Question 1

1. There are 7501 records in the dataset.

2. The maximum number of items a customer has bought is 20.

3.

Transactions	Items
1352	Cookies, water spray,
5304	frozen smoothie, cottage cheese ,brownies
3383	Eggs, melons, mint, green tea
4781	Chicken, cereals
3980	frozen vegetables, ground beef, spaghetti, mineral water

4.

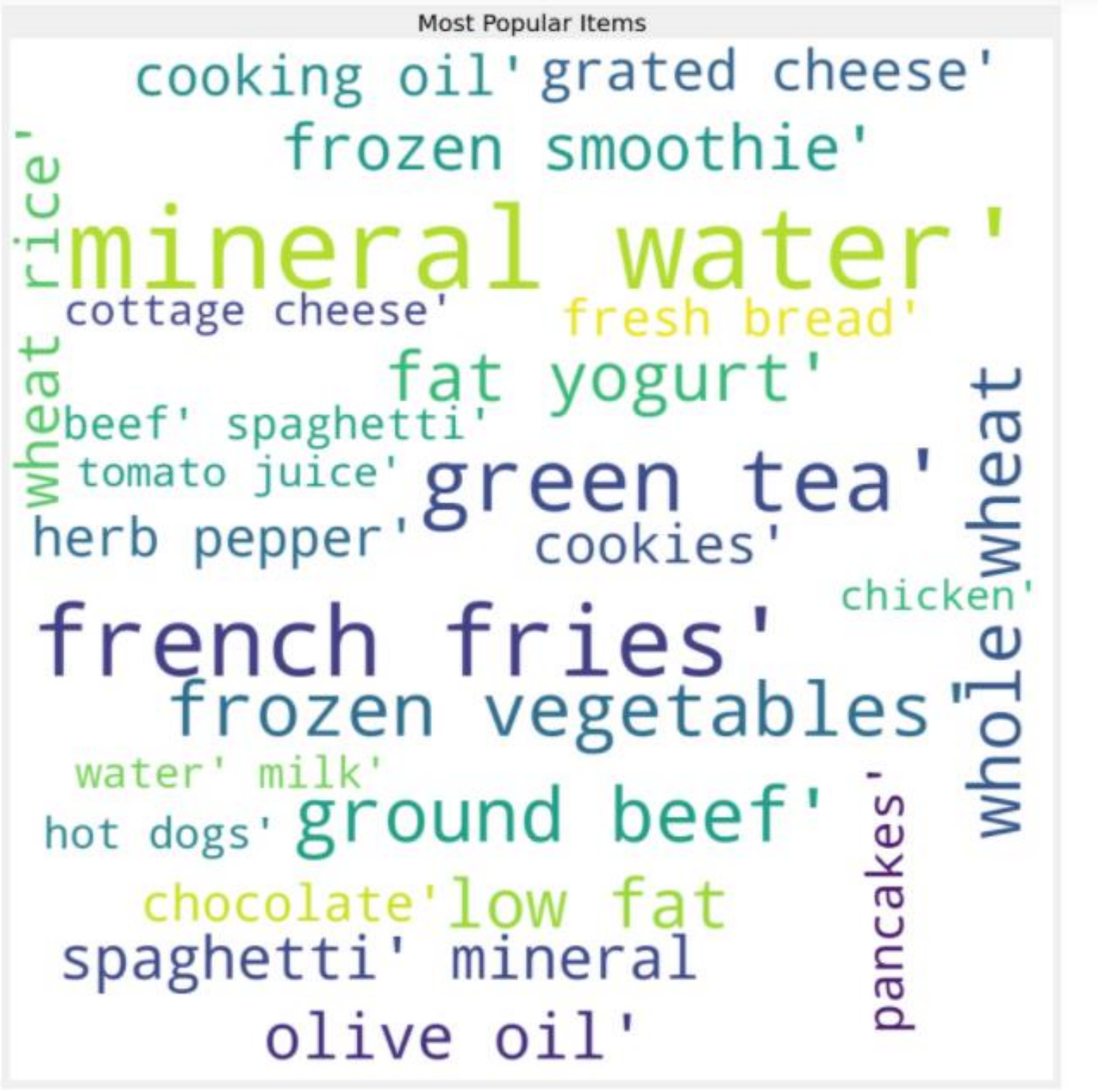


Figure 1. wordcloud with max_words set to 25

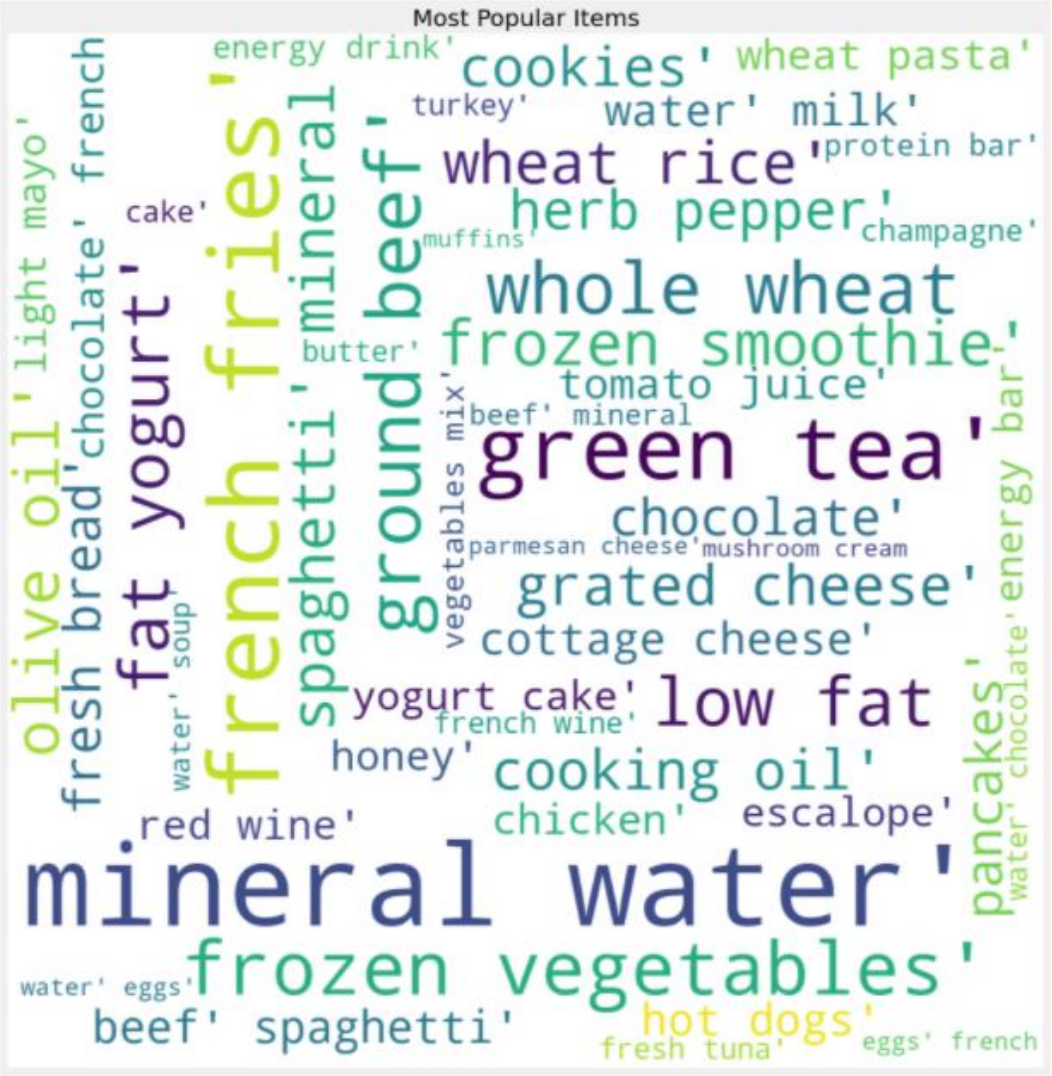
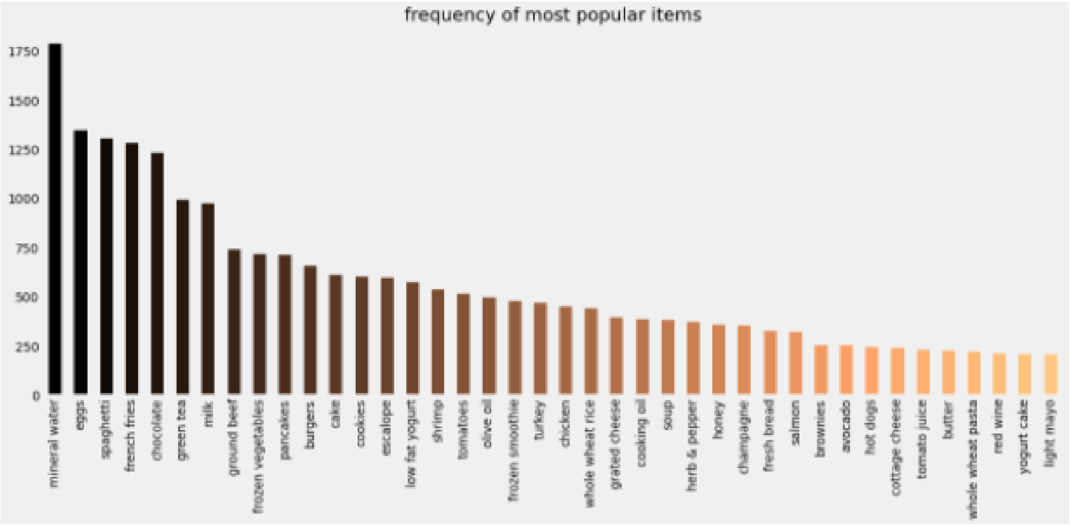


