CS 5525: Data Analytics

Project 2

Ankit Parekh
MS (Non-Thesis) Computer Engineering
ECE Department
ankitparekh@vt.edu

Problem 1 [17 points]

1. (4 points points) On the movie ratings dataset, k-means clustering assign users to two clusters: cluster 0 has users with more affinity for horror movies, and cluster 1 has users with more affinity for action movies. Given the cluster centroids, assign the following users to their respective cluster assignment:

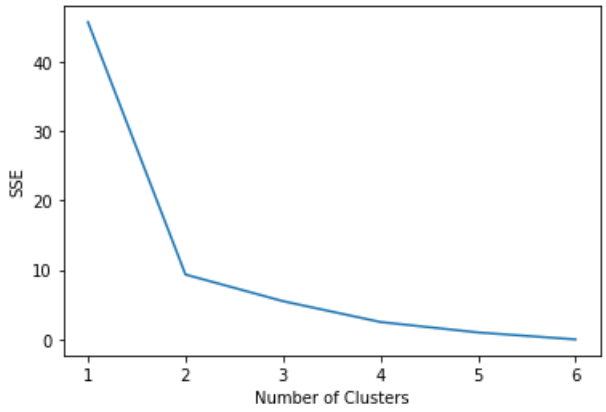
User	Exorcist	Omen	Star Wars	Jaws
Paul	4	5	2	4
Adel	1	2	3	4
Kevin	2	3	5	5
Jessi	1	1	3	2

A) The assigned clusters for the above users are mentioned in the below table:

	user	Jaws	Star Wars	Exorcist	Omen	Cluster ID
0	Paul	4	2	4	5	0
1	Adel	4	3	1	2	1
2	Kevin	5	5	2	3	1
3	Jessi	2	3	1	1	1

2. (2 points points) To determine the optimal value of K in K-means, a common approach is to use the Elbow Method, where the idea is to find a K value that shows the sharpest change in slope of the SSE curve. For the movie rating dataset, what value of K would you arrive at by applying the Elbow Method visually? Briefly explain your reasoning.

A) The plot of SSE vs K looks like an elbow with decreasing SSE with an increase in value of k. From a particular value of K, the graph starts to move almost parallel to X-axis i.e increase in K does not lead to a major decrease in SSE. This point is the optimal K value or the optimal number of Clusters.



For the given data, the sharpest change is at K=2 and hence the optimal number of clusters is 2. (Graph can be seen above). It can also be seen from the data that the ratings of the user are more inclined toward one of the categories of the movies and hence a cluster of horror and action is enough to
classify the data.

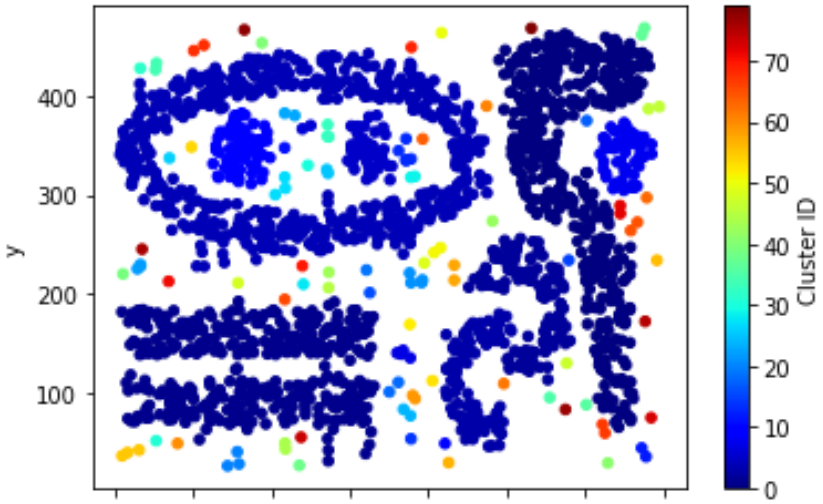
3. (4 points points) On the Vertebrate dataset, we illustrate the results of using three hierarchical clustering algorithms (1) single link (MIN), (2) complete link (MAX), and (3) group average. Given the class label in the original dataset, compute the cophenetic correlation coefficient of the clustering produced by each algorithm. Which clustering algorithm shows the best match with the class labels?

