Can BEYAZNAR
161044038

CSE454 DATA MINING HW01

1) Dataset

The name of the dataset I'm working on is "Search growth for Data Science terms". This dataset brings together the top area-related words and shows how much google they search weekly over the last 5 years. Numerical data it contains:

"week", "analytics", "api", "artificial intelligence", "big data", "clustering", "data mining", "data science", "data scientist", "data warehouse" , "deep learning", "etl", "excel", "github", "hadoop", "iot", "java", "machine learning", "matlab", "minitab", "modeling", "python" , "R", "regression", "sql", "statistician"

I tested my dbscan algorithm using data science and data mining data from this dataset.

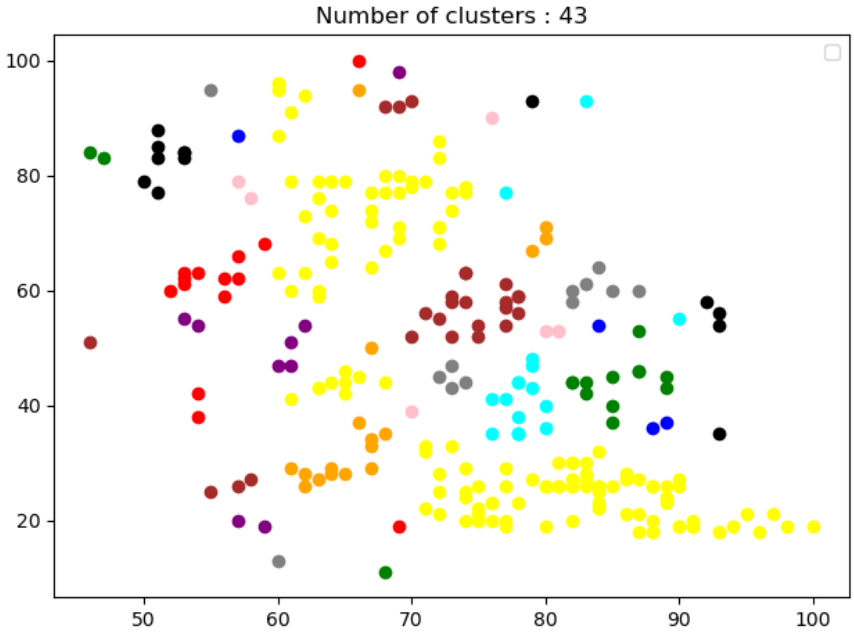
Link : https://www.kaggle.com/leonardopena/search-growth-for-data-science-terms

1) Result

In my resulting graphs, I changed my epsilon value 3 times and my min points value 3 times. My results are as follows. NOTE: If there are too many sets, the colors of the different sets in the chart may be the same.

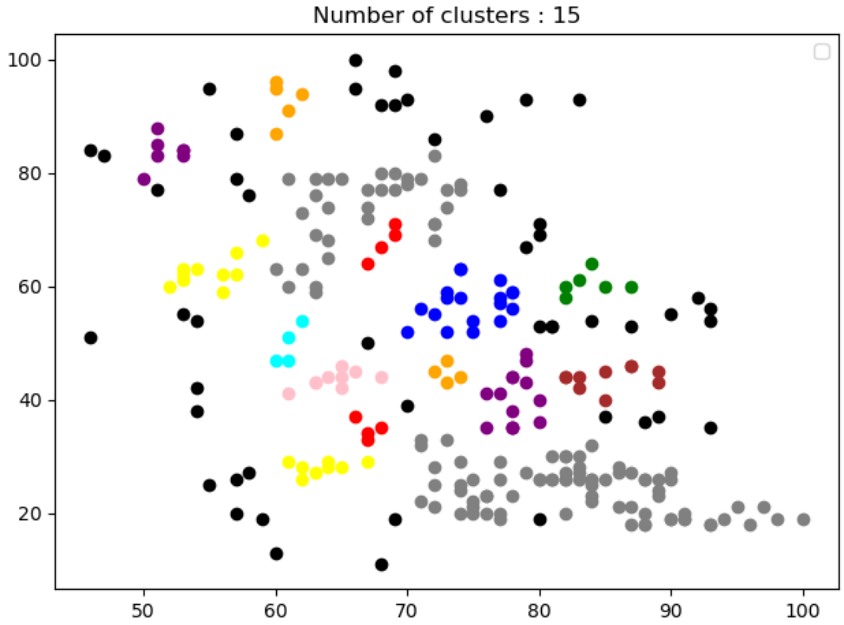
2-a) `dbscan(X, Epsilon=0.2, Min_points=1)`

A total of 43 clusters are formed in the graph. The main reason for this is because we make min points 1. Since the value of min points is 1, each data point visited is considered as a set. The clusters can expand according to the Epsilon value. There is no noise at all as each object is assumed as a cluster.



