P_5, P_6, (P_3, P_7)$

	P_1	P_2	P_4	P_5	P_6	(P_3, P_7)
P_1	0	0.2	0.4	0.3	0.2	0.4
P_2	0.2	0	0.1	0.4	0.1	0.175
P_4	0.42	0.18	0	0.3	0.2	0.10
P_5	0.3	0.4	0.3	0	0.2	0.490
P_6	0.2	0.1	0.2	0.2	0	0.2
(P_3, P_7)	0.4	0.1	0.104	0.490	0.256	0

step ÷ 3

sample

$(P_4, (P_3, P_7)) = 0.104$, $d(P_2, ((P_4, (P_3, P_7)) = 0.175$

New cluster : $P_1, P_2, P_5, P_6, (P_4, (P_3, P_7))$

Distance matrix

	P_1	P_2	P_5	P_6	$(P_4 (P_3, P_7))$
P_1	0	0.2	0.3	0.2	0.4
P_2	0.2	0	0.4	0.1	0.1
P_5	0.3	0.4	0	0.2	0.3
P_6	0.2	0.1	0.2	0	0.2
$(P_4, (P_3, P_7))$	0.4	0.1	0.3	0.2	00

step: 4

(P2, P6) = 0.132

clusters : {P1}, {P2}, {P5}, {P6}
{P4, (P3, P7)}

d{P1, (P2, P6)} min = 0.231

	P1	P5	(P4, (P3, P7))	(P2, P6)
P1	0	0.3	0.4	0.2
P5	0.3	0	0.3	0.2
(P4, (P3, P7))	0.425	0.3	0	~~0.17~~
(P2, P6)	0.2	0.2	0.1	0

step: 5

d(C P4, (P3, P7)), (P2, P6)) = 0.175 = C1

d(P1, C1) = 0.231

New cluster, P1, P5, ((P4, (P3, P7)), (P2, P6))

	P1	P5	C1
P1	0	0.3	0.2
P5	0.3	0	0.2
C1 = ((P4, (P3, P7)), (P2 , P6))	0.2	0.2	0

step: 6

d(C1, P1) = 0.231,

New cluster = (P1, ((P4, (P3, P7)), (P2, P6))) = C2

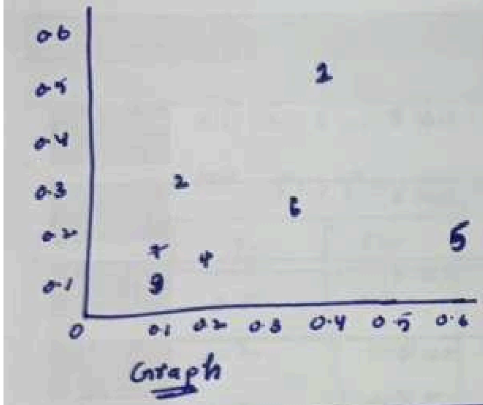
	P5	C4
P5	0.0	0.275
C4	0.295	0.0

step 7 (final)

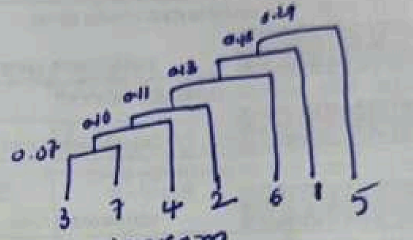
smallest distance = 0.295 between P5 and C2

(P5 , C2) = 0.295

C2 = (P1 , ((P4 , (P3 , P7)) , (P2 , P6))) .