A) Cophenetic correlation coefficient of Single Link (MIN) = 0.356
Cophenetic correlation coefficient of Complete Link (MAX) = 0.606
Cophenetic correlation coefficient of Group Average = 0.489
Complete Link (MAX) algorithm shows the best match with the class labels with highest coefficient.

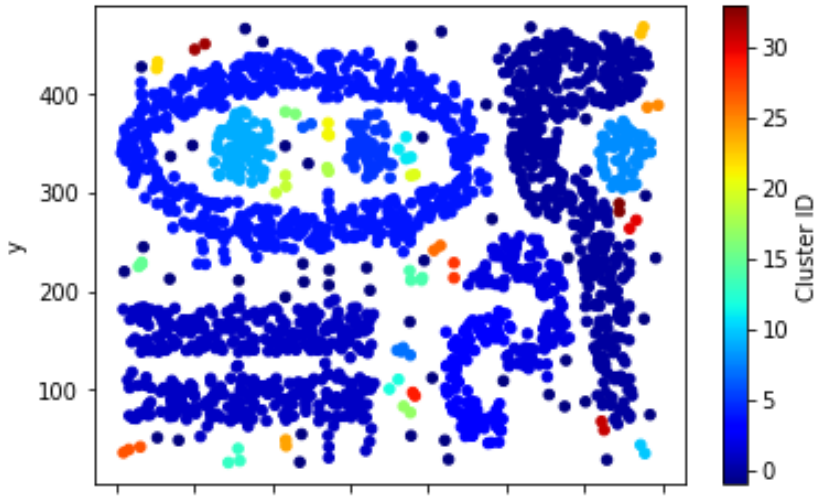
4. (5 points points) On the chameleon dataset, how many clusters are produced by DBSCAN when the minimum number of points (min samples) is set to 1, 2, 3, 4, and 5, respectively, while neighborhood radius (eps) is set to a constant value of 15.5. For each instance, copy and paste the plot of the clusters.

A) Minimum number of points = 1

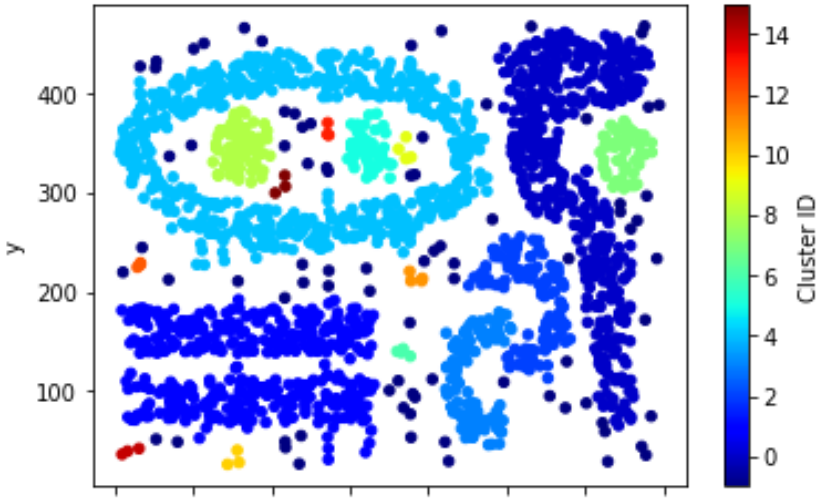
Number of clusters produced = 80



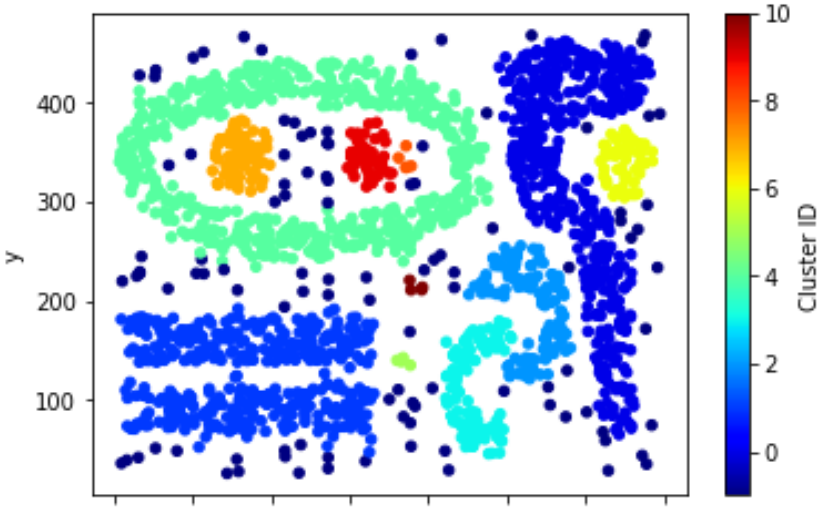
Minimum number of points = 2
Number of clusters produced = 34



