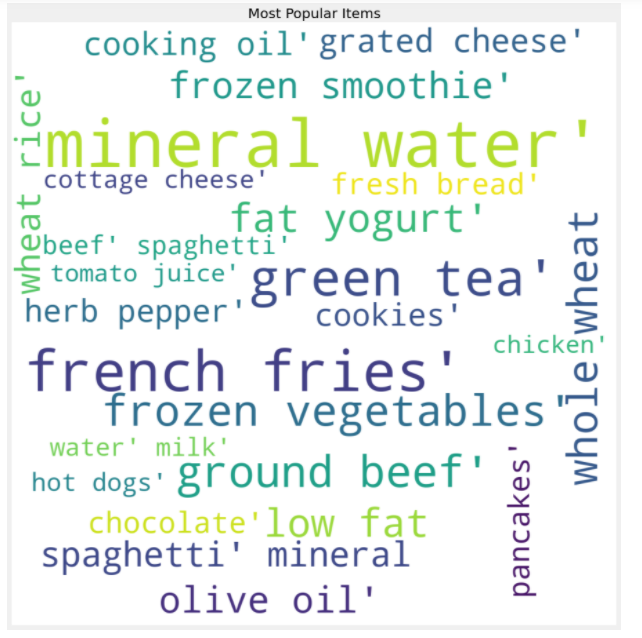
Question 1

1. There are 7501 records in the dataset.
2. The maximum number of items a customer has bought is 20.

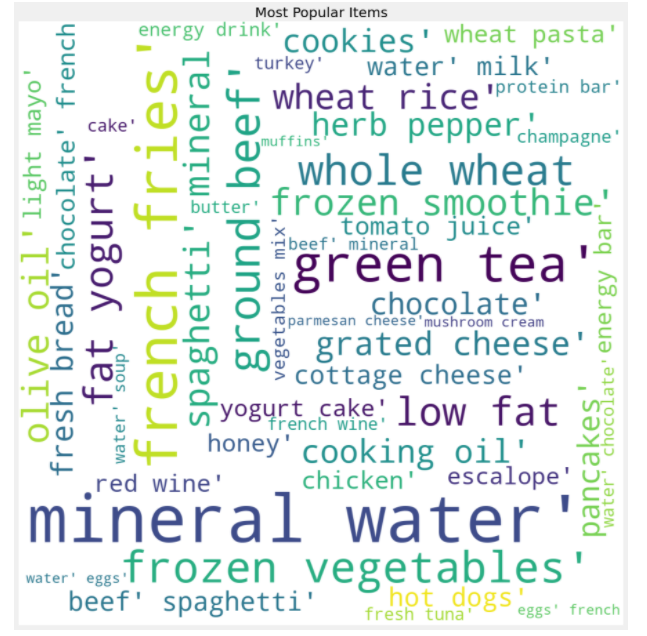
3.

|  |  |
| --- | --- |
| 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*

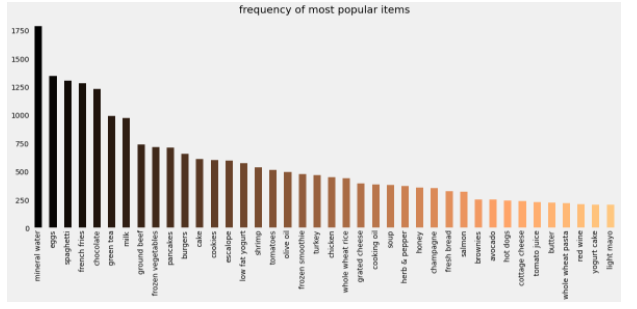
**

*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]

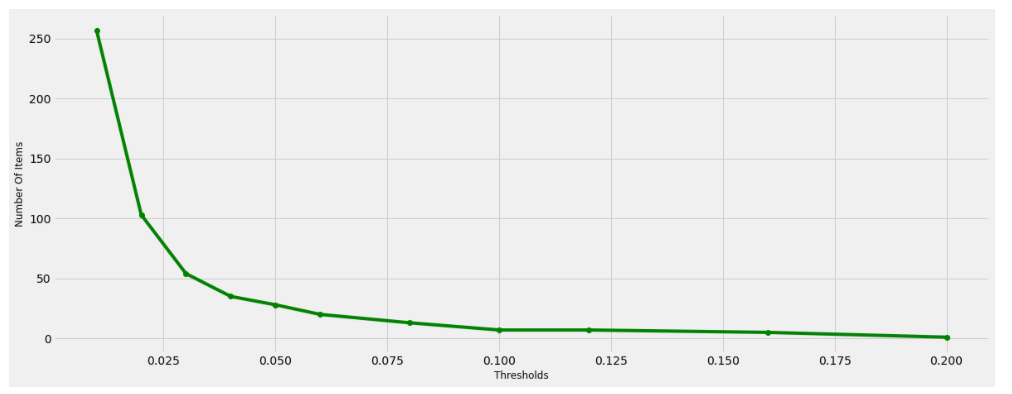
[False False True True True True]

[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.

