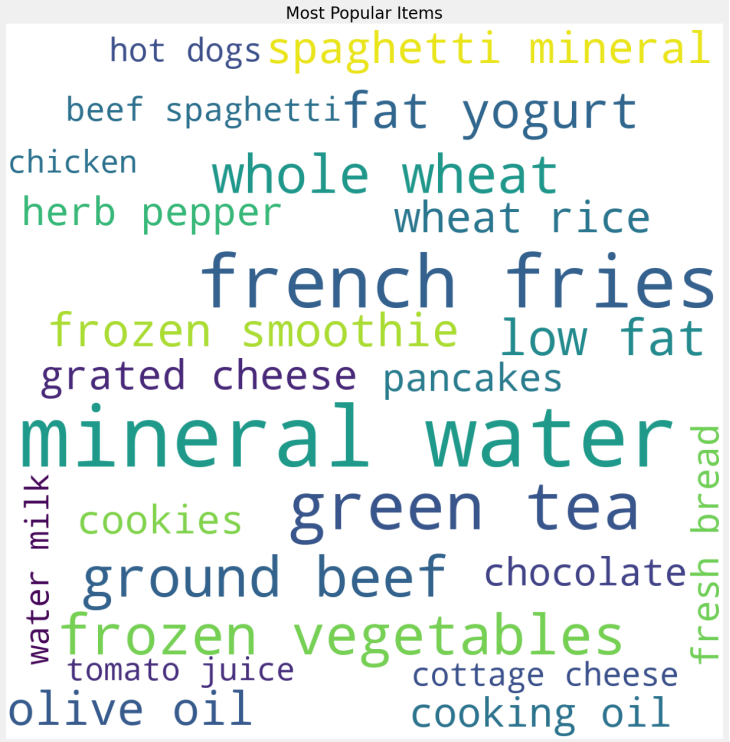
Introduction To Data Mining

Project 2

Sri Krishna Vamsi Koneru
5881358
koner033@umn.edu

Problem 1:

1) 7501
2) 20
3) Five of the transactions are,
   ['butter','light mayo','fresh bread'] , ['turkey','avocado'], ['chicken'] , ['cookies'] , ['tomatoes','milk','muffins','french fries']
4) maxwords=25



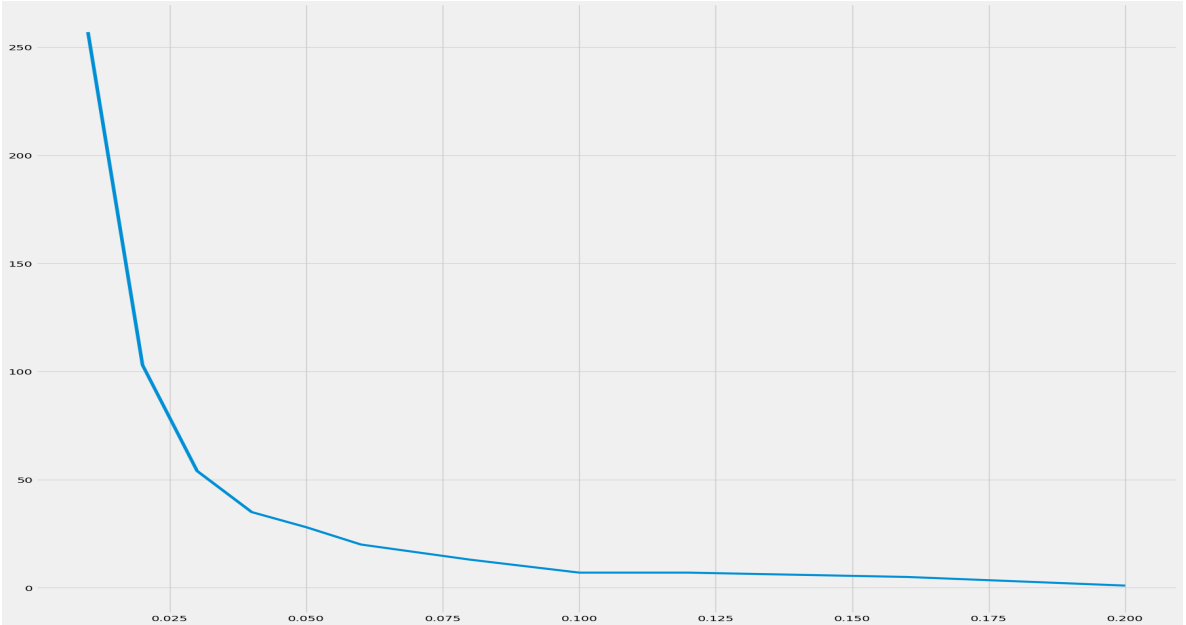
maxwords=50

They provide an overview of the most frequent/popular items in the transactions. The font size indicates how popular the items are like mineral water,french fries. Also, all words in maxwords=25 will be present in maxwords=50 as well because the former one is an even more specific subset with higher frequencies.

