Homework - 1

1) vectors $x = (1, 1, 0, 1, 0, 1)$
$y = (1, 1, 1, 0, 0, 1)$

a) Manhattan distance $\div$ $mod(x_1 - x_2)$

$$\Rightarrow |1-1| + |1-1| + |0-1| + |1-0| + |0-0| + |1-1|$$

$$= |0| + |0| + |-1| + |1| + |0| + |0|$$

$$= 0 + 0 + 1 + 1 + 0 + 0$$

$$= 2$$

b) Euclidean distance $\div$ $\sqrt{(x_2-x_1)^2 + (y_2-y_1)^2}$

$$Ed = \sqrt{(1-1)^2 + (1-1)^2 + (0-1)^2 + (1-0)^2 + (0-0)^2 + (1-1)^2}$$

$$= \sqrt{0^2 + 0^2 + (-1)^2 + (1)^2 + 0^2 + 0^2}$$

$$= \sqrt{1^2 + 1^2} = \sqrt{2}$$

c) Supremum distance : $\max(|x_i - x_j|)$

$= \max(|1-1|, |1-1|, |0-1|, |1-0|, |0-0|, |1-1|)$

$= \max(0, 0, 1, 1, 0, 0)$

$= 1$

d) Cosine distance = 1 - Cosine similarity

$$\text{Cosine similarity} = \frac{x \cdot y}{\|x\| \|y\|}$$

$x \cdot y = 1*1 + 1*1 + 0*1 + 1*0 + 0*0 + 1*1$

$x \cdot y = 1 + 1 + 0 + 0 + 0 + 1$

$x \cdot y = 3$

$\|x\| = \sqrt{1^2 + 1^2 + 0^2 + 1^2 + 0^2 + 1^2} = \sqrt{4} = 2$

$\|y\| = \sqrt{1^2 + 1^2 + 1^2 + 0^2 + 0^2 + 1^2} = \sqrt{4} = 2$

Cosine similarity $= \frac{3}{4}$

Cosine distance $= 1 - \frac{3}{4} = 1/4 = 0.25$ //

20)

element	S_1	S_2	S_3	S_4	$2x+3$	$3x+2$	$4x+1$
0	0	1	0	1	3	2	1
1	0	1	0	0	0	0	0
2	1	0	0	1	2	3	4
3	0	0	1	0	4	1	3
4	0	0	1	1	1	4	2