Figure 2. wordcloud with max_words set to 50

The wordcloud shows the frequency by the size of the word. With the larger size, the word would have higher frequency. So, the ‘minearl water’ has the highest frequency. When we expanded the search area, we could there are more words in second graph than those of the first graph.

5.



Top 5 most frequent items are Mineral water, eggs, spaghetti, french fries and chocolate.

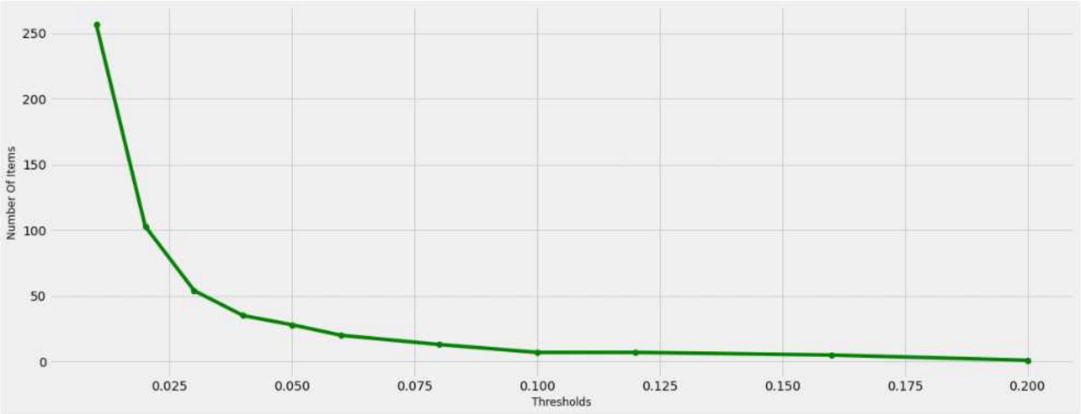
6.

The one-hot encoded Boolean array would be (Six columns are Apple, Bananas, Beer, Chicken, Milk and Rice, respectively):

```
[[True False True True False True]
[True False True False False True]
[True False True False False False]
[True True False False False False]
[False False True True True True]
[False False True False True True]
[False False True False True False]
[True True False False False False]]
```

7. There are 121 unique items in the input dataset.

8.

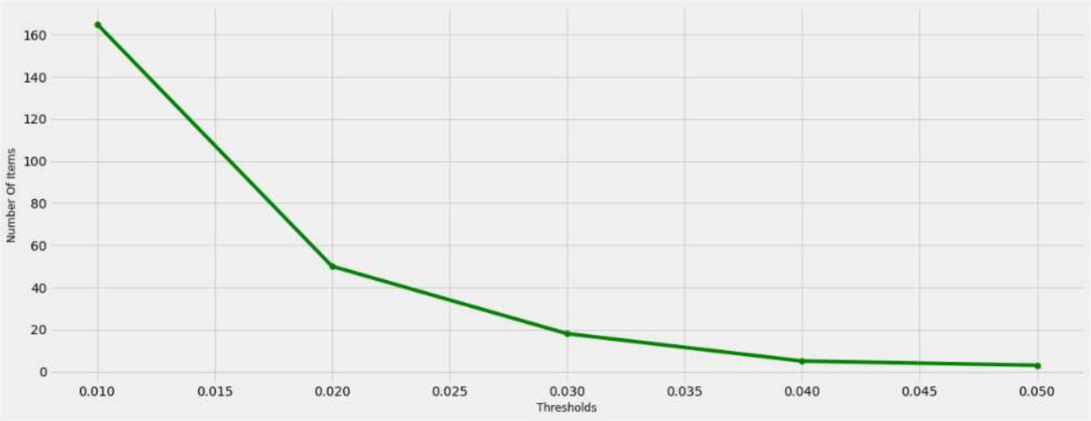


When the support threshold is increasing, the number of items is decreasing. Support threshold means that only the itemsets with support higher than or equal to the support threshold would be saved. So, when we increase the number of support threshold, fewer items are frequent itemsets and saved by the Apriori algorithm. Thus, the number of items decreases.

9.

The support of size 3 frequent itemset is greater than 1% but less than 2%. When we set the support threshold to 2%, the size 3 frequent itemset is no longer frequent itemset and be discarded by the Apriori. The support of size 1 and 2 frequent itemset is greater than 2%.

10.



When the support threshold is increasing, the number of itemsets of length 2 is decreasing. Support threshold means that only the itemsets of length 2 with support higher than or equal to the support threshold would be saved. So, when we increase the number of support threshold, fewer itemsets of length 2 are frequent itemsets and saved by the Apriori algorithm. Thus, the number of items decreases and we have a decreasing trend.

11.

Itemset	Support
Mineral Water	0.238
Chocolate	0.164
Eggs	0.180
Eggs, Mineral Water	0.051
Chocolate, Mineral Water	0.053

Question 2

1.

{apple,egg} has the minimum threshold of 3.

Step 1:

We find the frequent itemset of length 1(support >= 3):