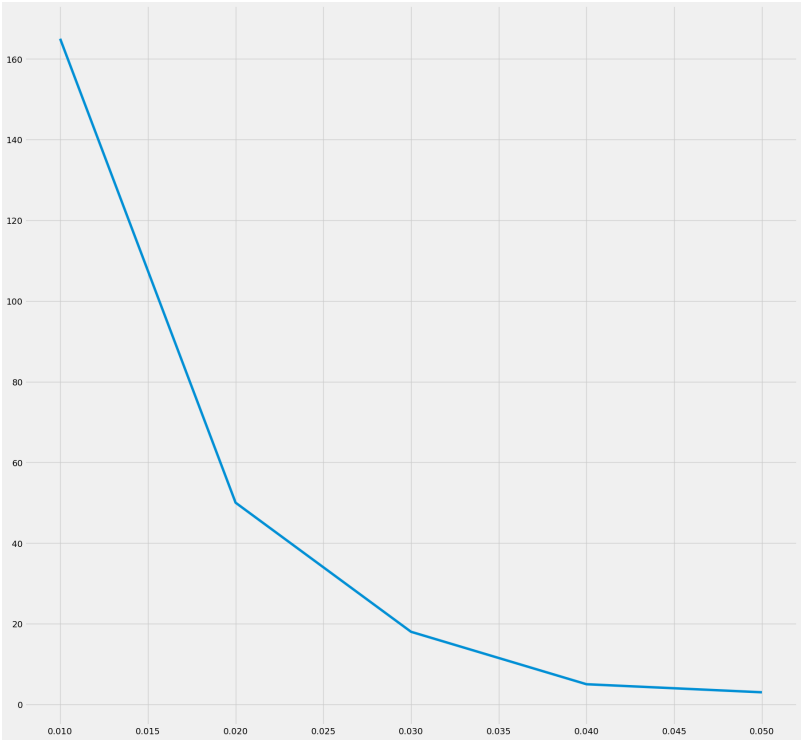
5) The top 5 most frequent items in the transactions in order are:
Mineral water, eggs, spaghetti, french fries,chocolate
6) The one hot-encoded Boolean array is:

	Apple	Bananas	Beer	Chicken	Milk	Rice	nan
0	True	False	True	True	False	True	False
1	True	False	True	False	False	True	True
2	True	False	True	False	False	False	True
3	True	True	False	False	False	False	True
4	False	False	True	True	True	True	False
5	False	False	True	False	True	True	True
6	False	False	True	False	True	False	True
7	True	True	False	False	False	False	True

7) 120
8) The plot between the support_thresholds and the frequent itemsets is like:
From the plot , it is clear that the number of frequent itemsets decreased as support threshold increased. Also, from around support=0.03 and 0.04 the frequent itemsets are very few. This is because the itemsets are not going to be present in as many transactions as required to meet the support criterion.



9) This is because there are no frequent itemsets of size three with support 0.02 while there are some which are frequent when the support threshold is 0.01. This is because their subsets can have become infrequent or the 3-sized itemset has also not been able to meet the requirements.
10) The plot between the support thresholds and the frequent 2-sized itemsets is below: Based on the plot, it is clear that as support threshold increases, the number of frequent itemsets decreases. There is a steep decrease observed from support=0.01 to support=0.02. And at support>=0.04 there are very few frequent itemsets.



11) Mineral Water-0.238
Chocolate-0.163
Eggs-0.179
Eggs,mineral water-0.0509

Chocolate,mineral water-0.052

Problem 2:

1. First, we calculate support of each individual item and then based on the 5 orders data we get the following:

Item	Support count
apple	4
egg	3
carrot	3
milk	2

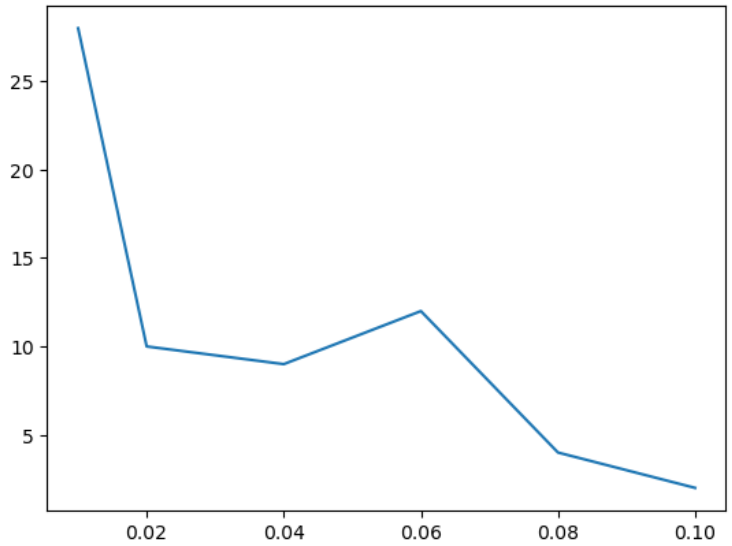
Now, we calculate the support count of the pairs having the items>minsup=3. So, we get following 2-size itemsets and their support counts

2-size itemset	Support count
[apple,egg]	3
[apple,carrot]	2
[egg,carrot]	1

Remaining do not exist and now based on this we can observe that only [apple,egg] meets the support criteria. Therefore it is the only pair that meets the minimum support threshold.

