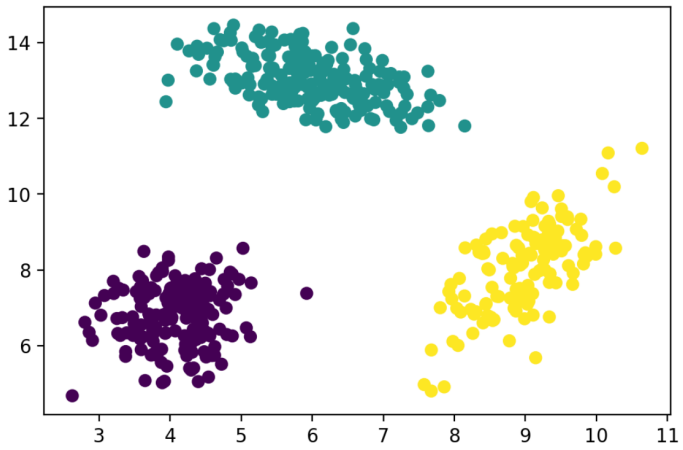
Q1

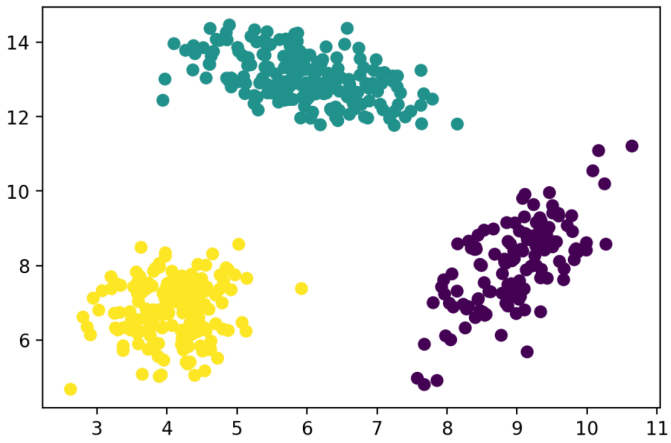
The updated centers and SSE for first two iteration and final are listed in the table below. Note that for computing the SSE, p = 1.

Iteration	Centers	SSE
1	C1 = [8.96785131, 8.71751639] C2 = [5.85290351, 13.04683287] C3 = [4.43499606, 7.18310891]	1040.7497
2	C1 = [8.86028426, 8.60252036] C2 = [5.89817085, 12.94199237] C3 = [4.34250289, 7.15303321]	1016.4344
…		
Final	C1 = [8.67125416, 8.36916046] C2 = [6.02665596, 12.78163089] C3 = [4.24652866, 7.12817881]	1016.0226

Clustering result:



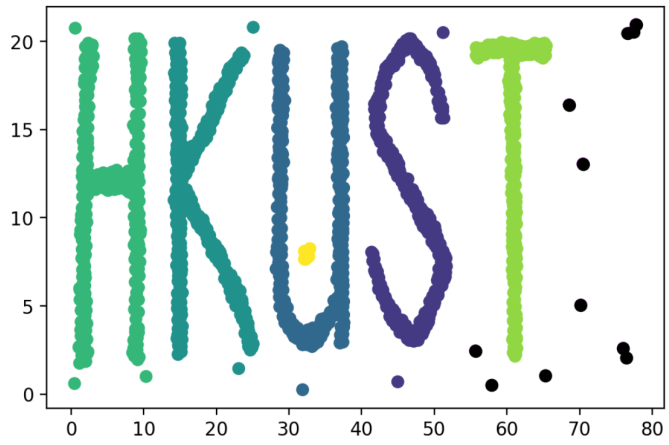
Scatter plot using original label:



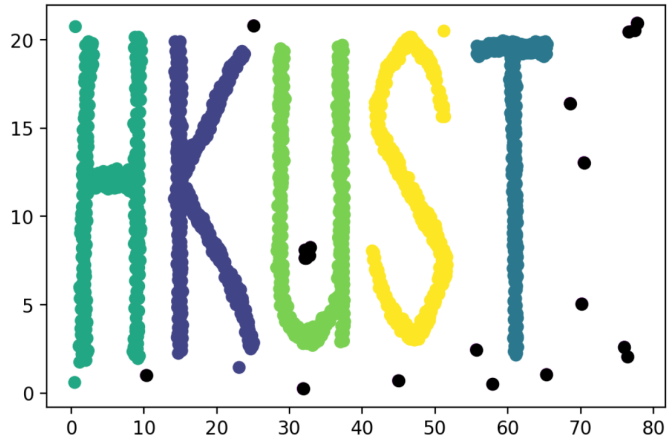
It is clear that the EM Algorithm has the same cluster results as the original label, which means the accuracy is 100%.