Apple: 4

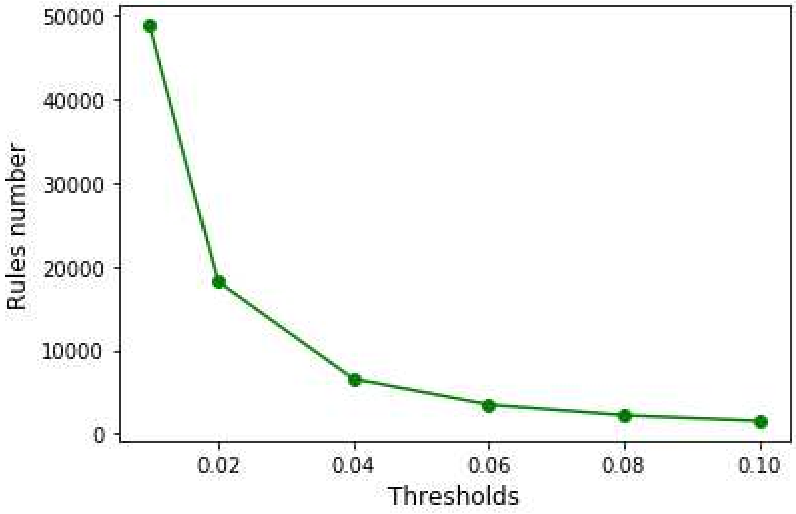
Egg: 3

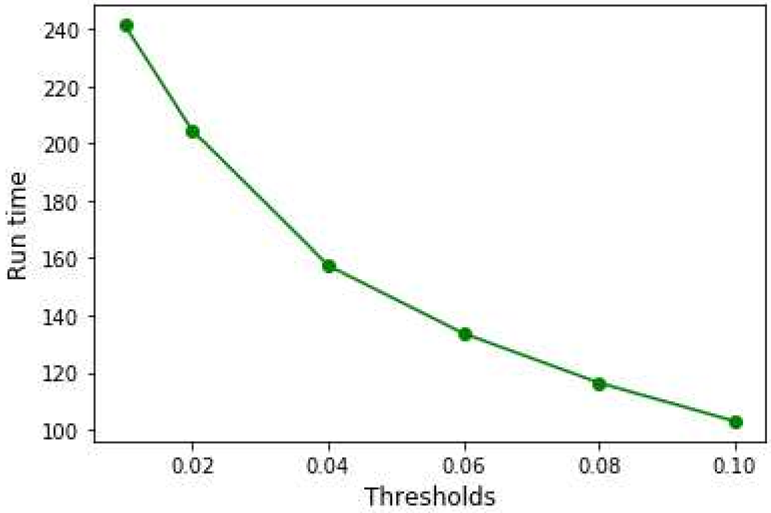
Carrot: 3
Step 2:
We find the frequent itemset of length 2(support >= 3):
{Apple, egg}: 3
{Apple, carrot}: 2
{Egg, carrot}: 1

Since we only have one frequent itemset of length 2, we do not need to consider about the frequent itemset of length 3 and the {apple, egg} is the only itemset having a minimum threshold of 3.

2.
There are 3214874 unique orders and 49677 unique items in the dataset. The number of records in the dataset is 32434489. Unique order is not same as the number of records in the dataset. A customer orders 84.25 items per order, on average.

3.

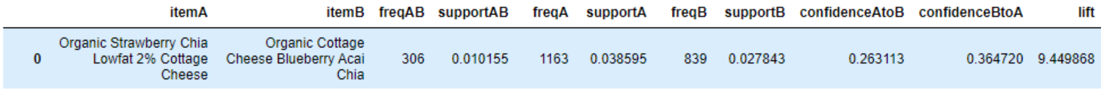




When the support threshold is increasing, the runtime is decreasing. Support threshold means that only the itemsets with support higher than or equal to the support threshold would be saved. So, when we increase the number of support threshold, fewer itemsets are frequent itemsets and saved by the Apriori algorithm. The non-frequent itemsets would be pruned by the Apriori algorithm. Thus, the runtime decreases, and we have a decreasing trend.

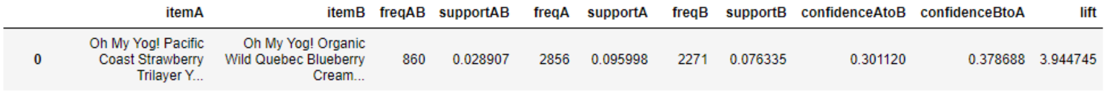
4.

When support threshold is 0.01:



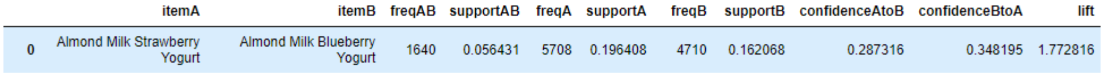
	itemA	itemB	freqAB	supportAB	freqA	supportA	freqB	supportB	confidenceAtoB	confidenceBtoA	lift
0	Organic Strawberry Chia Lowfat 2% Cottage Cheese	Organic Cottage Cheese Blueberry Acai Chia	306	0.010155	1163	0.038595	839	0.027843	0.263113	0.364720	9.449868

When support threshold is 0.02:



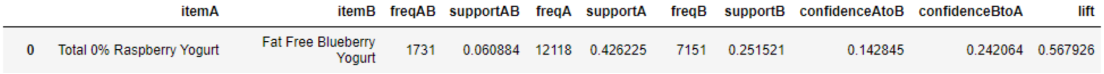
	itemA	itemB	freqAB	supportAB	freqA	supportA	freqB	supportB	confidenceAtoB	confidenceBtoA	lift
0	Oh My Yog! Pacific Coast Strawberry Trilayer Y...	Oh My Yog! Organic Wild Quebec Blueberry Cream...	860	0.028907	2856	0.095998	2271	0.076335	0.301120	0.378688	3.944745

When support threshold is 0.04:



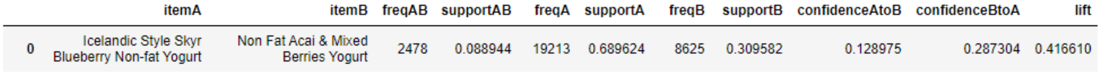
	itemA	itemB	freqAB	supportAB	freqA	supportA	freqB	supportB	confidenceAtoB	confidenceBtoA	lift
0	Almond Milk Strawberry Yogurt	Almond Milk Blueberry Yogurt	1640	0.056431	5708	0.196408	4710	0.162068	0.287316	0.348195	1.772816

When support threshold is 0.06:



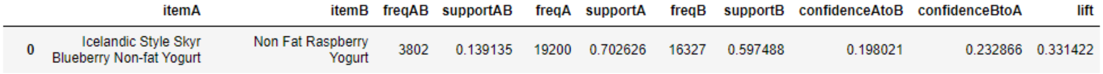
	itemA	itemB	freqAB	supportAB	freqA	supportA	freqB	supportB	confidenceAtoB	confidenceBtoA	lift
0	Total 0% Raspberry Yogurt	Fat Free Blueberry Yogurt	1731	0.060884	12118	0.426225	7151	0.251521	0.142845	0.242064	0.567926

When support threshold is 0.08:



	itemA	itemB	freqAB	supportAB	freqA	supportA	freqB	supportB	confidenceAtoB	confidenceBtoA	lift
0	Icelandic Style Skyr Blueberry Non-fat Yogurt	Non Fat Acai & Mixed Berries Yogurt	2478	0.088944	19213	0.689624	8625	0.309582	0.128975	0.287304	0.416610

When support threshold is 0.1:



	itemA	itemB	freqAB	supportAB	freqA	supportA	freqB	supportB	confidenceAtoB	confidenceBtoA	lift
0	Icelandic Style Skyr Blueberry Non-fat Yogurt	Non Fat Raspberry Yogurt	3802	0.139135	19200	0.702626	16327	0.597488	0.198021	0.232866	0.331422

For the above results, I would analyze them mostly in three aspects: lift, confidenceAtoB and confidenceBtoA.

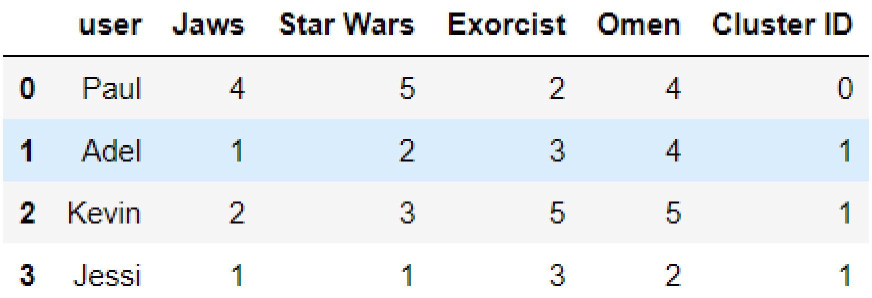
First, let's focus on the itemset with lift > 1, which means the two products have positive relationship. For example, Organic Strawberry Chia Lowfat 2% Cottage Cheese(A) and Organic Cottage Cheese Blueberry Acai Chia(B) have lift = 9.45, which is much higher than 1. Furthermore, these two products have confidenceBtoA equals to 0.364. Combining the lift = 9.45 and confidenceBtoA = 0.364 , we could decrease the price of

Organic Cottage Cheese Blueberry Acai Chia to increase the sales of the Organic Cottage Cheese Blueberry Acai Chia. Then the sales of Organic Strawberry Chia Lowfat 2% Cottage Cheese would also increase. This strategy is also suitable for Oh My Yog! Pacific Coast Strawberry Trilayer Yogurt and Oh My Yog! Organic Wild Quebec Blueberry Cream Top Yogurt & Fruit, and Almond Milk Strawberry Yogurt and Almond Milk Blueberry Yogurt.