Graph



Diagram

Average (method)

step: 1

clusters : {P1} {P2} {P3} {P4} {P5} {P6} {P7}

{P3, P7} = 0.071

merge : C1 = {P3, P7}

$C_1 = \left(\frac{0.10+0.11}{2} , \frac{0.08+0.15}{2} \right) = (0.105, 0.115)$

d(P4, C1) = 0.105

New cluster , P1, P2, P4, P5, P6 , (P3, P7)

	P1	P2	P4	P5	P6	(P3, P7)
P1	0	0.2	0.4	0.39	0.23	0.48
P2	0.2	0	0.18	0.42	0.13	0.28
P4	0.4	0.18	0.00	0.24	0.200	0.105
P5	0.39	0.424	0.282	0.00	0.245	0.427
P6	0.231	0.182	0.200	0.295	0.00	0.281
(P3, P7)	0.456	0.206	0.106	0.497	0.279	0.00

step 2

$(P_4, (P_3, P_7)) = 0.105$

$C_2 = (P_4, (P_3, P_7))$

$$C_2 = \left(\frac{0.21 + 0.10 + 0.11}{3}, \frac{0.12 + 0.08 + 0.15}{3} \right) = (0.14, 0.1167)$$

$d(X, C_2) = \| \text{centriod}(X) - \text{centroid}(C_2) \|$

- $d(P_2, C_2) \simeq (0.193$

New cluster $P_1, P_2, P_5, P_6, (P_4, (P_3, P_7))$

	P1	P2	P5	P6	(P4, (P3, P7))
P1	0.0	0.2	0.3	0.231	0.463
P2	0.2	0.0	0.4	0.1	0.193
P5	0.3	0.42	0.0	0.29	0.462
P6	0.2	0.132	0.285	0.0	0.251
(P4, (P3, P7))	0.4	0.193	0.482	0.251	0.0

step 3

$P_2, P_6) = 0.132$

$$C_3 = (P_2, P_6) = \left(\frac{0.20 + 0.33}{2}, \frac{0.30 + 0.28}{2} \right) = 0.265, 0.29$$

New cluster list : $P_1, P_5, (P_4, (P_3, P_7)), (P_2, P_6)$

	P1	P5	(P4 (P3, P7))	(P2, P6)
P1	0	0.3	0.4	0.25
P5	0.3	0	0.4	0.3
(P4 (P3, P7))	0.4	0.4	0.0	0.2
(P2, P6)	0.2	0.3	0.2	0.0

step 4 :

$(P4, (P3, P7)), (P2, P6) = 0\cdot214 = C4$

$C4 = (0\cdot95/5, 0\cdot93/5) = (0\cdot19, 0\cdot186)$

$d(P1, C4) = 0\cdot378$

	P1	P5	C4
P1	0·0	0·394	0·378
P5	0·394	0·0	0·411
C4	0·378	0·411	0·0

step 5

$(P1, C4) = 0\cdot378$

New cluster $C5 = (P1, C4)$

$C5 = (1\cdot35/6, 1\cdot43/6) \simeq (0\cdot225, 0\cdot2383$

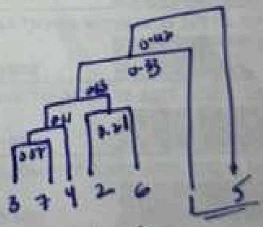
$d(P5, C5) \approx 0\cdot383$

Final

	P5	C5
P5	0.0	0·383
C5	0·383	0·0

All points are now in one cluster.