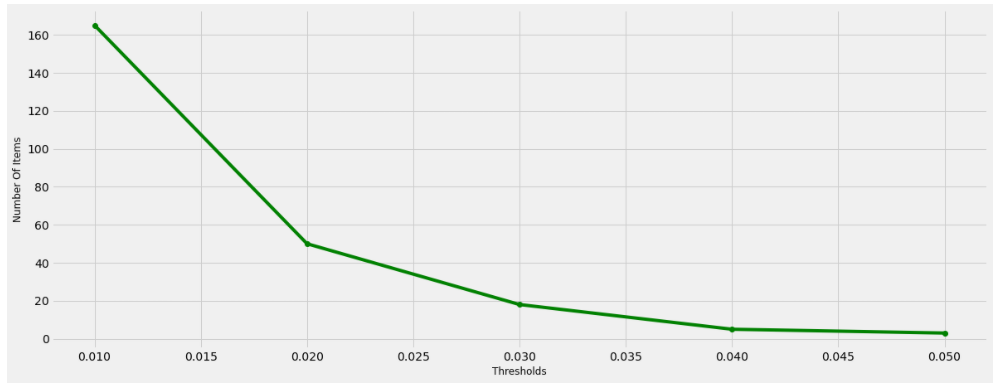
8. 

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

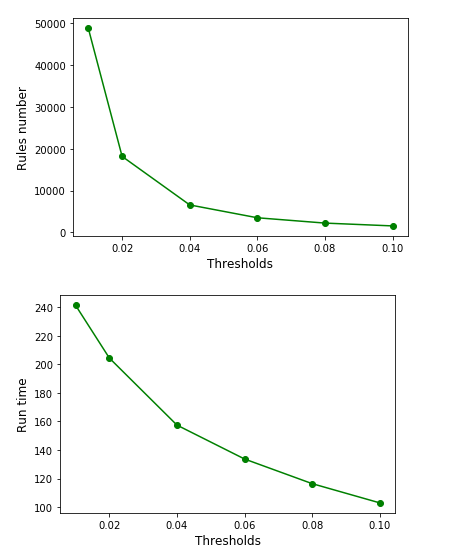
{Egg, carrot}: 1

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.

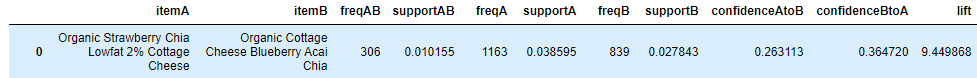
3.



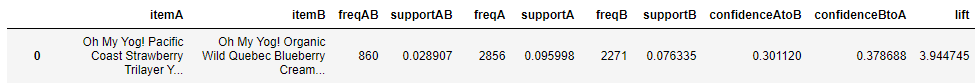
When the support threshold is increasing, the runtime is decreasing. Support threshold means that only the itemsets with support higher than or equal to the support threshold would be saved. So, when we increase the number of support threshold, fewer itemsets are frequent itemsets and saved by the Apriori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

4.

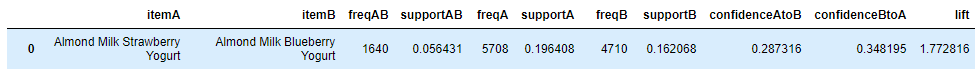
When support threshold is 0.01:



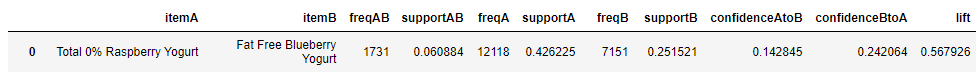
When support threshold is 0.02:



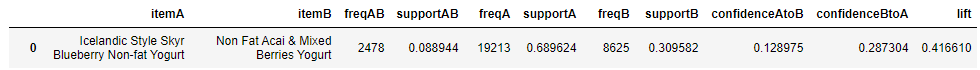
When support threshold is 0.04:



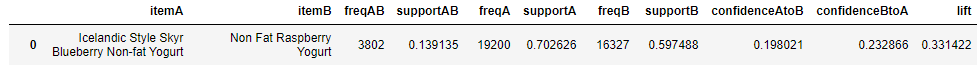
When support threshold is 0.06:



When support threshold is 0.08:



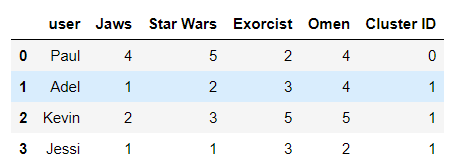
When support threshold is 0.1:



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

Question 3

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.

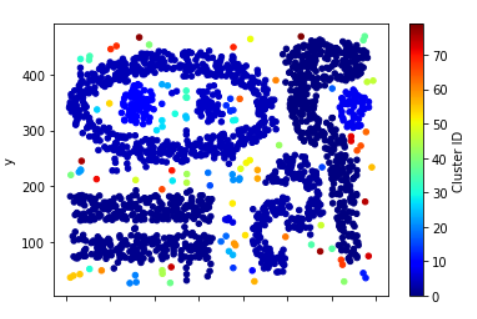
For Single link, it links two clusters based on the minimum distance between 2 elements and it tends to produce long thin clusters since the link based on only 2 points.

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.

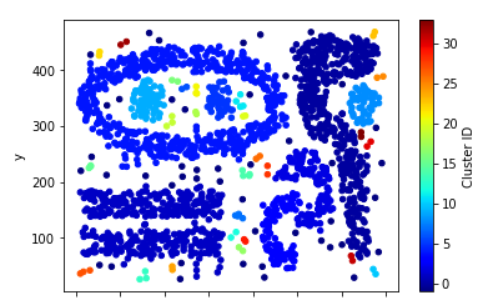
From the results, we know that the average link is the best among these three hierarchical clustering algorithms. The average link balances the breadth and depth and gives us a reasonable clustering.

4.