Step 1 :-

	S_1	S_2	S_3	S_4
h_1	∞	∞	∞	∞
h_2	∞	∞	∞	∞
h_3	∞	∞	∞	∞

Step 2 :-

	S_1	S_2	S_3	S_4
h_1	∞	3	∞	3
h_2	∞	2	∞	2
h_3	∞	1	∞	1

Step 3 :-

	S_1	S_2	S_3	S_4
h_1	∞	0	∞	3
h_2	∞	0	∞	2
h_3	∞	0	∞	1

Step 4 :-

	S_1	S_2	S_3	S_4
h_1	2	0	∞	2
h_2	3	0	∞	2
h_3	4	0	∞	1

Step 5 :-

	S_1	S_2	S_3	S_4
h_1	2	0	4	2
h_2	3	0	1	2
h_3	4	0	3	1

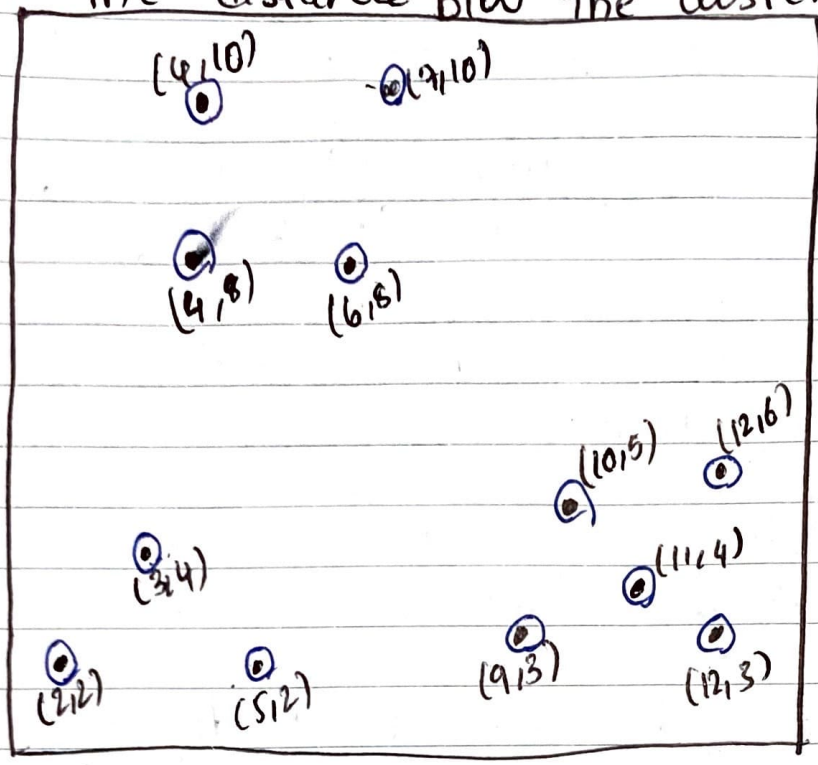
Step 6 :-

	S_1	S_2	S_3	S_4
h_1	2	0	1	1
h_2	3	0	1	2
h_3	4	0	2	1

Pair	True Jaccards similarity	estimated
S_1, S_2	0	0 ✓
S_1, S_3	0	0 ✓
S_1, S_4	1/3	0 ✓
S_2, S_3	0	0
S_2, S_4	1/4	0
S_3, S_4	1/4	1/3

* At first each point is its own cluster
Find the distance blw the cluster using centroid.

30)

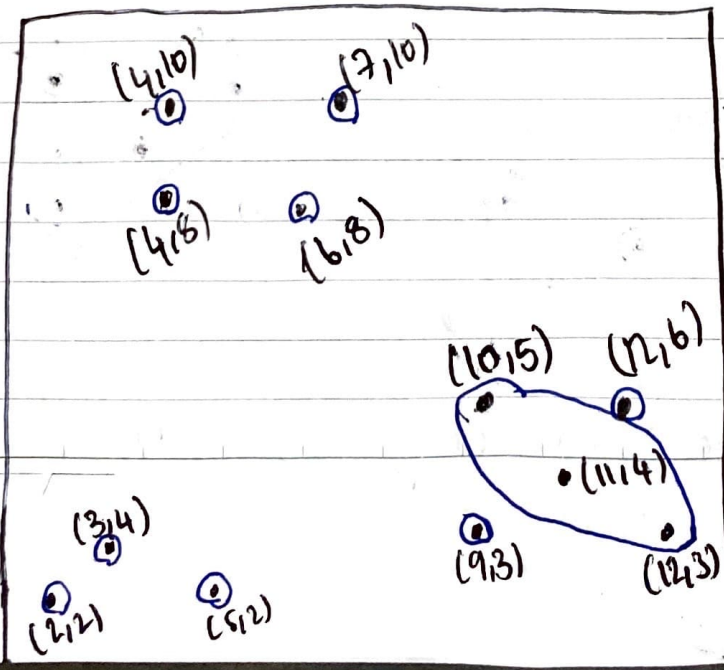


There are 13 clusters

consider the first iteration we can see that the nearest points (or) the distance between (11,4) (10,5) = $\sqrt{1^2+1^2}$ = $\sqrt{2}$ = 1.41

(11,4) (12,3) = $\sqrt{1^2+1^2} = \sqrt{2}$ = 1.41

Hence combine (10,5) (11,4) (12,3) as cluster & consider it as 'A'



Here there are 10 clusters

The centroid of the cluster is Ps (11,4)