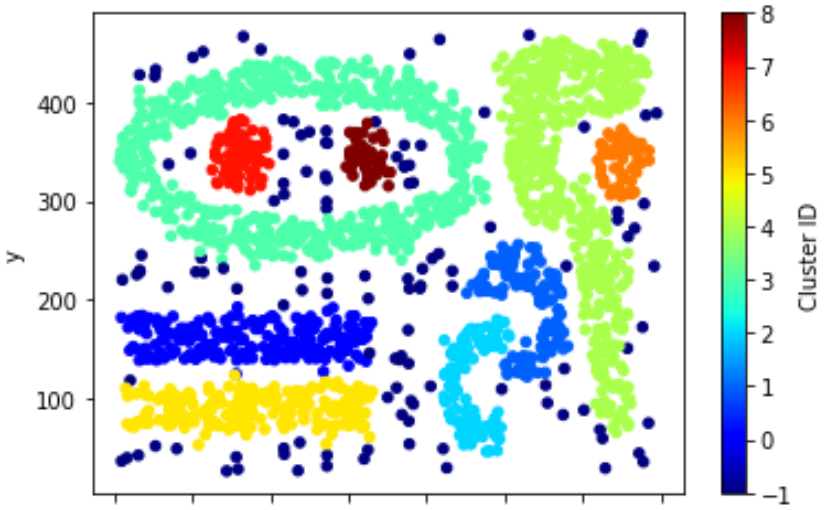
Minimum number of points = 3
Number of clusters produced = 16



Minimum number of points = 4
Number of clusters produced = 11

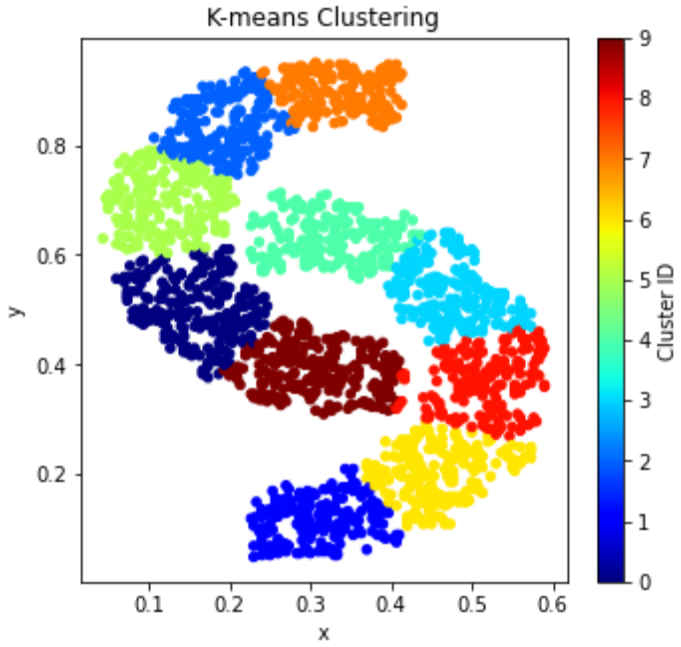


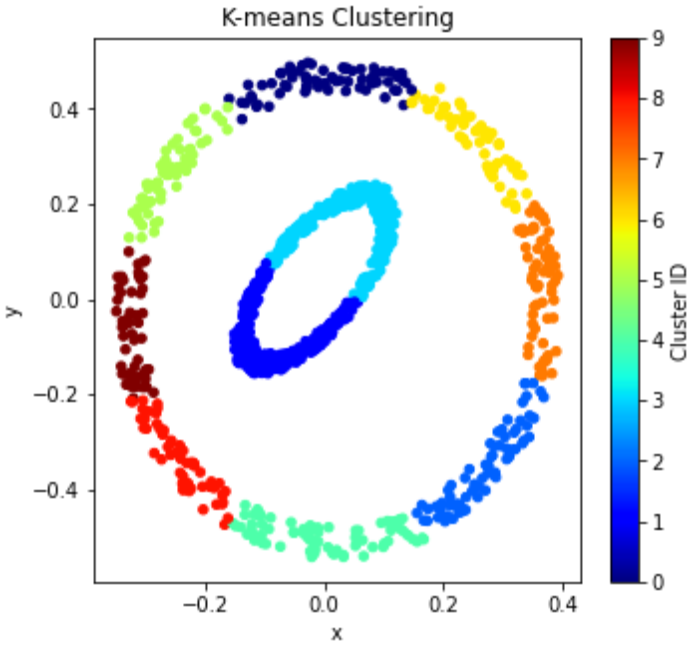
Minimum number of points = 5
Number of clusters produced = 9