Then, for the itemset with lift < 1, which means the two products have negative relationship. For example, Total 0% Raspberry Yogurt and Fat Free Blueberry Yogurt have lift = 0.568. Since both items have high support, medium confidence and negative relationship, we may bundle them together and sell them at a discount. This strategy is also suitable for Icelandic Style Skyr Blueberry Non-fat Yogurt and Non Fat Acai & Mixed Berries Yogurt, and Icelandic Style Skyr Blueberry Non-fat Yogurt and Non Fat Raspberry Yogurt.

Question 3

1.



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

2.

The optimal value of K is 2. For Elbow Method, When the selected k value is less than the real value, the corresponding sum-of-squared errors value will be greatly reduced every time k increases by 1, and when the selected k value is greater than the real K, every time k increases by 1, the corresponding sum-of-squared errors value will not change so significantly. In this way, we know that optimal value of K is 2.

3.

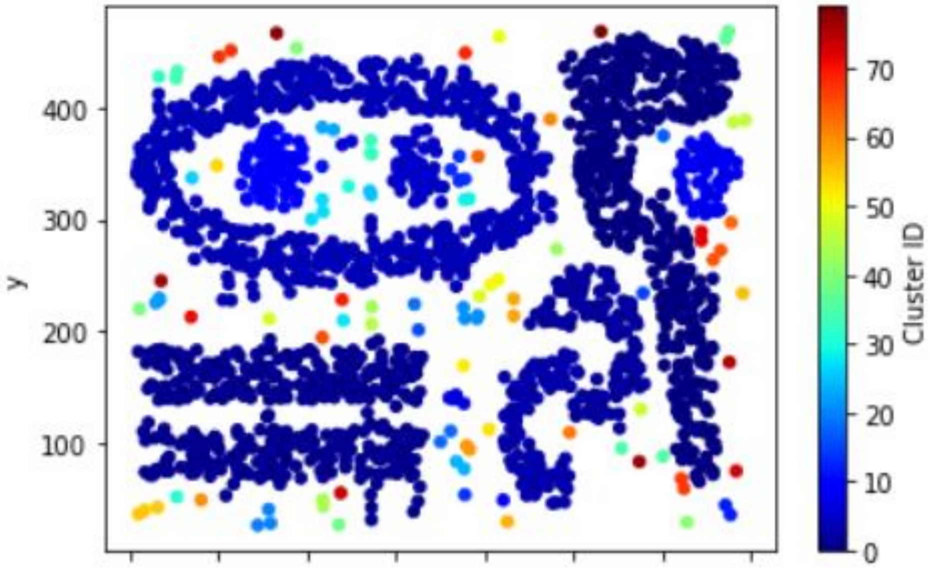
For Single link, it links two clusters based on the minimum distance between 2 elements and it tends to produce long thin clusters since the link based on only 2 points.

For Complete link, it links two clusters based on the max distance between 2 elements and it tends to be overly conservative.

For Average link, it takes the distance between every pair of elements.

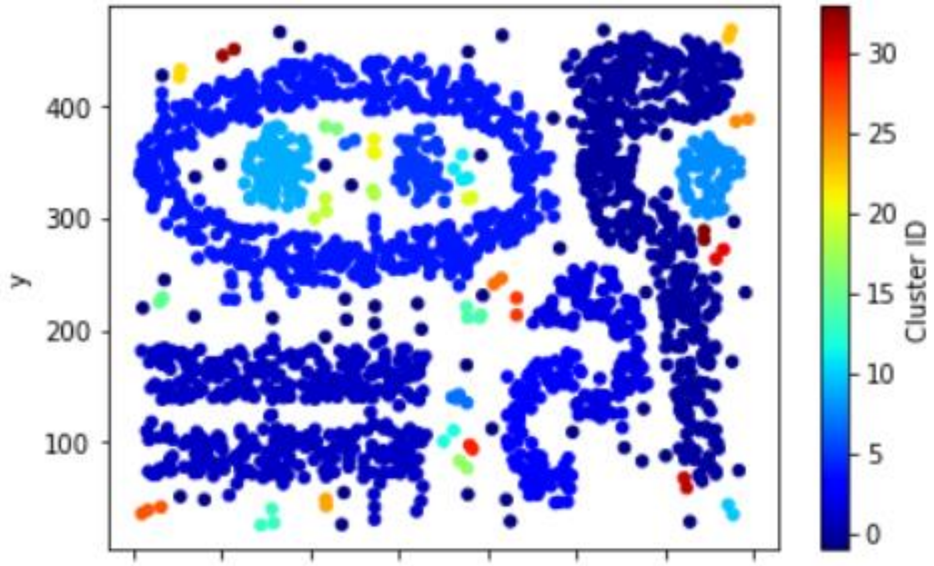
From the results, we know that the average link is the best among these three hierarchical clustering algorithms. The average link balances the breadth and depth and gives us a reasonable clustering.

4.