Alternatively, we can generate all itemset pairs and then get the support counts and even then we can conclude that [‘apple’,’egg’] only meets the criteria.

2. There are 3,214,874 unique orders and unique items are 49,677. The unique orders number is different from total number of orders, 32,434,489. The average width of the transactions can be found out by dividing the total orders by unique orders which is 10.088.
3. The plot between the various support thresholds and the number of association rules is,



Runtime decreases as support increases which is explainable because the number of itemsets and thus the rules generated decreases. However, at support threshold=0.1 the runtime is slightly increased. This can be observed from the various times obtained in the cells when executed for various supports. At s=0.08(i got cpu time 2.83 sec) but at 0.1 i got (3.59 sec). That was the only deviation from the trend.

4. At support threshold=0.01, Banana-> Strawberries

item 1	item 2	freq AB	Support AB	freq A	Support A	freq B	Support B	Conf AtoB	Conf BtoA	lift
Banana	Strawberries	29182	0.017	401800	0.241341	122223	0.0734	0.072628	0.238760	0.989306

At support threshold=0.02, Banana-> Large Lemon

item 1	item 2	freq AB	Support AB	freq A	Support A	freq B	Support B	Conf AtoB	Conf BtoA	lift
Banana	Large Lemon	28796	0.024	341573	0.2855	131742	0.110144	0.084304	0.218579	0.765397

At support threshold=0.04, Banana-> Strawberries

item	item	freq	Sup	freq	Sup	freq	Sup	Conf	Conf	lift

1	2	AB	port AB	A	port A	B	port B	Ato B	Bto A	
Banana	Strawberries	29182	0.046428	234940	0.37378	83098	0.132209	0.124210	0.351176	0.939503

At support threshold=0.06, Organic Strawberries-> Organic Baby Spinach

item 1	item 2	freq AB	Support AB	freq A	Support A	freq B	Support B	Conf Ato B	Conf Bto A	lift
Organic Strawberries	Organic Baby Spinach	20068	0.0659	150887	0.4957	132748	0.43616	0.133	0.1511	0.3049

At support threshold=0.08, Banana-> Organic Strawberries

item 1	item 2	freq AB	Support AB	freq A	Support A	freq B	Support B	Conf Ato B	Conf Bto A	lift
Banana	Organic Strawberries	38564	0.32	57125	0.481	117577	0.991826	0.675081	0.327989	0.6806

At support threshold=0.10, Banana-> Bag of Organic Bananas

item 1	item 2	freq AB	Support AB	freq A	Support A	freq B	Support B	Conf Ato B	Conf Bto A	lift
Banana	Bag of Organic Bananas	654	0.556	1176	1.0	1176	1.0	0.556	0.556122	0.556122

As a data scientist, I would suggest taking some measures like posting related ads from the rules we observed as there is enough confidence to deduce that people are actually buying the products together. Also, the app

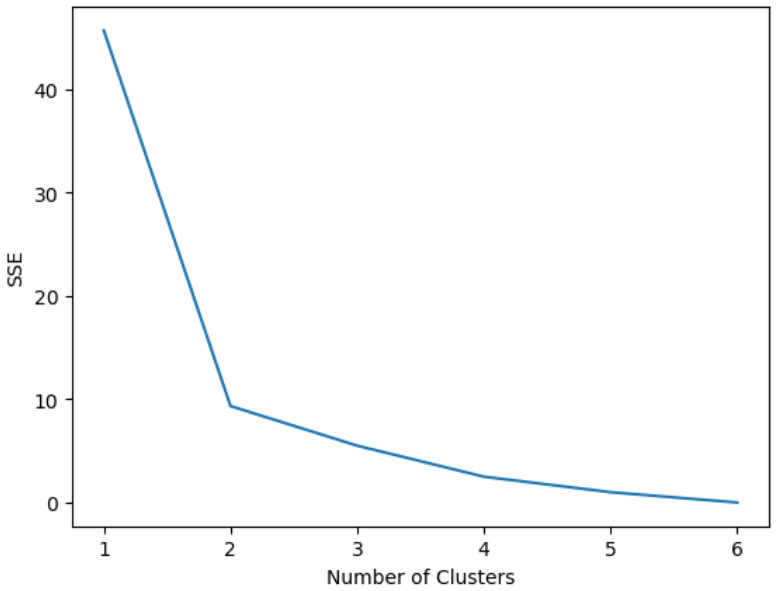
can work out on some promotional offers and bundles where they put together popular items like bananas, strawberries(a few that I observed over the various support thresholds) along with some other related products like organic fruit products based on what I observed in the aprioiri analysis.