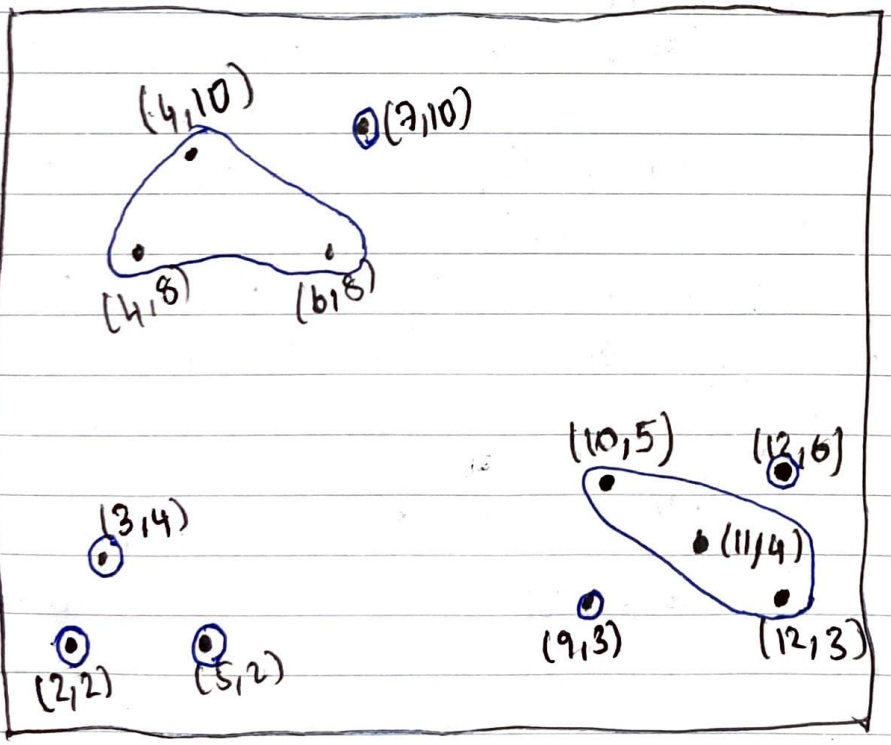
distance blw $\left.\begin{matrix}(11,4)-(9,3)\\(11,4)-(12,6)\end{matrix}\right\} = 2.24$

But we can see here that the distance blw

$(4,10)\ (4,8) = \sqrt{0^2+2^2} = 2$

$(4,8)\ (6,8) = \sqrt{2^2+0^2} = 2$

These are the two closest distances found Hence make (4,10), (4,8), (6,8) as a cluster.



Here there are only 8 cluster's

Consider the cluster (4,8) (6,8) (4,10) as B
Find the centroid cluster B and that is $\left(\frac{4+6+4}{3}, \frac{8+8+10}{3}\right) = (4.66, 8.66)$

Now we have to find the distance blw the centroid of B & (7,10)

$(4.66, 8.66), (7,10) = \sqrt{(7-4.66)^2 + (10-8.66)^2}$

$= 2.69$

But this greater than the distance blw centroid A & (9,3), (12,6) => 2.24

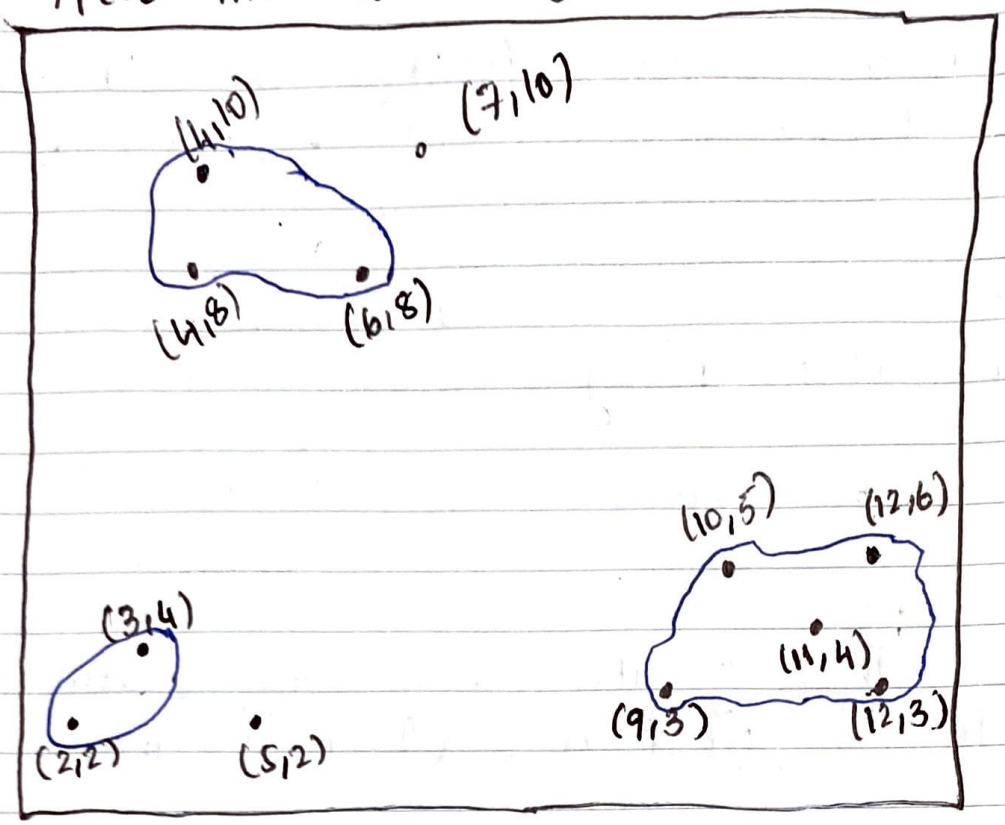
calculate distance blw (2,2) (3,4)
(2,2) (5,2)
(3,4) (5,2)