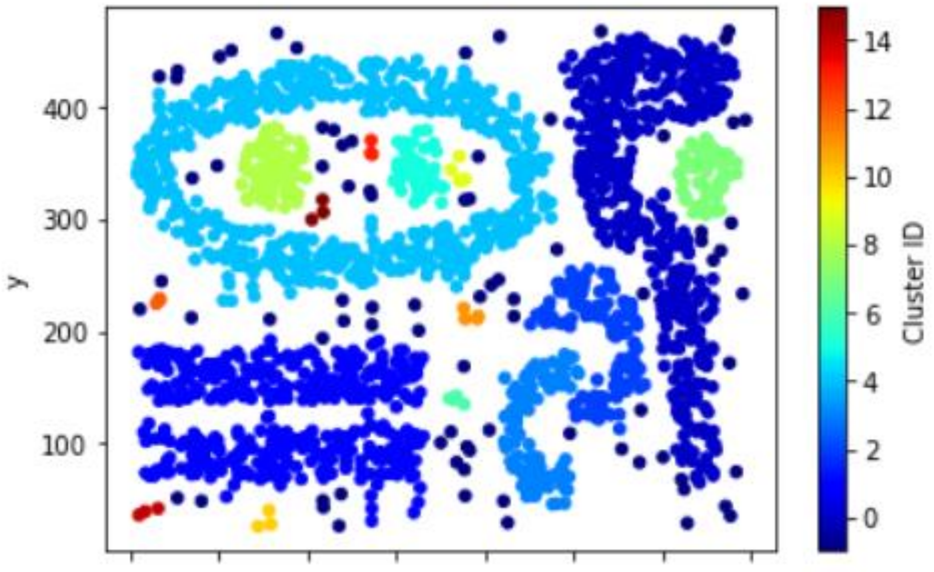
Minimum number of points set to 1

78 clusters are formed. Max: 77 Min: 0



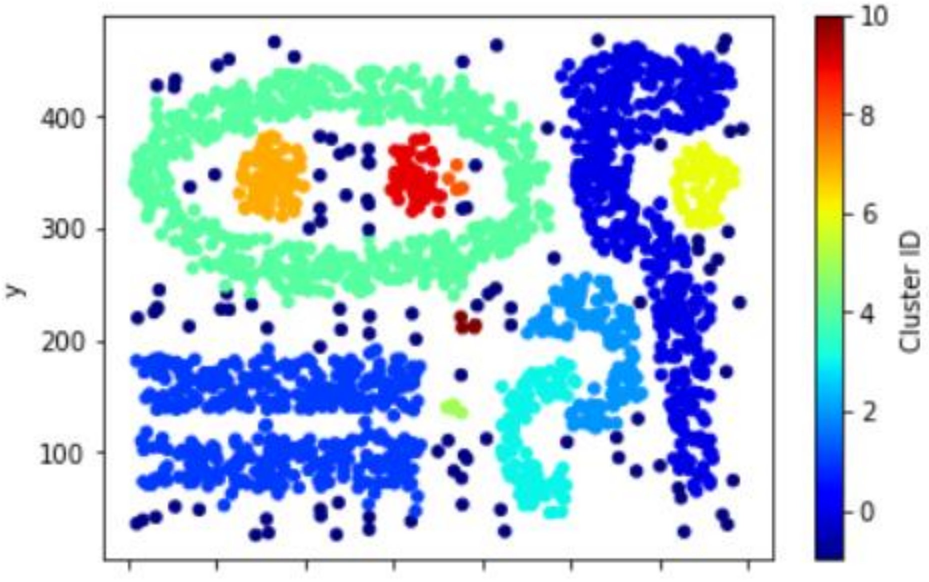
Minimum number of points set to 2

37 clusters are formed. Max: 35 Min: -1



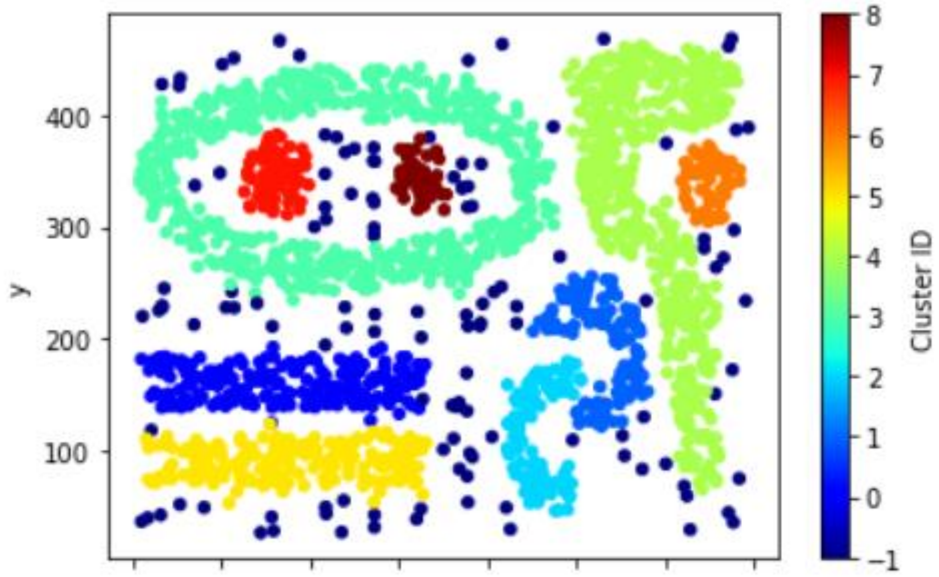
Minimum number of points set to 3

17 clusters are formed. Max: 15 Min: -1



Minimum number of points set to 4

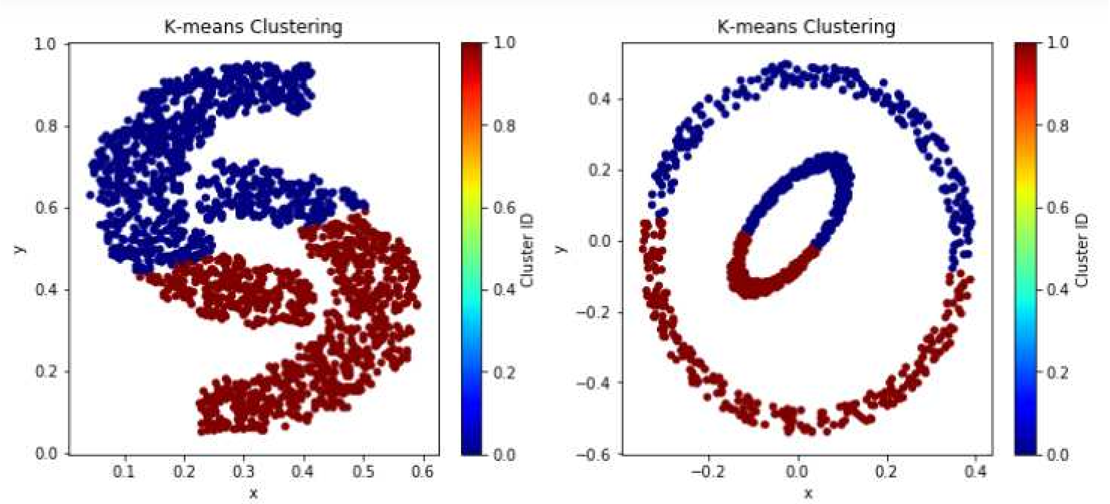
12 clusters are formed. Max: 10 Min: -1



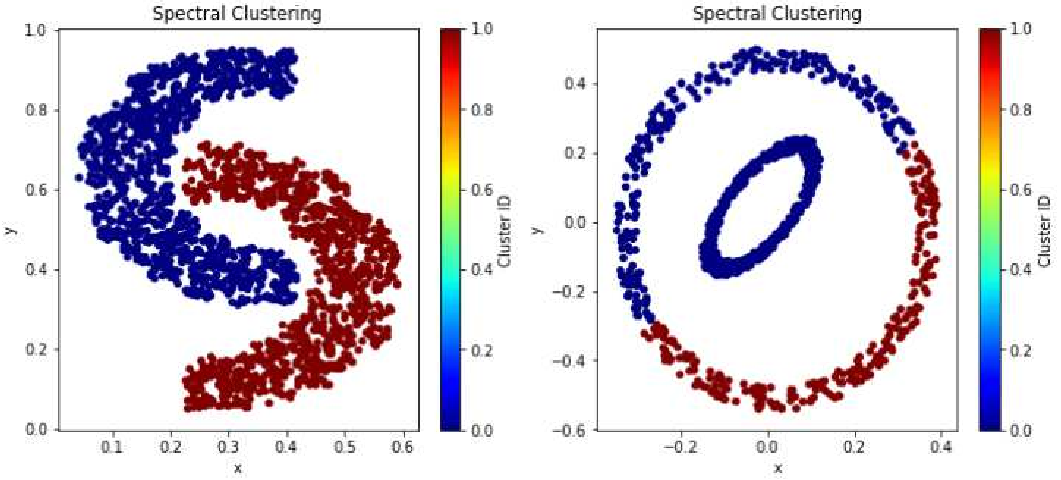
Minimum number of points set to 5

10 clusters are formed. Max: 8 Min: -1

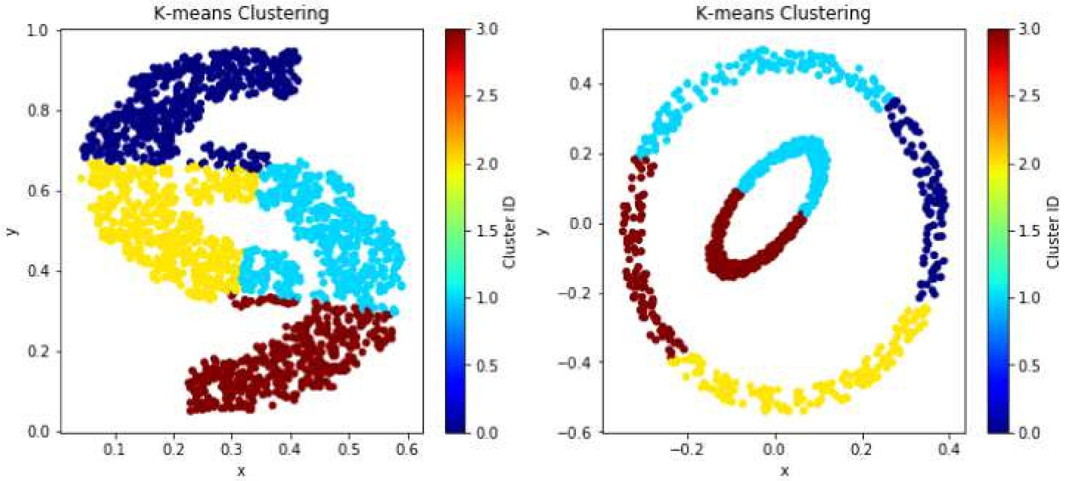
5.