5. (2 points points) For elliptical and 2D data, we applied k-means with k = 2. What happens if we use k = 10 for both these datasets? Copy and paste the clusters formed.

A) Clusters formed for 2D data and Elliptical data with k = 10 are shown below:





Problem 2 [4 points]

1. (2 points points) Let us look at the correspondence between the cluster labels and the original activity class labels. We see that each cluster has points coming from multiple classes, and is thus impure. Let's look at the maximum percentage of points in a cluster that are coming from a single class, which we can call as the 'purity' metric. For example, if a cluster consisting of 300 points has the following distribution of class labels:

- class 1 - 200
- class 3 - 50
- class 6 - 50

then the purity metric for this cluster will be 200/300, which is approximately 0.67. A higher value of this metric for a cluster signifies higher purity of the cluster. Compute this metric for all of the 6 clusters produced by running Kmeans with K = 6 on the given dataset. What is the maximum purity metric across all 6 clusters?

A) Purity of Cluster 1 is 0.462
Purity of Cluster 2 is 0.511
Purity of Cluster 3 is 0.502
Purity of Cluster 4 is 0.418
Purity of Cluster 5 is 0.702
Purity of Cluster 6 is 0.945
Maximum Purity = 0.945

2. (2 points points) What is the maximum purity metric for any cluster if we run Kmeans with K = 10 on the same dataset? Explain the rise/fall in purity as we increase K from 6 to 10.

A) Purity of Cluster 1 is 0.568
Purity of Cluster 2 is 0.878
Purity of Cluster 3 is 0.759
Purity of Cluster 4 is 0.958
Purity of Cluster 5 is 0.661
Purity of Cluster 6 is 0.638
Purity of Cluster 7 is 0.510
Purity of Cluster 8 is 0.536
Purity of Cluster 9 is 0.561
Purity of Cluster 10 is 0.753
Maximum Purity = 0.958.

Maximum Purity increases with an increase in the number of clusters because points are classified finely into clusters. In general, purity increases as the number of clusters increases. For instance, if we have a model that groups each observation in a separate cluster, the purity becomes one.

Problem 3 [18 points]

1. (2 points points) After handling duplicates, what is the count, mean, standard deviation minimum, and maximum values for the abstract word count and body word count?

A)