Diagram



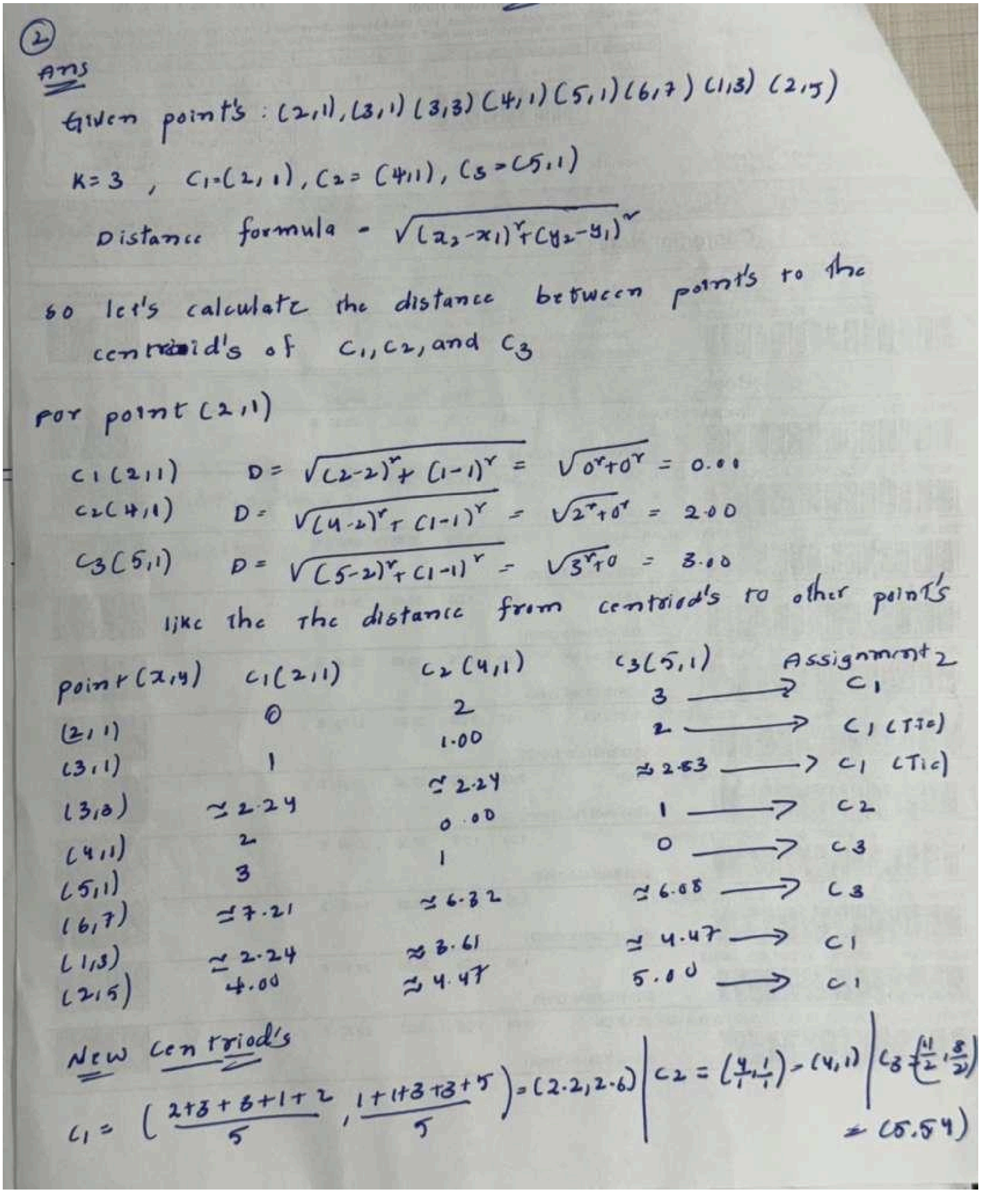
Q2. We have the following 2D data points:

Points: (2,1), (3,1), (3, 3), (4, 1), (5, 1), (6,7), (1,3), (2,5)
for K =3:
Centroid1: (2,1)
Centroid 2: (4, 1)
Centroid 3: (5, 1)

Euclidean Distance:

$$\text{Distance} = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

Ans :



②

Ans

Given point's : $(2,1), (3,1)\ (3,3)\ (4,1)\ (5,1)\ (6,7)\ (1,3)\ (2,5)$

$K=3$, $C_1=(2,1), C_2=(4,1), C_3=(5,1)$

Distance formula - $\sqrt{(x_2-x_1)^2+(y_2-y_1)^2}$

so let's calculate the distance between point's to the centroid's of C_1, C_2, and C_3

For point (2,1)