Problem 3:

1) Since Adel,Kevin,Jessi prefer action over horror movies they have been assigned to a different cluster than Paul who prefers horror movies.

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

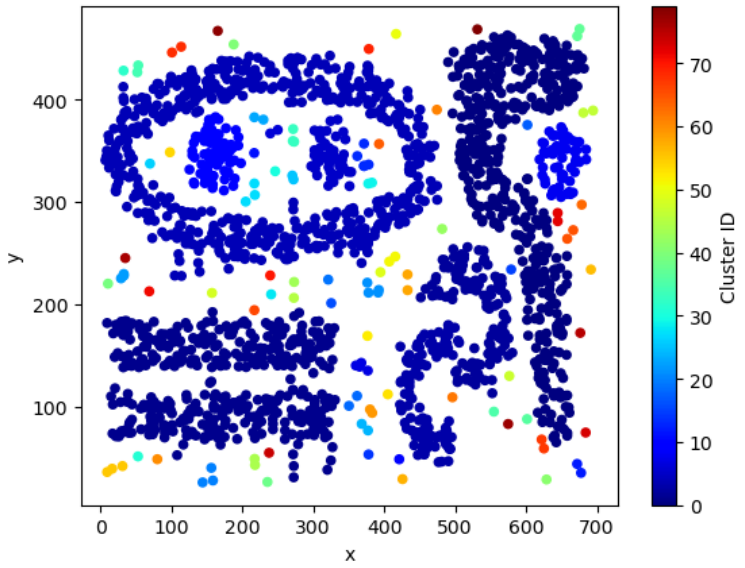
2) The optimal number of clusters is when k=2. This is because as can be observed in the below SSE vs number of clusters graph, the elbow is at two clusters and from then it is continuous decline. Hence, there is not benefit of adding more clusters.



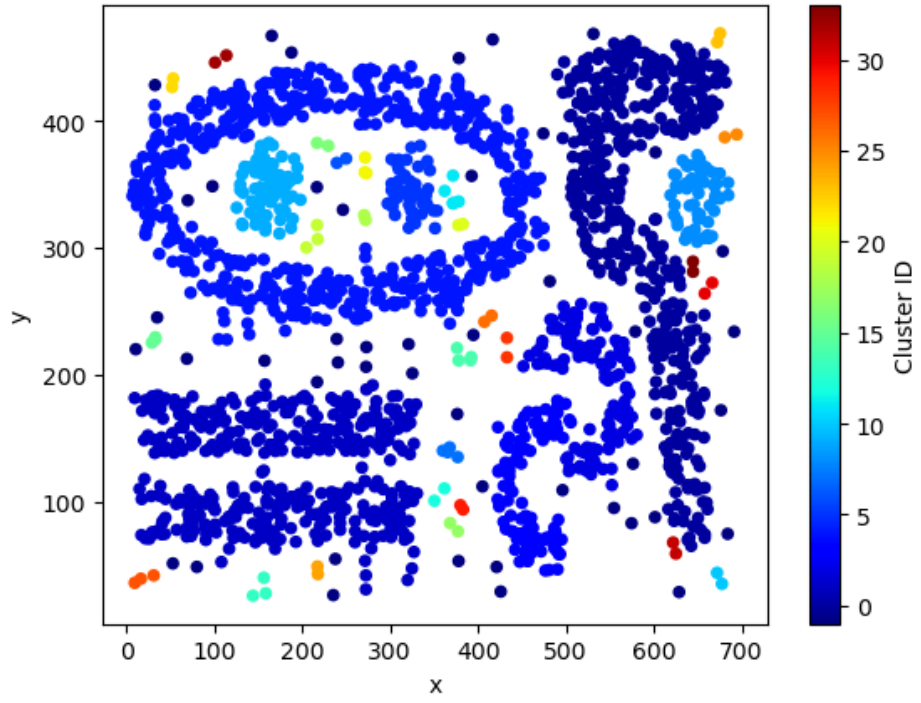
3) Group Average. This is because if we take a closer look at the dendograms, compared to other MIN method MAX and group average are clustering way better. In both of them local similarities are given precedence when merging and finally they are being merged towards the end on basis of attributes like has legs, is aquatic etc. However, I feel Group Average slightly performs better. This is mainly what I deduced because of the way I felt the object “Python” is being merged. In MAX, Python is being merged with frog and salamander in the early steps indicating good similarity between them but in reality, python is reptile while other two are amphibians. Even putting that aside and looking at attributes python and frog have only Hibernates as common. Also, there are some more like the clustering of bat and pigeon together

while belonging to different classes. Even though not perfect, I feel Group Average is a better clustering.