Q2

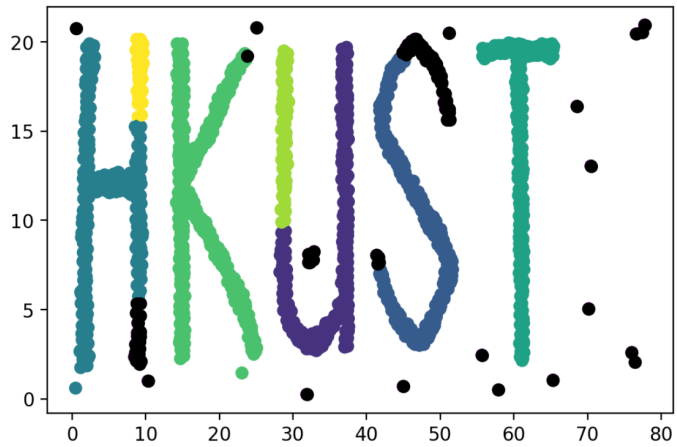
ϵ = 3, Minpoints = 5:



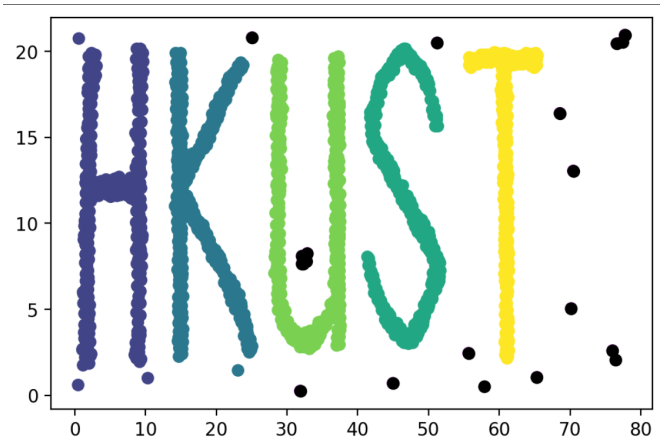
ϵ = 2.5, Minpoints = 20:



ϵ = 2, Minpoints = 20:



ϵ = 2, Minpoints = 15:



Parameters	# of clusters	# of outliers
ϵ = 3, Minpoints = 5	6	11
ϵ = 2.5, Minpoints = 20	5	19
ϵ = 2, Minpoints = 20	7	30
ϵ = 2, Minpoints = 15	5	18

The slightly change of the combination of parameters has huge impact of the final cluster results. When Minpoints increase from 5 to 20, the number of core points & boarder points will decrease, and more points are considered as outliers. While keep Minpoints unchanged and decrease ϵ, # of clusters increase (such as 'H' and 'U' are divided into two parts). In the other hand, the data are more sensitive to ϵ rather than Minpoints.

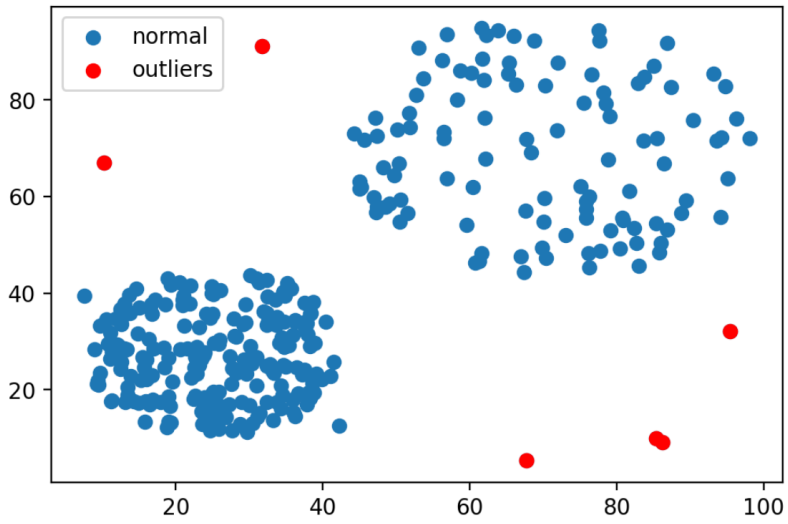
I think ϵ = 2, Minpoints = 15 is the best setting. Since the 'HKUST" can be precisely clustered into five groups, and most of points do not belong to these five characters (includes those four points in the 'U') are detected as outliers. If we try increasing the Minpoint or decreasing ϵ, probably # of clusters will increase.

Q3

Parameters:

Radius r = 10, fraction threshold = 0.01

Coordinates of outliers
[95.41160594, 32.02898551]
[31.71390013, 91.15942029]
[86.23481781, 9.13043478]
[85.42510121, 10]
[10.25641026, 66.95652174]
[67.7462888, 5.36231884]


normal
outliers
80
60
40
20
20
40
60
80
100