$(2,2), (3,4) = \sqrt{1^2 + 2^2} = 2.24$
$(2,2), (5,2) = \sqrt{3^2 + 0} = 3$
$(3,4), (5,2) = \sqrt{2^2 + 2^2} = \sqrt{8} = 2.83$

Now we can see that there 2 smallest distances that are ~~2.24~~ between
(2,2), (3,4)
(11,4) (9,3) & (11,4), (12,6) } 2.24

∴ cluster Both these (2,2), (3,4) & Add (9,3) & (12,6) to the cluster 'A'

Here there are only 5 clusters



Let the cluster (2,2) & (3,4) be 'C'

we know the centroid of A, B => (4.66, 8.66)

The centroid of A is $\left(\frac{10+11+9+12+12}{5}, \frac{5+6+4+3+3}{5}\right)$

$= (10.8, 4.2)$

The centroid of C is $\left(\frac{5}{2}, \frac{6}{2}\right)$

$(2.5, 3)$

Now calculate the distance blw

$(2.5, 3), (5, 2) = \sqrt{6.25 + 1}$

$= \sqrt{7.5} = 2.73$

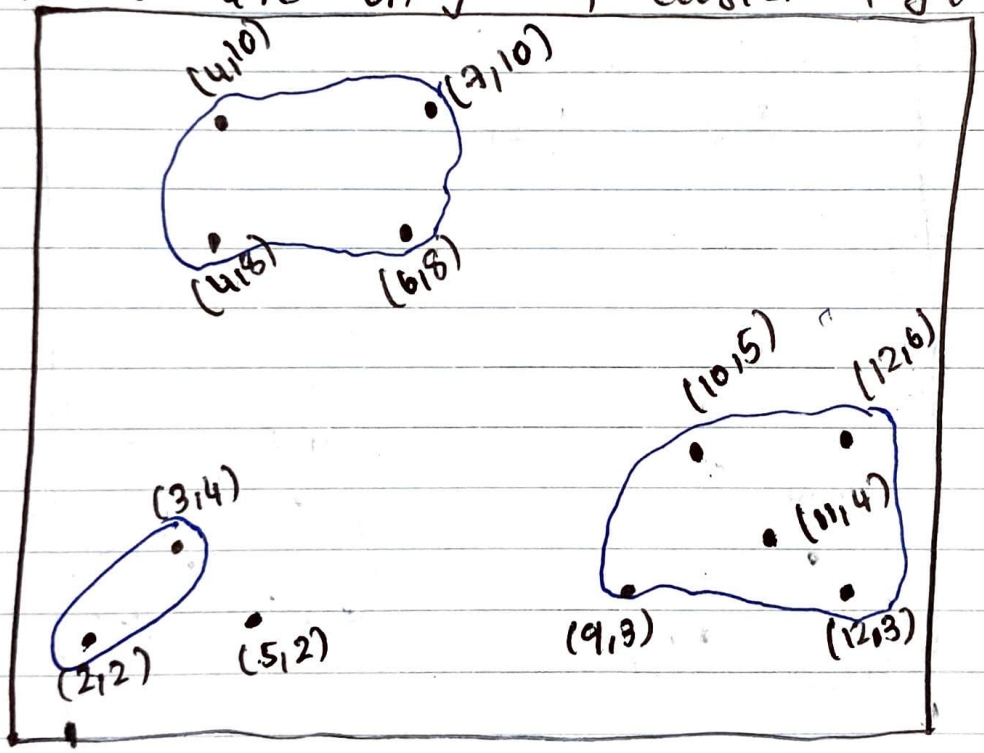
The distance b/w (4·66, 8·66) (7,10) is

= 2·69

Here we can see that

distance between cluster B & (7,10) is near Hence, add (7,10) with the cluster 'B'