	Abstract word count	Body word count
Count	24584	24584
Mean	216.447	4435.475

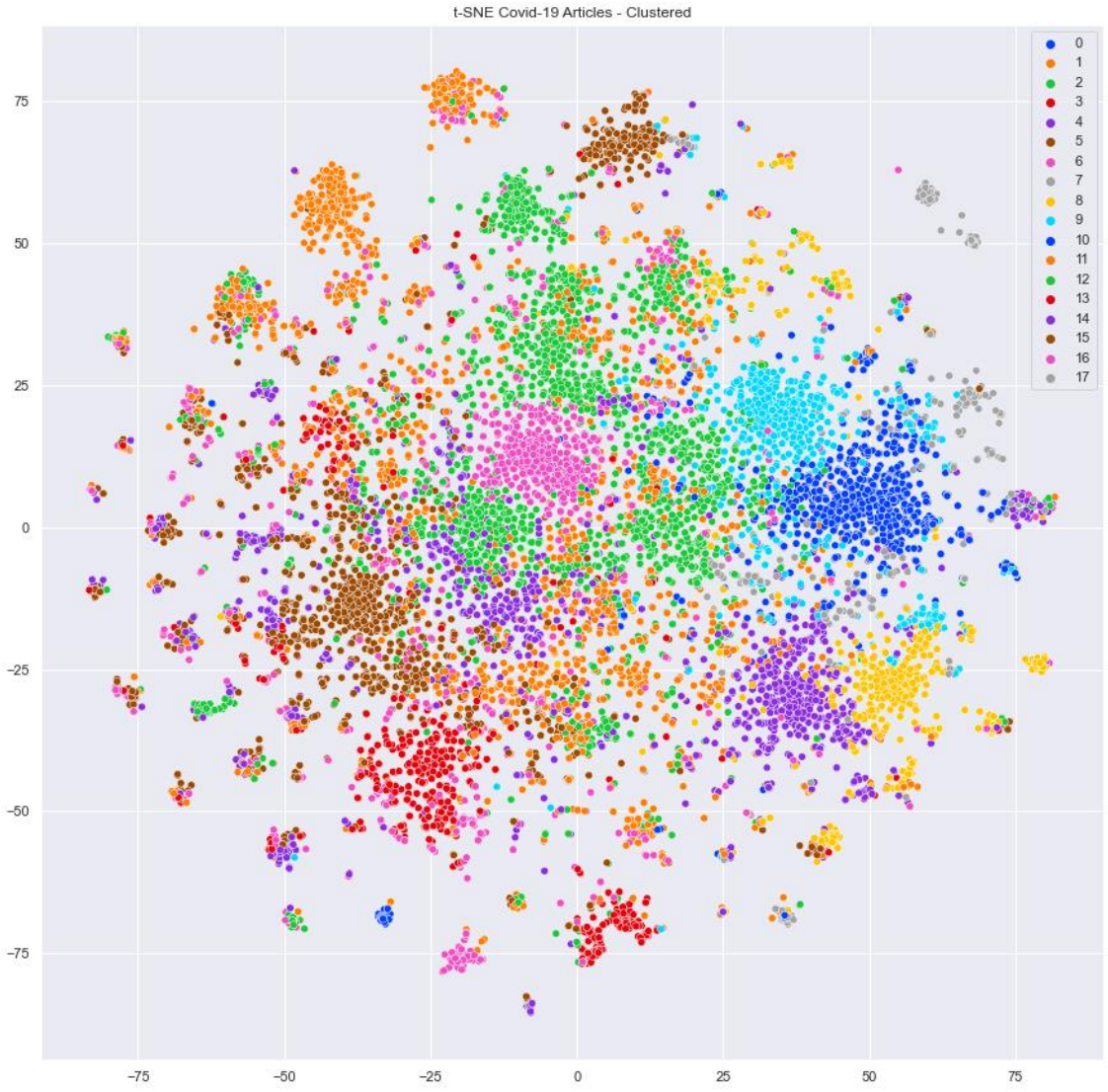
Standard Deviation	137.065	3657.421
Minimum	1	23
Maximum	3694	232431

2. (2 points points) Given the following word list: [’the’, ’2019’, ’novel’, ’coronavirus’, ’sarscov2’, ’identified’, ’as’, ’the’, ’cause’], what is its corresponding list of 2-grams?

A) List of 2-grams for the given list of words is:
['the2019', '2019novel', 'novelcoronavirus', 'coronavirussarscov2', 'sarscov2identified', 'identifiedas', 'asthe', 'thecause']