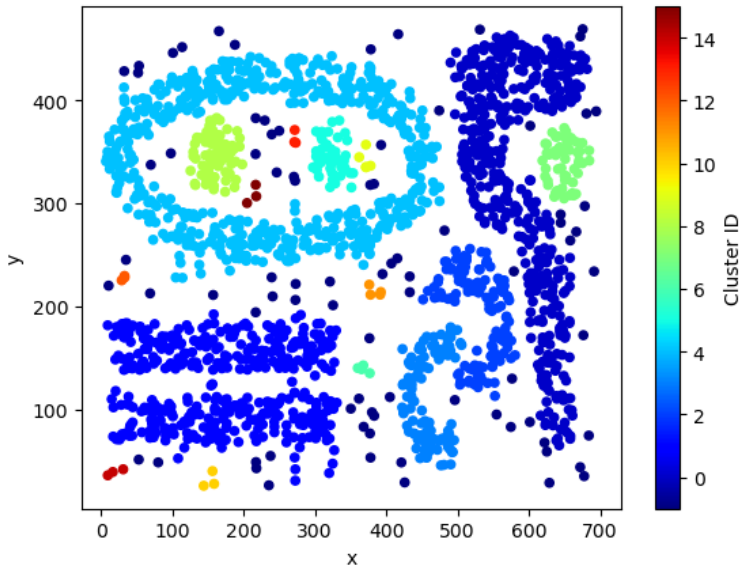
4) When min number of points=1, the number of clusters=80



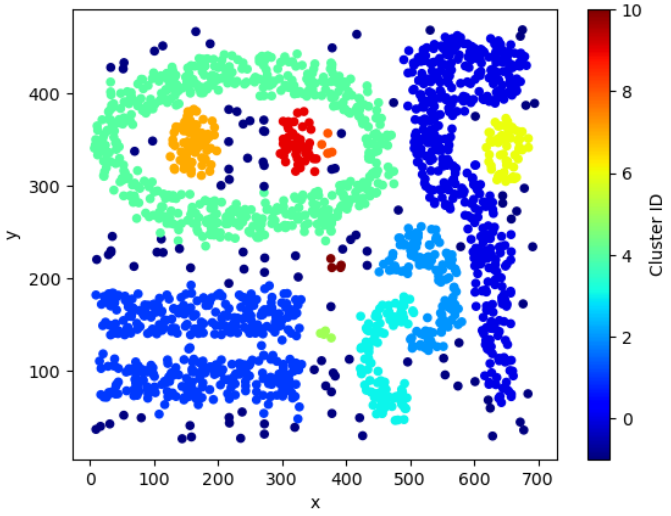
When min no.of points=2, the number of clusters is 35



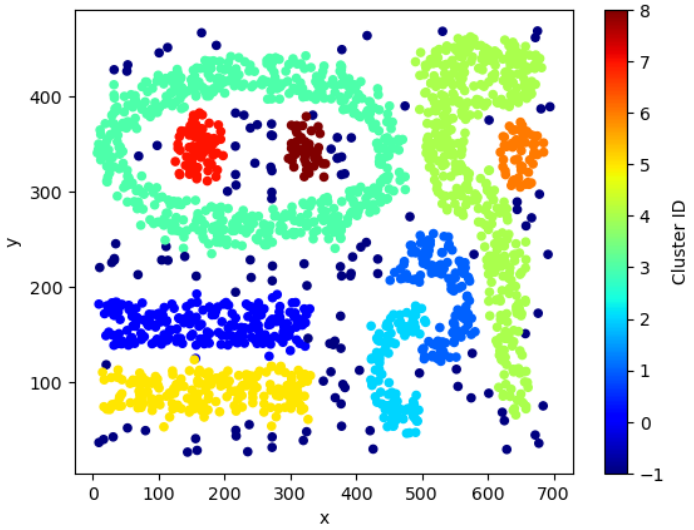
When min no.of points=3, the no.of clusters is 17