2-b) `dbscan(X, Epsilon=0.2, Min_points=4)`

A total of 15 clusters are observed in the chart. Objects that were close to each other turned into larger clusters due to the epsilon value. We observe that as the min points increase, the number of clusters decreases. Here we see that the noisy objects are more than the previous values.

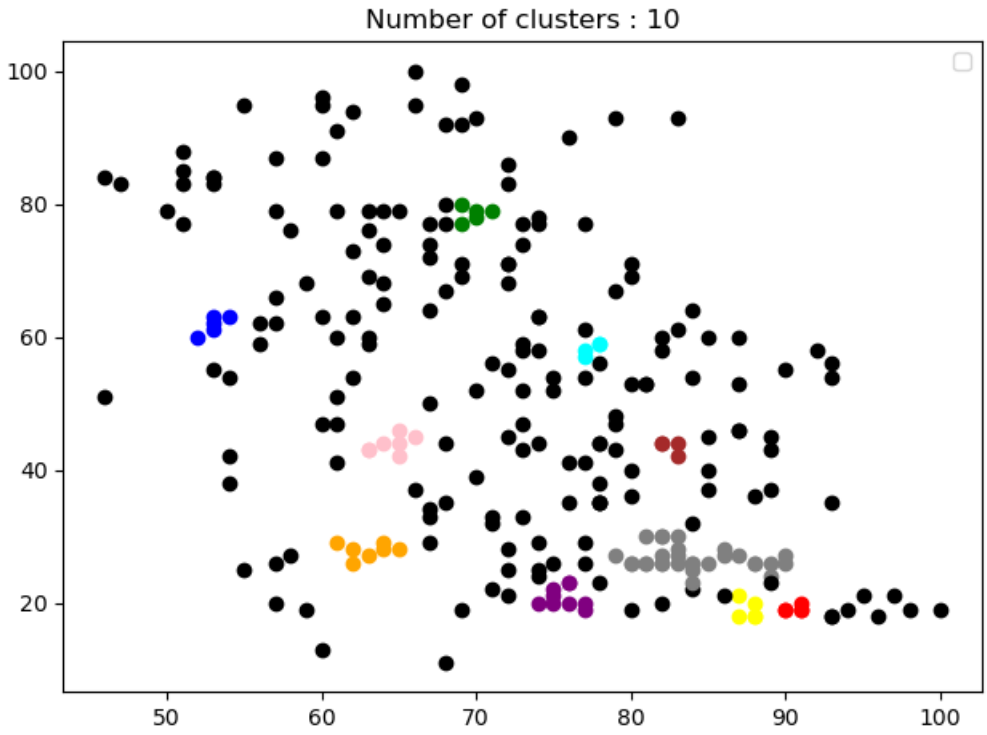


2-c) `dbscan(X, Epsilon=0.2, Min_points=9)`

A total of 2 clusters are observed in the chart. The reason for this is because we give a very high value to the min points parameter. In the clusters taken according to epsilone among objects, the number of clusters was low because the number of objects did not exceed the min points threshold. Here we see a lot of noisy objects.

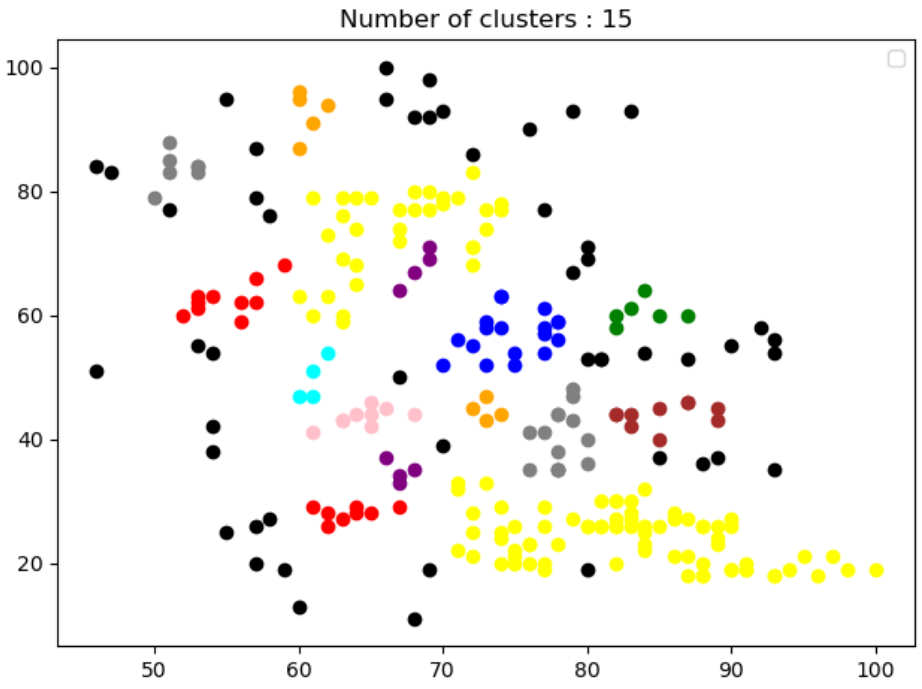
2-d) dbscan(X, Epsilon=0.1, Min_points=4)

After changing the min points value, we change the epsilon value. First, our epsilon value is 0.1. When you give the values in this way, 10 clusters are formed in total. The number of large clusters is almost non-existent in this chart. The reason for this is that the number of objects in epsilon proximity to each other is very low. We can only accept a set of gray objects as a large set. Also, here we see a lot of noisy objects.