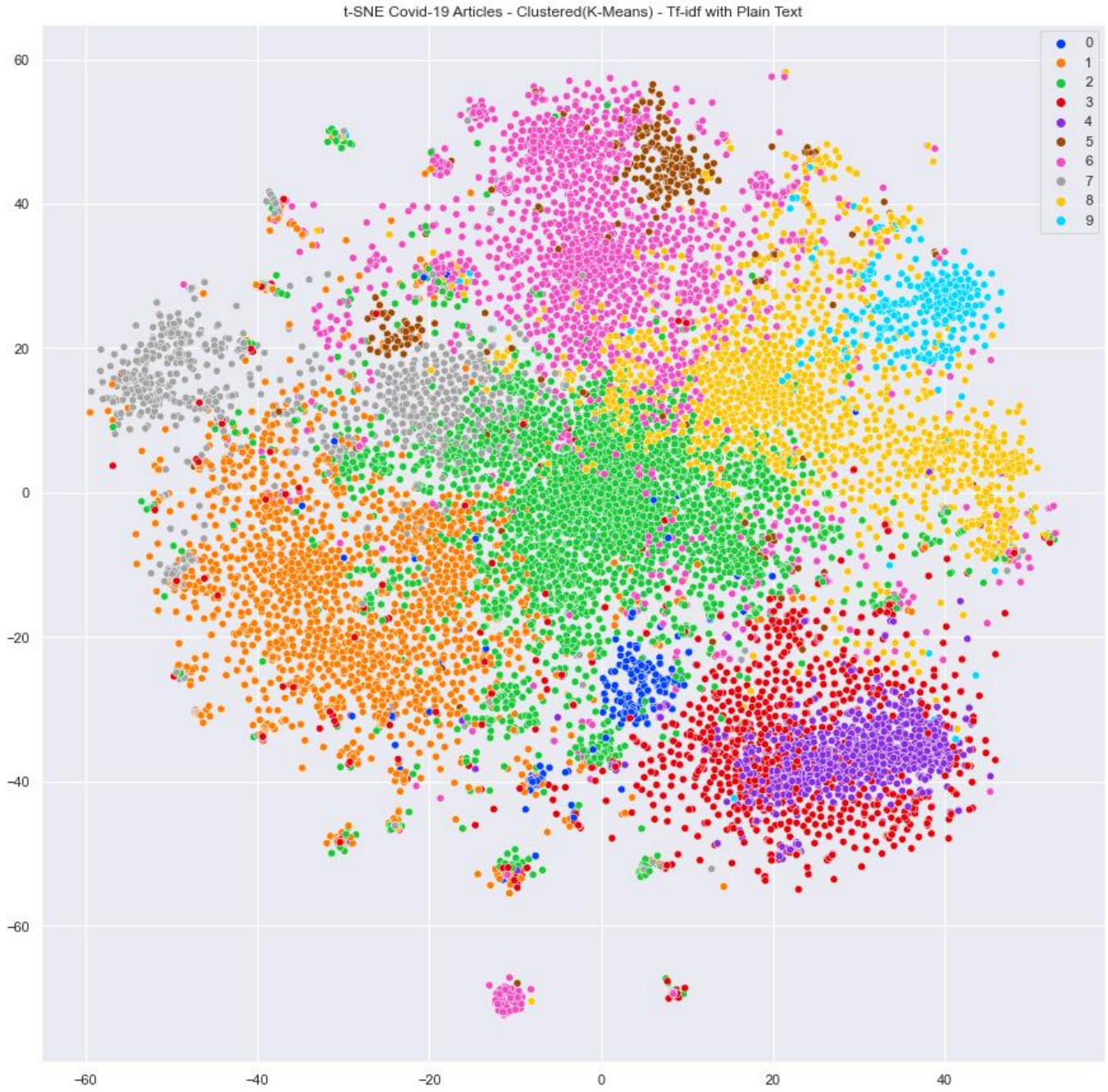
3. (4 points points) When we applied k-means clustering with K = 10 on the data created using HashingVectorizer features from 2-grams, we could see that some clusters still had some overlap in the t-SNE plot. Can you improve this by changing the number of clusters? What value of K visually leads to good separation among the clusters in the t-SNE plot? Copy and paste the corresponding t-SNE plot.

A) K=18 leads to a good separation of clusters in the t-SNE plot.



4. (4 points points) By using tf-idf vectorizer and plain text features instead of 2-grams, we could see that the clusters obtained from K-means clustering (with K = 10) are more separable in the t-SNE plot. What happens when we apply the tf-idf vectorizer on the 2-gram representation of documents instead of plain text, and then apply K-means clustering with K = 10? Copy and paste the corresponding t-SNE plot.

A) Using 2-gram representation and tf-idf vectorizer, and putting K=10, we get the below t-SNE plot:



5. (6 points points) In the interactive t-SNE with 20 clusters, can you do a manual analysis of different clusters to see what articles are clustered together? Choose any 5 clusters and write 4-5 keywords that describe it. Hover your mouse over the cluster point and you can see the article that it refers. You can use the box zoom feature to choose to display points of only one cluster in the plot, to simplify your analysis. Also, name the clusters that include articles involving social and economic impacts of the coronavirus?

A)

Cluster	Keywords
C19	gene, infection, antibody, response, protein
C18	lung, virus, test, influenza, cause
C17	probe, dna, target, respiratory, clinical
C16	virus, structure, membrane, viral, fusion
C15	group, delivery, gene, epitope, strain

Clusters that include articles involving social and economic impacts of the coronavirus are:

C7, C10, C0, C1, C2, C-18