When min no.of points=4, the number of clusters is 12

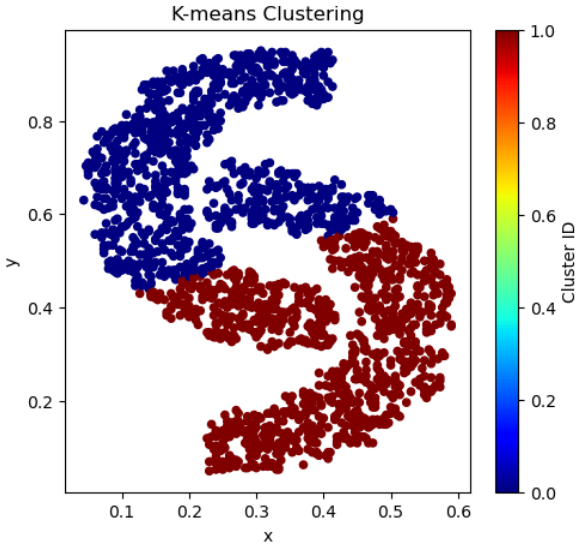


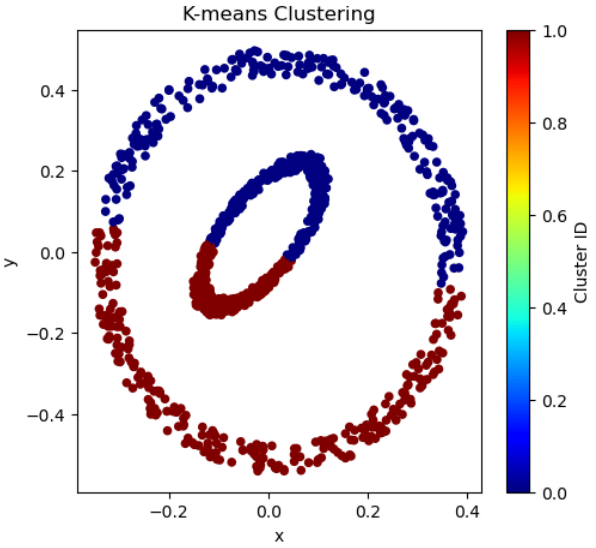
When min no.of points=5, the number of clusters is 10.



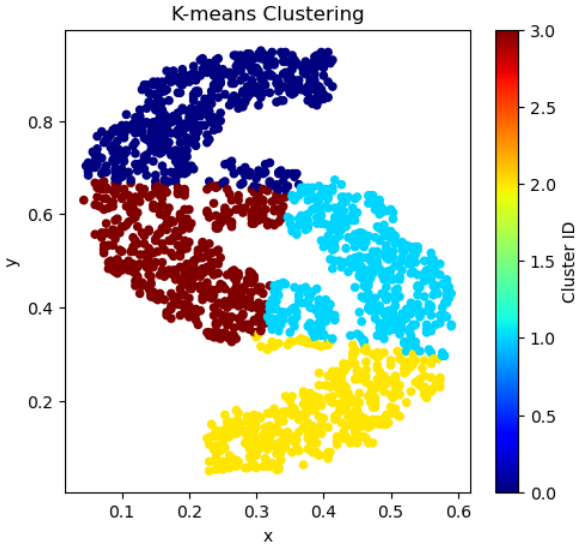
I have included the outliers/noise points also as acluster in the calculations.