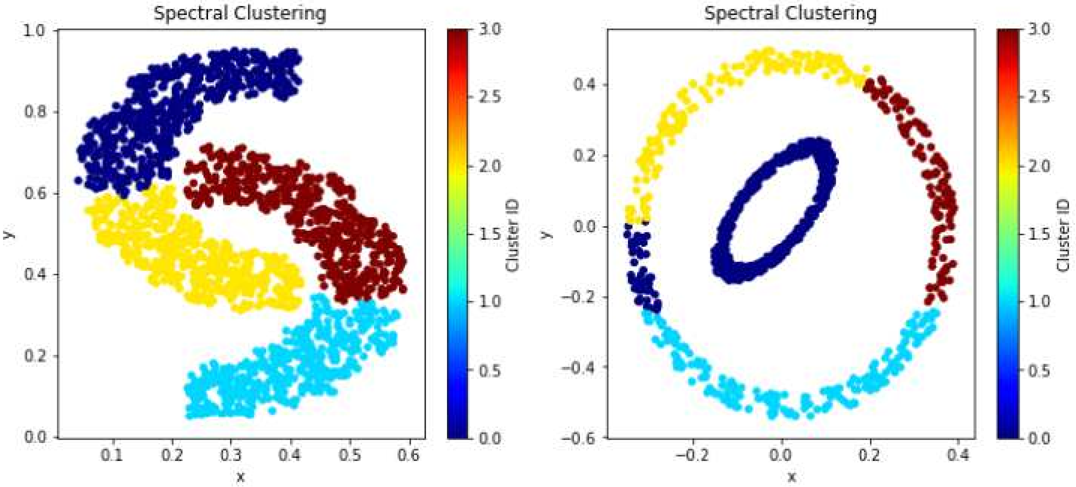
K-means Clustering(k=2)



Spectral Clustering(k=2)



K-means Clustering(k=4)



Spectral Clustering(k=4)

SSE of k-means for data 1 when k = 2: 62.837973033162726
SSE of k-means for data 2 when k = 2: 56.30529204176325
SSE of k-means for data 1 when k = 4: 29.1746326877093
SSE of k-means for data 1 when k = 4: 29.058653796979797

From the SSE for k-means, we can easily find that the k-means performs better for both datasets when k equals to 4 than k equals to 2.

When k = 2, data set = 2D data, the spectral clustering works better than k-means since it clearly separates two parts.

When k = 2, data set = Elliptical data, both algorithms do not separate the inner circle from the outer circle and perform not well.

When k = 4, data set = 2D data, the spectral clustering works better than k-means since it clearly separates four parts -left top, left bottom, right top and right bottom. K-means misclassify the parts in the middle.

When k = 4, data set = Elliptical data, the spectral clustering works better than k-means since it separates the most part of the inner circle from the outer circle. K-means divides the inner circle into 2 parts and each part is the same as part of the outer circle.

In general, the spectral clustering works better than k-means.

Question 4
1.

	Abstract word count	Body word count
Count	24584	24584
Mean	216.45	4435.48
Standard Deviation	137.07	3637.42
Minimum	1	23
Maximum	3694	232431

2.

The data pre-processing steps:

1. Remove the duplicate articles because of some authors submitting the article to multiple journals.
2. Drop Null values.
3. Limit number of articles to speed up computation.
4. Remove punctuation from each text.
5. Convert each text to lower case.

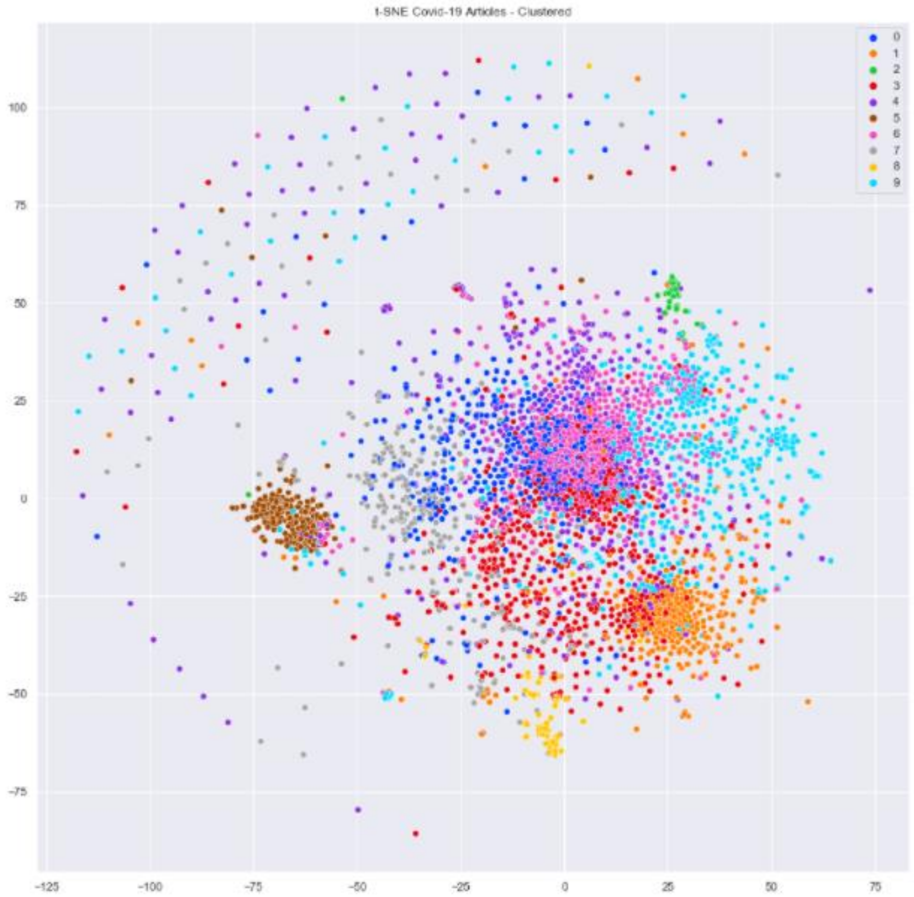
The real-world data is incomplete and inaccurate. So, we need data pre-processing steps to clean up the text and help format and organize the raw data.