$C_1(2,1)$ $D=\sqrt{(2-2)^2+(1-1)^2} = \sqrt{0^2+0^2} = 0.00$

$C_2(4,1)$ $D=\sqrt{(4-2)^2+(1-1)^2} = \sqrt{2^2+0^2} = 2.00$

$C_3(5,1)$ $D=\sqrt{(5-2)^2+(1-1)^2} = \sqrt{3^2+0} = 3.00$

like the the distance from centriod's to other point's

point (x,y)	$C_1(2,1)$	$C_2(4,1)$	$C_3(5,1)$	Assignmnt 2
(2,1)	0	2	3 $\longrightarrow$	C_1
(3,1)	1	1.00	2 $\longrightarrow$	C_1 (Tie)
(3,3)	$\simeq$ 2.24	$\simeq$ 2.24	$\simeq$ 2.83 $\longrightarrow$	C_1 (Tie)
(4,1)	2	0.00	1 $\longrightarrow$	C_2
(5,1)	3	1	0 $\longrightarrow$	C_3
(6,7)	$\simeq$ 7.21	$\simeq$ 6.32	$\simeq$ 6.08 $\longrightarrow$	C_3
(1,3)	$\simeq$ 2.24	$\simeq$ 3.61	$\simeq$ 4.47 $\longrightarrow$	C_1
(2,5)	4.00	$\simeq$ 4.47	5.00 $\longrightarrow$	C_1

New centriod's

$C_1 = \left(\frac{2+3+3+1+2}{5}, \frac{1+1+3+3+5}{5}\right) = (2.2, 2.6)$ | $C_2 = \left(\frac{4}{1}, \frac{1}{1}\right) = (4,1)$ | $C_3 = \left(\frac{11}{2}, \frac{8}{2}\right) = (5.5, 4)$

Part B — Short-Answer

Q1.

a) Describe agglomerative hierarchical clustering.

Ans :Agglomerative hierarchical clustering is a "bottom-up" approach to clustering. It starts by treating every single data point as its own individual cluster. It then iteratively merges the two clusters that are the most similar or nearest to each other based on a defined distance function. This process continues until all data points are merged into a single, large cluster, forming the root of the hierarchical tree

b) Describe divisive hierarchical clustering.

Ans : Divisive hierarchical clustering is a "top-down" approach. It begins with all data points grouped together in one cluster . In each step, it splits the cluster into a set of child clusters. This splitting procedure is applied recursively until only singleton clusters of individual data points remain.

c) Which one is more commonly used and why?

Ans :Agglomerative clustering is more popular than divisive methods. The reason is generally related to computational and conceptual simplicity. In agglomerative clustering, the primary task is simply to find the minimum distance between existing clusters and merge them. Divisive clustering requires a more complex method to determine the optimal way to split a cluster to ensure the resulting sub-clusters are coherent, which can be computationally intensive

Q2.

a) To improve clustering quality, should inter-cluster distance be maximized or minimized?

Ans : The inter-cluster distance is the measure of separation between different clusters. By maximizing this distance, you ensure that the clusters are as far apart as possible, meaning the data instances in one cluster are very different ("far away") from the data instances in other clusters. This is essential for good separation and a clear distinction between the identified groups.

b) Same question for intra-cluster distance — explain the reasoning.

Ans : The intra-cluster distance is the measure of cohesion, or tightness, within a single cluster. It measures how similar ("near") the data instances are to each other inside the same cluster. By minimizing this distance, you ensure that all points belonging to a single cluster are very similar and tightly grouped, leading to high internal homogeneity.

Q3.

a) Define single link, complete link, and average link.

Ans :

Single Link (MIN): The distance between two clusters is defined by the minimum distance between the closest pair of points, where one point is in the first cluster and the other is in the second

Complete Link (MAX): The distance between two clusters is defined by the maximum distance between the farthest pair of points, where one point is in the first cluster and the other is in the second

Average Link: The distance between two clusters is defined as the average distance between all pairs of points across the two clusters. This is typically calculated as the distance between the cluster centroids

b) Explain one strength and one weakness of single-link clustering.

Ans : Strength:

- Can discover non-elliptical shapes (such as elongated or curved clusters). Because it only looks for the nearest neighbor between two clusters, it can "chain" points together to form clusters of arbitrary shapes.

Weakness:

- Susceptible to Noise (Chaining Effect): Single-link clustering is highly sensitive to noise or outliers and suffers from the "chaining effect". A single point lying between two otherwise well-separated clusters can act as a bridge, forcing the two distinct clusters to merge prematurely at a very small distance

Q4.

a) What is the role of tokenization and give one example.

Ans 👍The primary role of tokenization is to break down a sequence of text (like a document, paragraph, or sentence) into smaller, meaningful units called tokens. These tokens are the fundamental building blocks used for further processing in natural language processing (NLP) tasks.