There are only 4 cluster's left



The centroid of A is (10.8, 4.2)

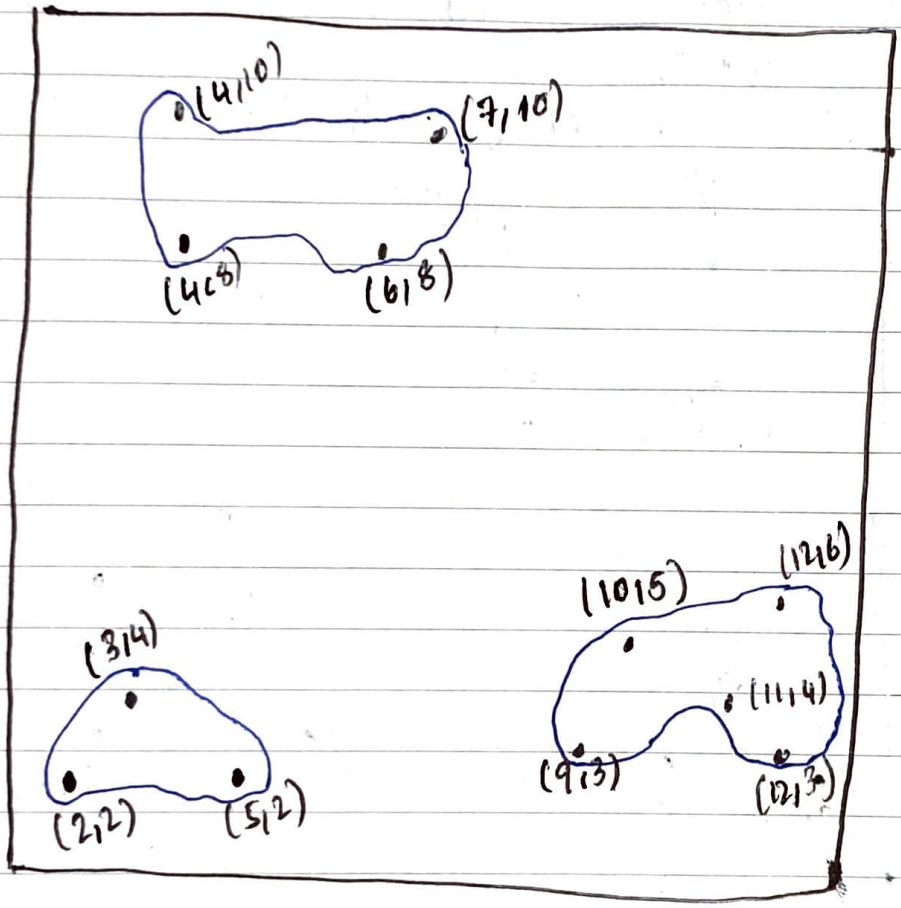
The centroid of B is $\left(\frac{4+4+7+6}{4}, \frac{10+10+8+8}{4}\right)$

$= (5.25, 9)$

Now calculate the distance b/w the centroid of C & (5,2)

$(2.5,3)\ (5,2) = \sqrt{7.5} = 2.73$

Now Add (5,2) to the cluster 'C'



This is the 5th iteration & there are 3 clusters A, B, C

40) The first 5-ten shingles

* Considering white_spaces as underscore the shingles must be set i.e there should not be any shingles that are duplicate
* The given sentences can be written as

" The_most_effective_way_to_represent_documents_as_sets_for_the_purpose_of_identifying_lexically-similar_documents_is_to_construct_from_the_document_the_set_of_short_strings_that_appear_within_it. "

The Shingles are

1) " The_m "
2) " he_mo "
3) " e_mos "
4) " _most "
5) " most_ "
6) " ost_e "
7) " st_ef "
8) " t_eff "
9) " _effe "
10) " effec "

The main usage of shingling is to convert document into sets.