3.
We focus on the body text of the articles, more specifically, the n-grams of the body text.

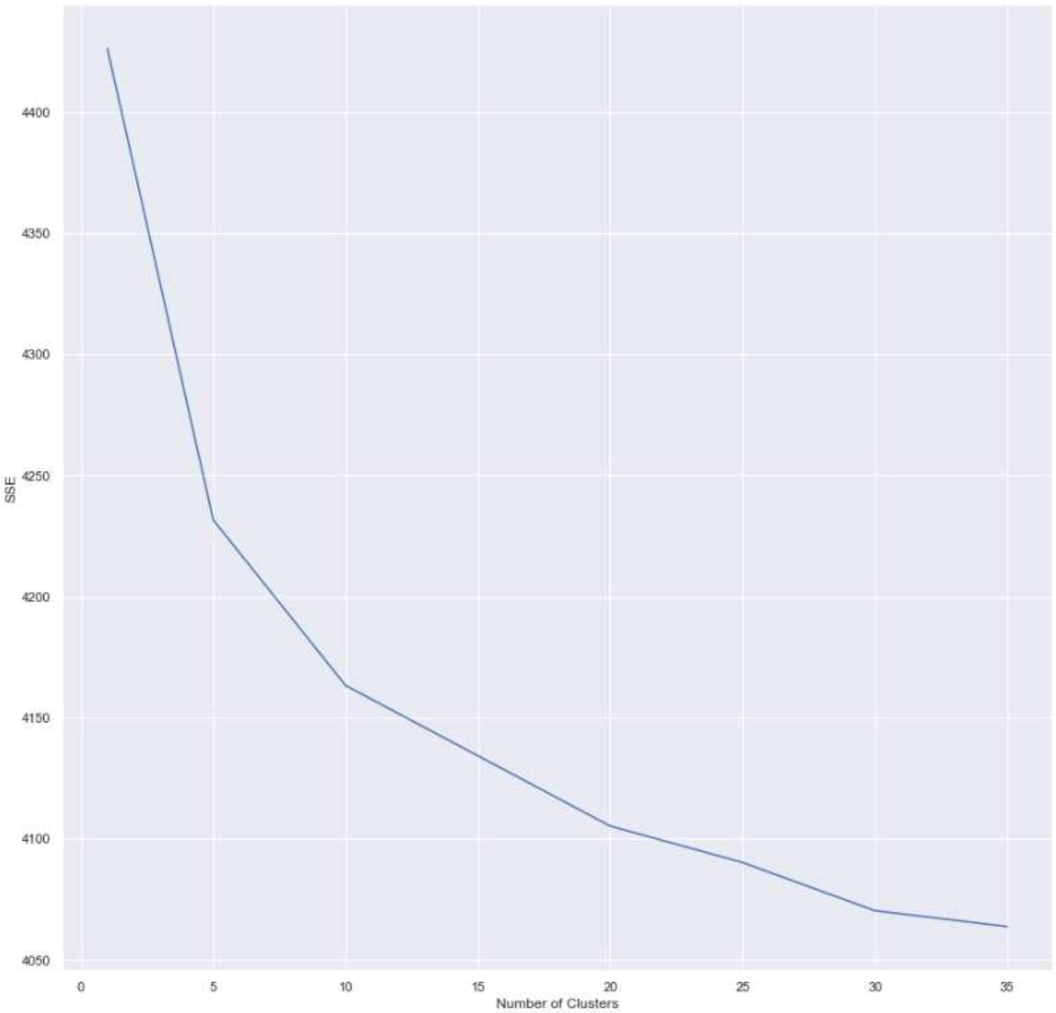
4.
An N-gram means a sequence of N words.
The 2-gram of [’the’, ’2019’, ’novel’, ’coronavirus’, ’sarscov2’, ’identified’, ’as’, ’the’, ’cause’, ’of’] would be:

['the2019', '2019novel', 'novelcoronavirus', 'coronavirussarscov2', 'sarscov2identified', 'identifiedas', 'asthe', 'thecause', 'causeof']
5.
HashingVectorizer is used to create the features vector. The feature size of HashingVector would be 4096.

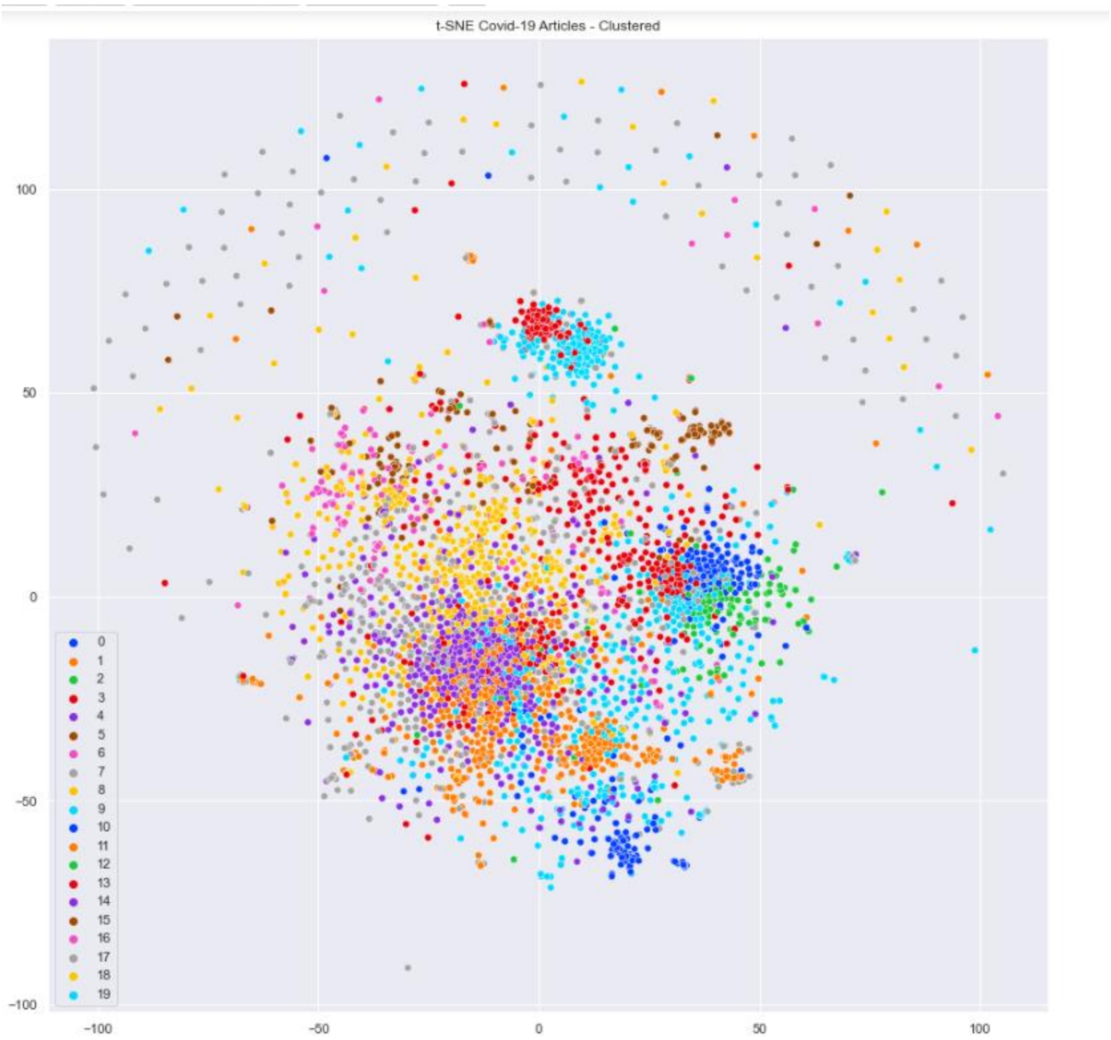
6.