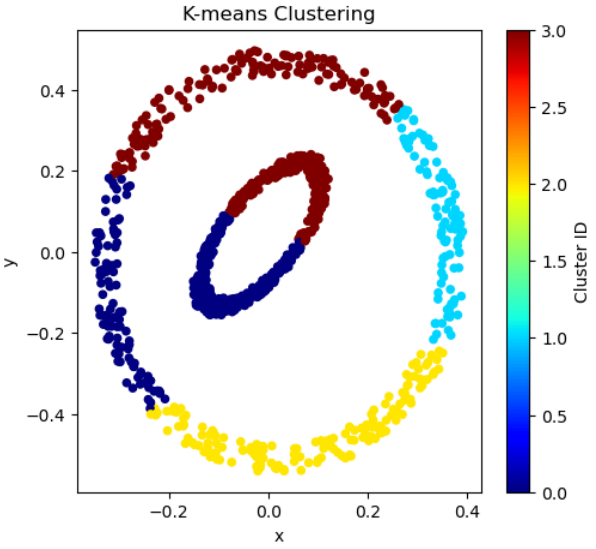
5) K-means with k=2



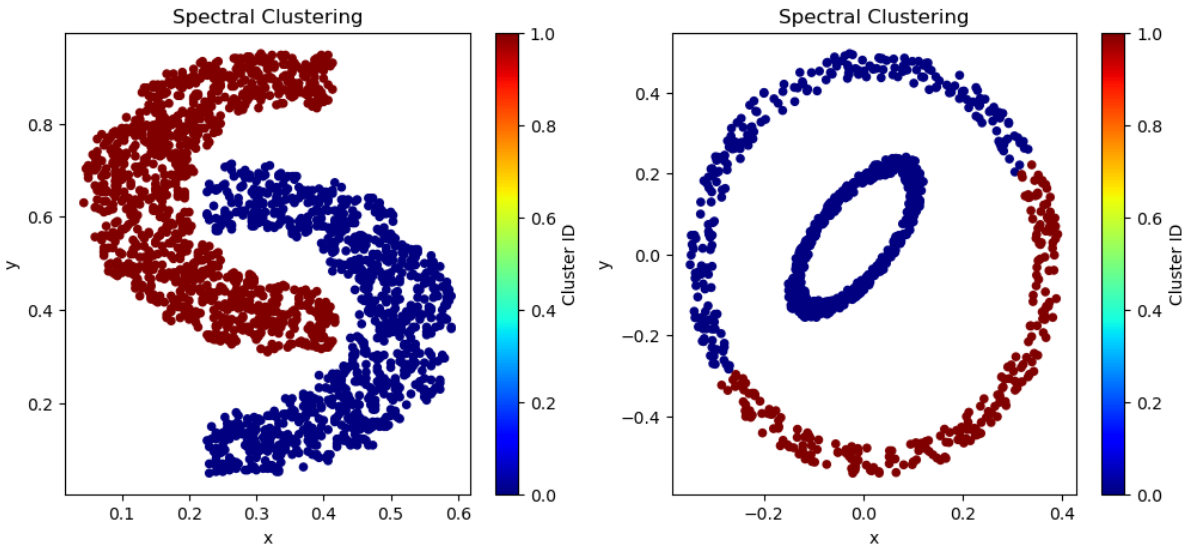


K-means with k=4

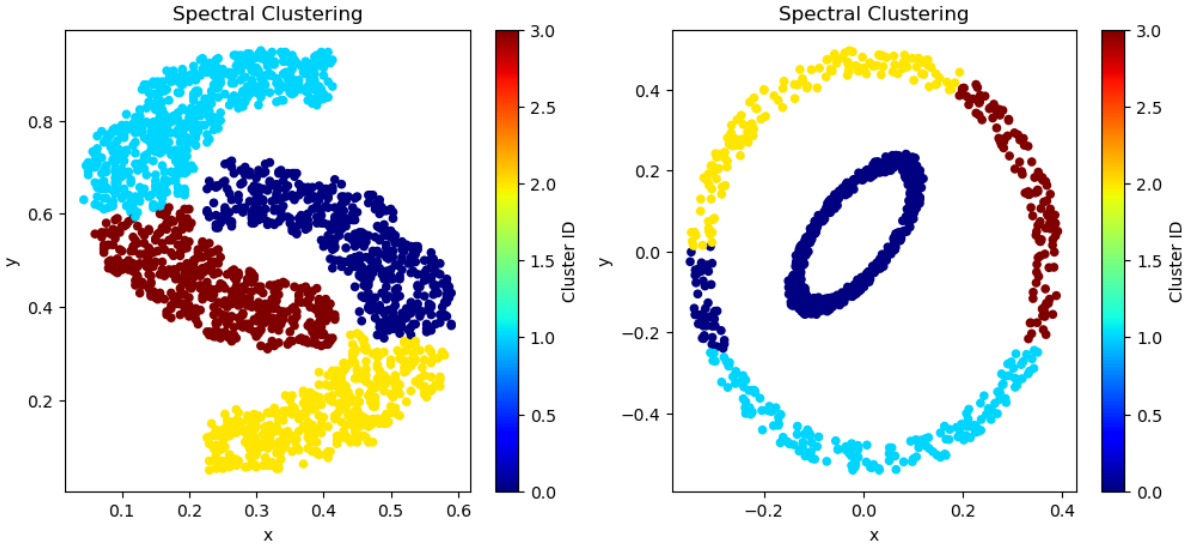




Spectral Clustering with k=2



Spectral Clustering with k=4,



We know in each image, left side is for data1(2D dataset) and right side is for data2(Elliptical dataset).

At k=2, for 2D dataset clearly Spectral Clustering does a better job but for Elliptical dataset both perform pretty similarly but K-means does a slightly better job because spectral is not clustering the inner circle datapoints properly.

At k=4, for 2D dataset again Spectral clustering performs better. But for the elliptical dataset, it seems like K-means does a better job again as spectral is not clustering the inner circle again properly.

Problem 4:

1)

	Abstract Word Count	Body Word Count
count	24584	24584
mean	216.4466	4435.475
Standard Deviation	137.065	3657.421
Minimum	1	23
Maximum	3694	232431

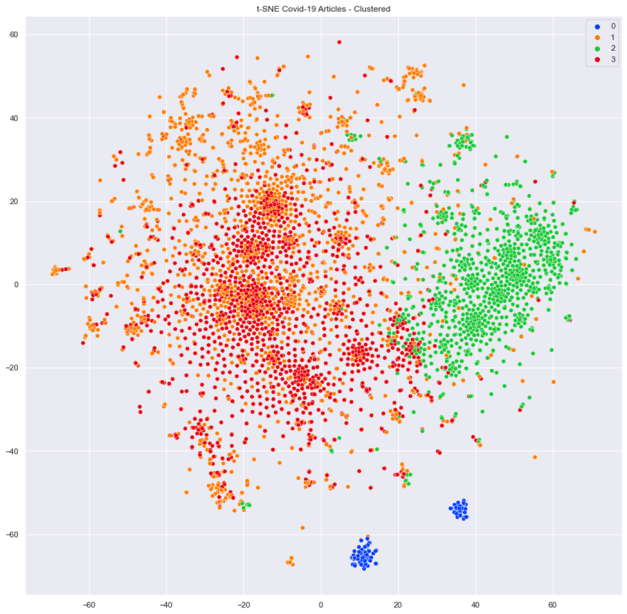
2) For cleaning up the text,
   First we remove the punctuation from each text and then we convert the text to lower case.