Role: To transform unstructured text data into a format that can be analyzed by a machine learning model.

Example:

Original Text: "Clustering is a great technique!"

Tokens (Word Tokenization):
"Clustering", "is", "a", "great", "technique", "!"

b) Compare stemming vs. lemmatization in terms of speed and accuracy.

Ans:

Feature	**Stemming**	**Lemmatization**
Speed	**Faster**	**Slower**
Accuracy	**Lower**	**Higher**
Reasoning	Uses a **crude heuristic process** (set of simple rules) to chop off word endings, without checking if the resulting root is a real word.	Uses a **linguistic process** (vocabulary/dictionary and morphological analysis) to correctly identify the dictionary form of the word (the lemma).

Q5.

a) Explain what word sense ambiguity is and provide an example.

Ans 👍

a) Word Sense Ambiguity

Word sense ambiguity occurs when a single word can have multiple distinct meanings or interpretations, and the correct meaning can only be determined by its context within a sentence.

Explanation: A word is polysemous (has multiple related meanings) or homonymous (has different, unrelated meanings but the same spelling). Without sufficient context, a computational model cannot confidently determine which "sense" of the word is intended.

Example: The word **"bank"**

Sense 1: "I need to go to the **bank** to deposit a check." (Financial institution)

Sense 2: "We saw the deer drinking water near the river **bank**."

b) Explain why pronoun reference ambiguity can confuse a model.

Ans : **Pronoun reference ambiguity** (or anaphora resolution) confuses a model because the pronoun's antecedent (the noun it refers back to) is not explicitly clear. Models struggle when a pronoun, such as **"they"** or **"it"**, could logically refer to more than one noun in the preceding text. For example, in the sentence, "The old city bus passed the tree because it was blocking the street," the pronoun **"it"** could refer to either the **"bus"** or the **"tree."** Resolving this requires the model to apply complex reasoning and world knowledge, which is a significant challenge for purely syntactic or pattern-based models.

Q6.

a) Why can't NLP tasks like POS tagging be solved by predicting each token independently?

Ans : NLP tasks like **Part-of-Speech (POS) tagging** cannot be solved by predicting each token independently because the grammatical role of a word is often **mutually dependent** on the words immediately preceding or following it in a sentence. A word may have multiple possible parts of speech (e.g., "drive" can be a noun or a verb), and the correct tag depends entirely on the surrounding context. Predicting each token in isolation would ignore these dependencies, leading to syntactic errors and a low-accuracy analysis of the sentence structure. Models must utilize information about adjacent tags to maintain **sequence coherence**.

b) Give one example where decisions are mutually dependent in a sentence.

A common example where tagging decisions are mutually dependent involves tagging sequences that contain **homographs** (words spelled the same but having different tags).

Sentence: "Can you **can** the tuna?"

Token	**Possible Tags**	**Correct Tag (Dependent Decision)**
can	Noun (The physical container) / **Modal Verb** (Auxiliary)	**Modal Verb** (Followed by another verb, "can")
can	**Noun** (The physical container) / Verb (To put in a container)	**Verb** (Preceded by a modal verb, "can")
tuna	**Noun** (Object of the verb "can")	**Noun**

The decision to tag the first "**can**" as a **Modal Verb** forces the second "**can**" to be tagged as a **Main Verb**, illustrating their mutual dependence. If the first "can" was tagged independently as a Noun, the rest of the sentence structure would break grammatically.

Part C — Coding

Q1.

Write Python code to perform the following steps:

1. Segment into tokens
2. Remove stopwords

3. Apply lemmatization (not stemming)
4. Keep only verbs and nouns (use POS tags)

Input text:

```
"John enjoys playing football while Mary loves reading books in the library."
```

Q2.

Use Python and any NLP model to perform:

1. Named Entity Recognition (NER)
2. Disambiguation prompt**:**
   If the text contains a pronoun ("he", "she", "they"), print:

   "Warning: Possible pronoun ambiguity detected!"

Input text:

```
"Chris met Alex at Apple headquarters in California. He told him about the new
iPhone launch."
```