*Feature size = 2**12, k = 10*



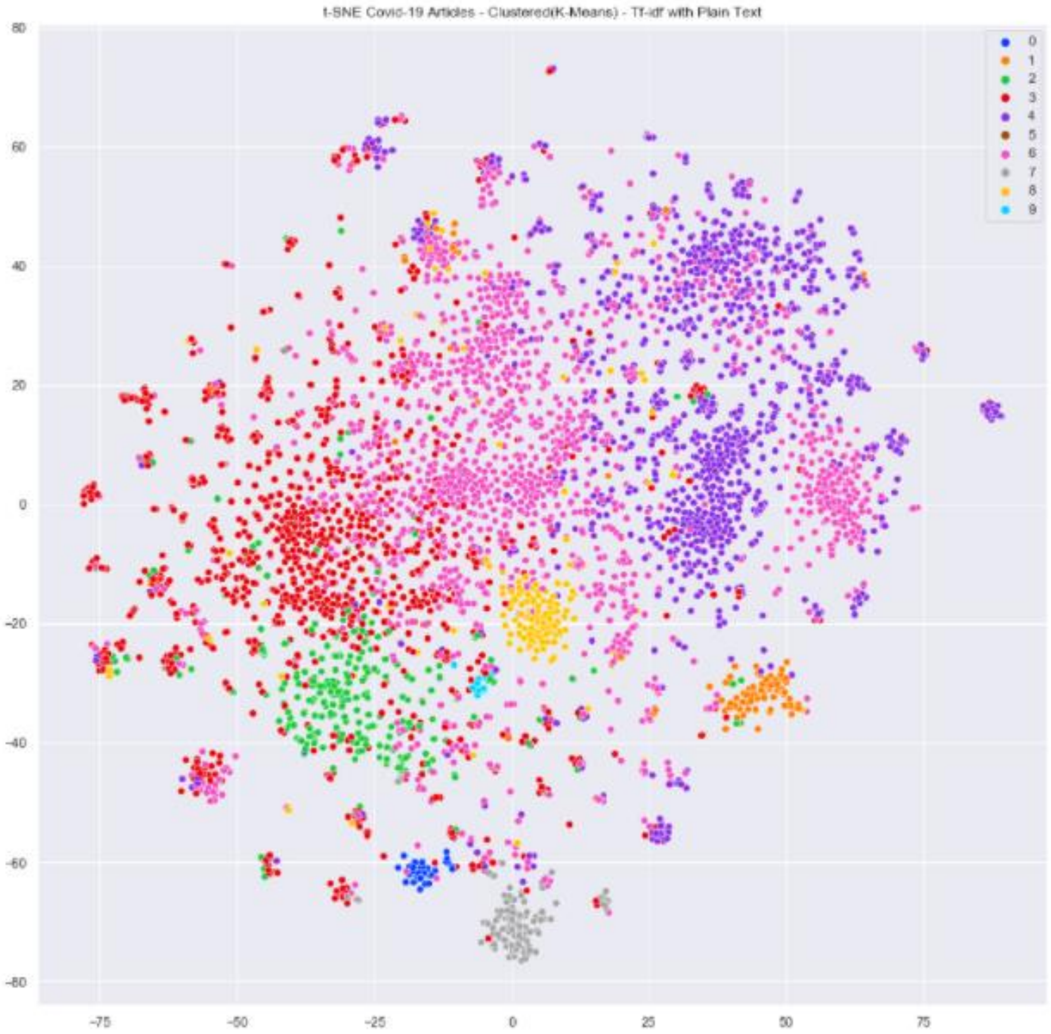
Elbow Method



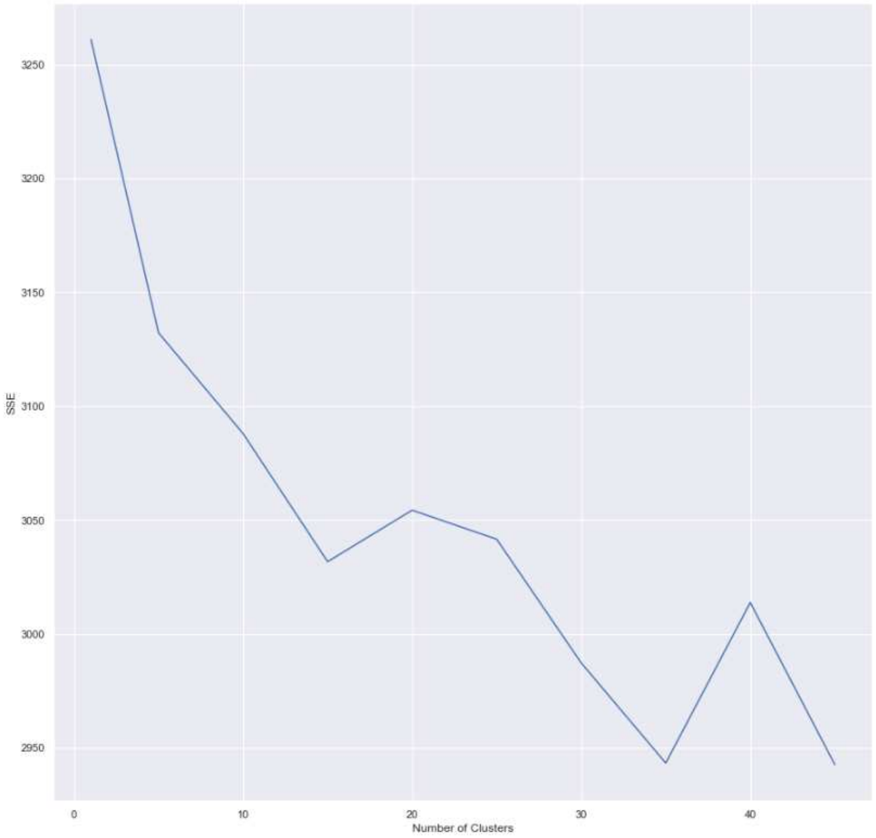
*Feature size = 2**12, k = 20*

I first tried to set the feature size to 2**15 and 2**13 to reduce the collusions and improve the accuracy, but I failed for the lack of the memory. Then I set the feature size to 2**12 and use Elbow Method to determine the k value. From the Elbow Method, it has a smooth line and I picked k = 20. So, I set the feature size to 2**12 and k equals to 20. The result seems better than before but there are still overlaps.

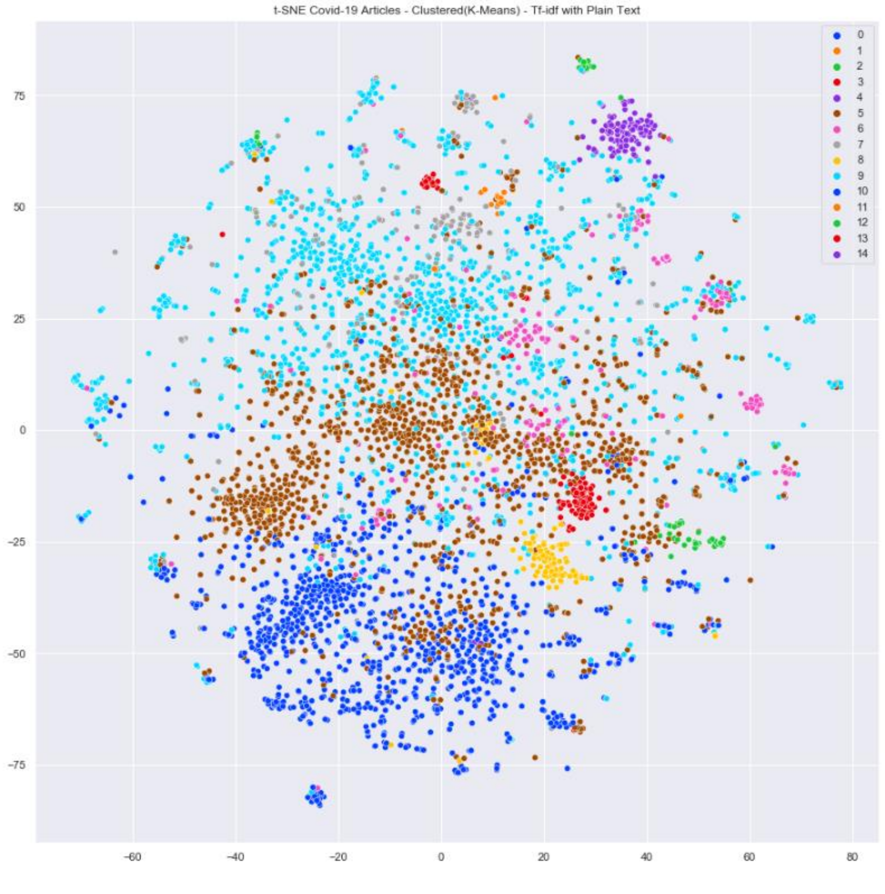
7.



*Feature size = 2**12, k = 10*



Elbow Method



*Feature size = 2**13, k = 15*

I first tried to set the feature size to 2**20 and 2**15 to reduce the collusions and improve the accuracy, but I failed for the lack of the memory. Then I set the feature size to 2**13 and use Elbow Method to determine the k value. From the Elbow Method, I know that when k equals to 15, the decreasing trend slows down. So the feature size is set to 2**13 and k equals to 15 and the result seems better than before.

8.
5 Clusters:
C1: mers, patient, outbreak, human, cell
C2: pneumonia, asthma, influenza, rsv, infant
C3: rna, ibv, strain, protein, cell, structure
C4: mouse, viral, antibody, immune, gene
C5: virus, cell, antiviral, inhibitor, plant

There are many clusters containing the social and economic impacts of the coronavirus, I pick top 5 clusters which contain the greatest number of articles about social and economic impacts. There are: C12,C8, C18,C11 and C3.