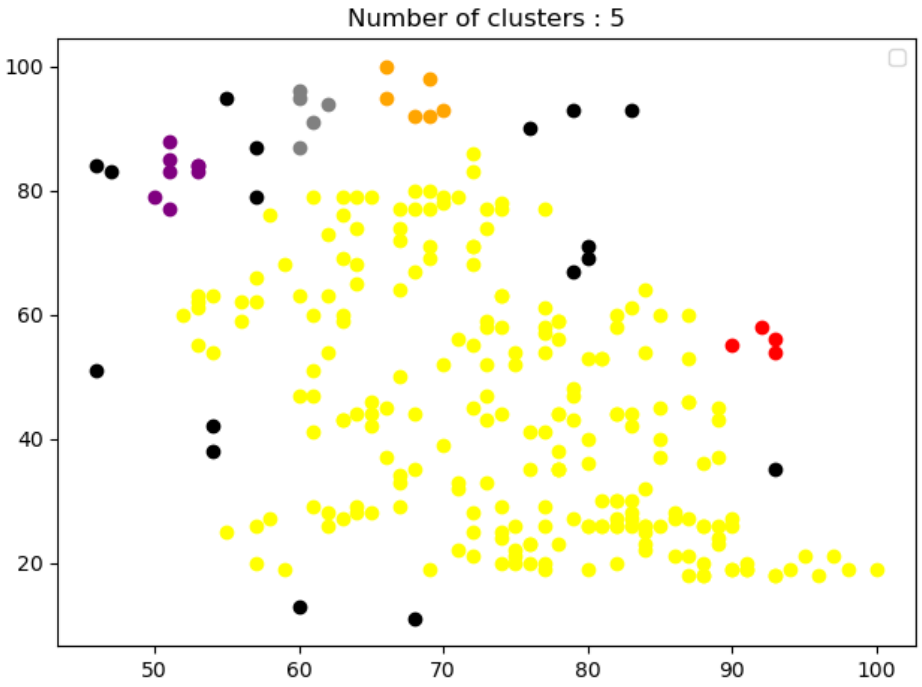
2-e) dbscan(X, Epsilon=0.2, Min_points=4)

Here we increased the epsilon value and an increase was observed in the number of clusters. Based on this, we can think that the distances between the objects we have are large. However, if we increase the epsilon value in the following steps, the number of clusters may decrease. In this graph, we see that the number of noisy objects is low.



2-f) dbscan(X, Epsilon=0.3, Min_points=4)

Here we increased the epsilon value and a decrease in the number of clusters was observed. The reason for this is that we can think that the number of clusters decreases with the merger of more than one cluster. As clusters expand, the number of clusters is decreasing. Clusters contain more objects because of the greater distance between objects. And naturally our cluster number is decreasing. And we can easily see that the number of noisy objects has decreased.



3) In the report, give a technique to automatically decide on the parameters of DB-Dcan?

I searched for methods to automatically decide the Epsilon and Min points parameters. I have read a few articles for these methods. Choosing the min points parameter generally depends on the data sets used. Therefore, if the data set used has normal distributions, Min points are chosen empirically. If the data set does not have normal distributions, the Min points parameter can be selected using the Chebyshev inequality. For the Eps parameter, the main purpose is to accurately determine the sharp increments of distances. Therefore, first of all, knee-row distances must be specified precisely. Then the point corresponding to sharp increases in distances is determined. Depending on this point and the knee size, the correct value of the eps parameter can be calculated. Apart from these, DBSCAN proposes a new hybrid approach consisting of Binary Differential Evolution (BDE) and DBSCAN clustering algorithm as BDE DBSCAN to quickly and automatically select very suitable Eps and MinPts parameters. BDE DBSCAN performance is evaluated using various datasets with different densities and shapes. In addition, BDE-DBSCAN outperforms other algorithms such as Binary Harmony Search (BHS), Binary Bees (BBees) algorithm, Tournament Selection Binary Genetic Algorithm (GA-TS) and Roulette Wheel (GA-RW) and Binary Particle Swarm Optimization. As a result of the articles I have read, I have determined that these methods are suitable for determining the Min points and epsilon parameters appropriately. Of course, there are different solutions other than these methods. I think the data set is important for selecting appropriate values for the parameters. Epsilon and min points parameters may vary depending on the status of the data set.