3) For our first try while creating the feature vector, we tried grabbing the body part of the article. Since we succeeded we proceeded with that. Hence, we can conclude that we focused on the body text part of the article.

4) An n-gram is a sequence of N-words. It can be described as a contiguous sequence of tokens, however in our specific case this can be considered as words. N-grams are used in ML where we want to classify documents. The 2gram for given set of words is:
   the2019,2019novel,novelcoronavirus,coronavirussarscov2,sarscov2identified,identifiedas,asthe,thecause,causeof

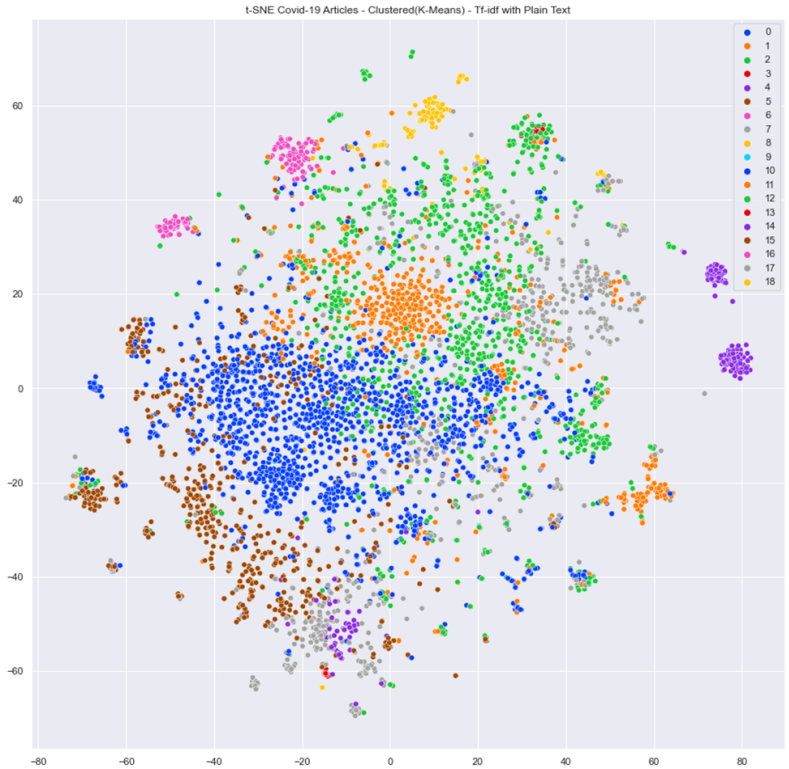
5) HashingVectorizer is used to create the featurevector X.It converts the text document into a sparse matrix taking note of token occurrence. In our case, we used the size 2^12=4096 as the feature size.

6) I have tried running the cell with varying values of features and clusters. I first tried some combinations with features 2^14 and 2^10 but I didnt find results better than the ones I did when I ran it with 2^12. Hence, I reverted back to 2^12. As for values of k, I tried values ranging from 4 to 20. However,I think it suits better when k is of a lower value however not too low. Alternatively, this could be concluded better based on the Silhouette coefficient and other measures which can be used to determine cluster validity. However, it was very challenging



k=4, features= 2^{12}

7) I have tried changing the max features value and the cluster size(value of k) from 10-20 and I found out the k=19 to be good with feature size being 2^{12}. I have tinkered around with it but I could not find any better satisfactory result.



k=19, no.ofmax_features= 2^{12}

#For 6&7 also I had difficulties running in my notebook due to large times taken for execution. It ran sometimes without any problem but on last run, it took too much time and did not complete execution.

8) # My notebook had issues while running this. I tried installing Bokeh but faced some issues.. I referred to the github link you gave along with the notebook to solve this question.

   So, in the interactive t-SNE with 20 clusters, on doing manual analysis and observing, the clusters had articles with some specific keywords. A few are,

   Cluster 1- patient, outbreak,vaccine,test,antibody
   Cluster 3- protein, cell, structure,genome, sample
   Cluster 5- antiviral,inhibitor, reaction,concentration, bind
   Cluster 10- cat, dog, rabbit, pig, feline
   Cluster 15- smallpox, dengue,influenza,strain,clinical

   Clusters that include articles on social, economic impacts of coronavirus are Clusters 12,8,18,3,11,14.