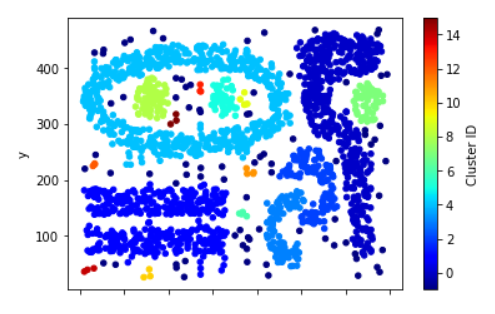
Minimum number of points set to 1

78 clusters are formed. Max: 77 Min: 0



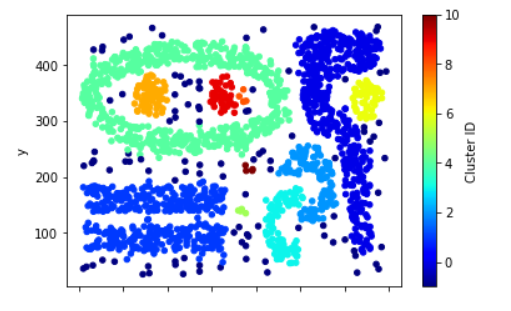
Minimum number of points set to 2

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



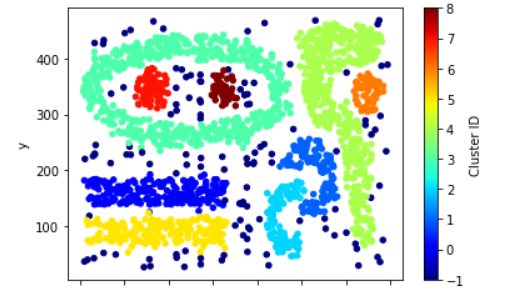
Minimum number of points set to 3

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



Minimum number of points set to 4

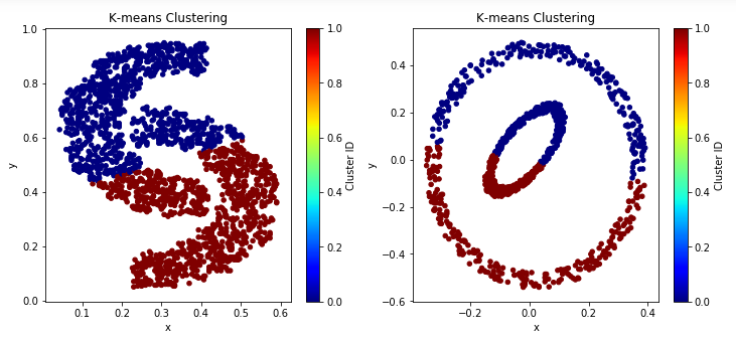
12 clusters are formed. Max: 10 Min: -1



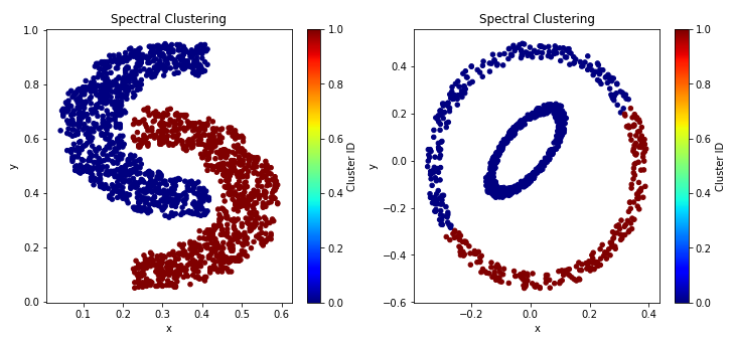
Minimum number of points set to 5

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

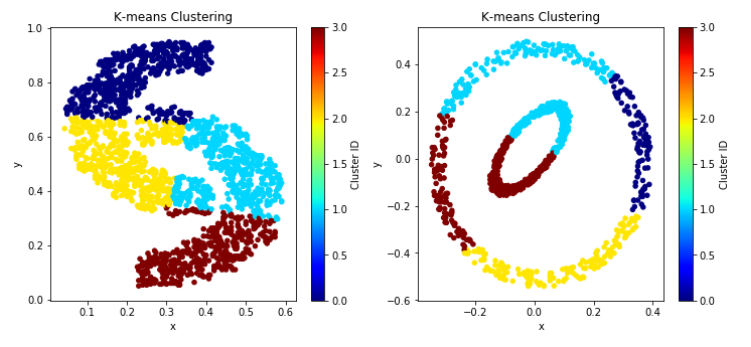
5.



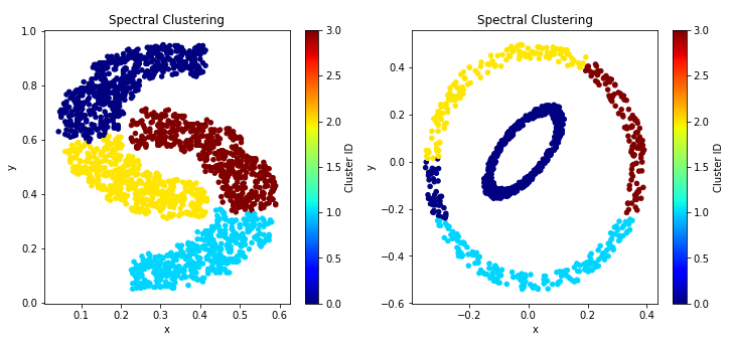
K-means Clustering(k=2)



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

The k-mean clustering divided the inner circle to two different clusters, each has a large part of the outer circle. So, we could know that when k-4 the Spectral Clustering work better on the elliptical data.

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.

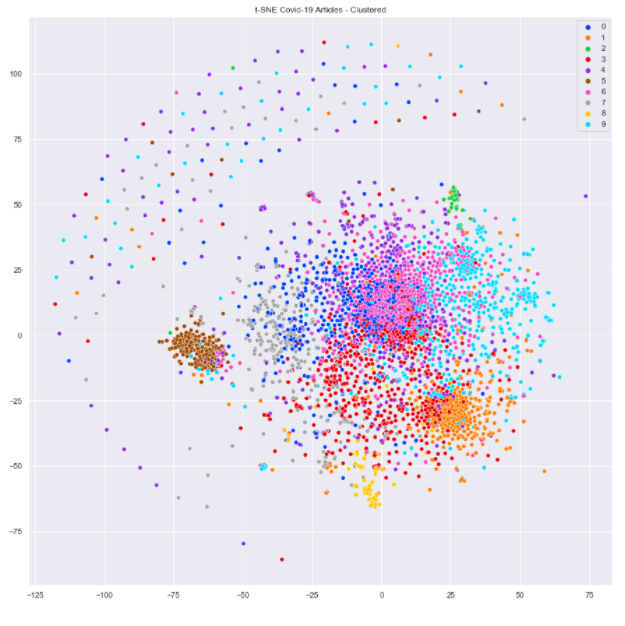
The 2-gram of [’the’, ’2019’, ’novel’, ’coronavirus’, ’sarscov2’, ’identified’, ’as’, ’the’, ’cause’, ’of’] would be:

['the2019', '2019novel', 'novelcoronavirus', 'coronavirussarscov2', 'sarscov2identified', 'identifiedas', 'asthe', 'thecause', 'causeof']

5.

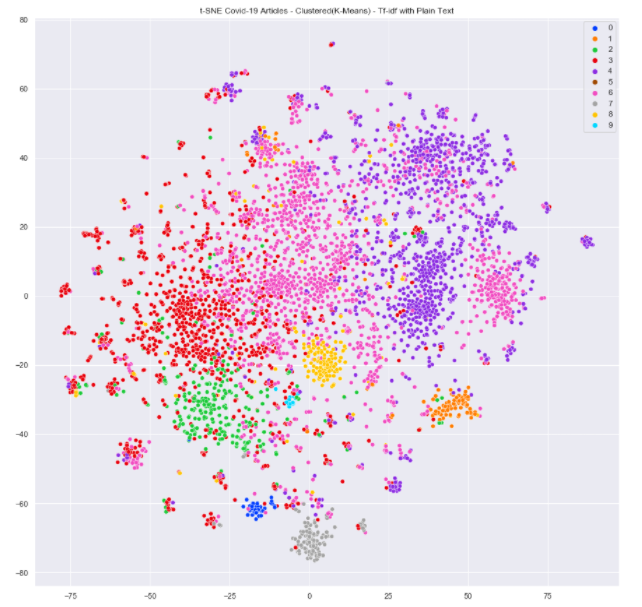
HashingVectorizer is used to create the features vector. The feature size of HashingVector would be 4096.

6.

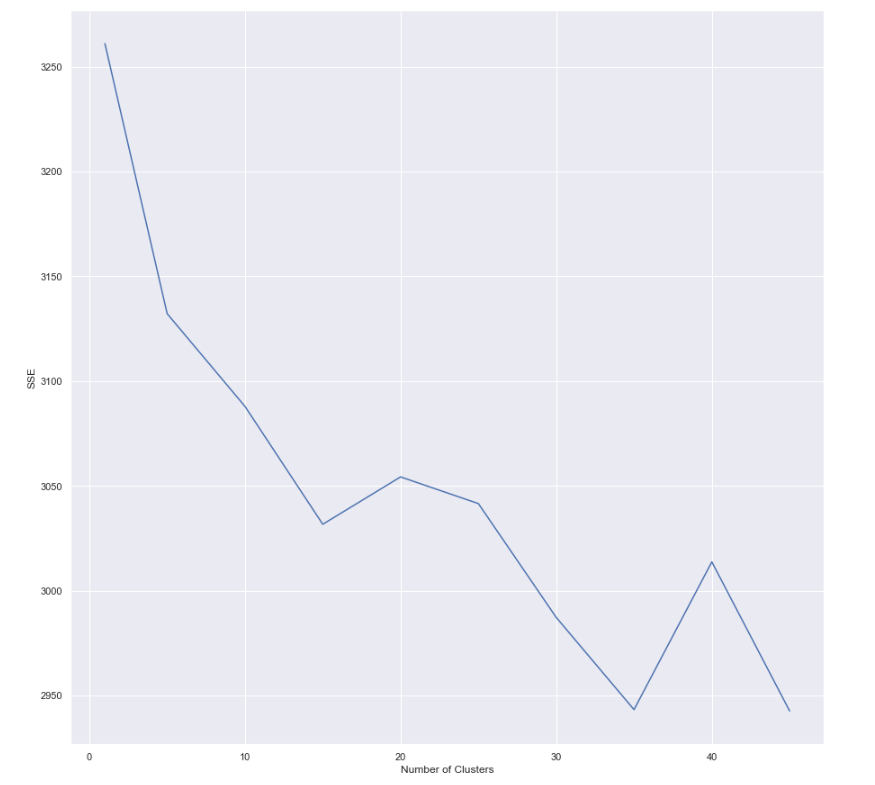


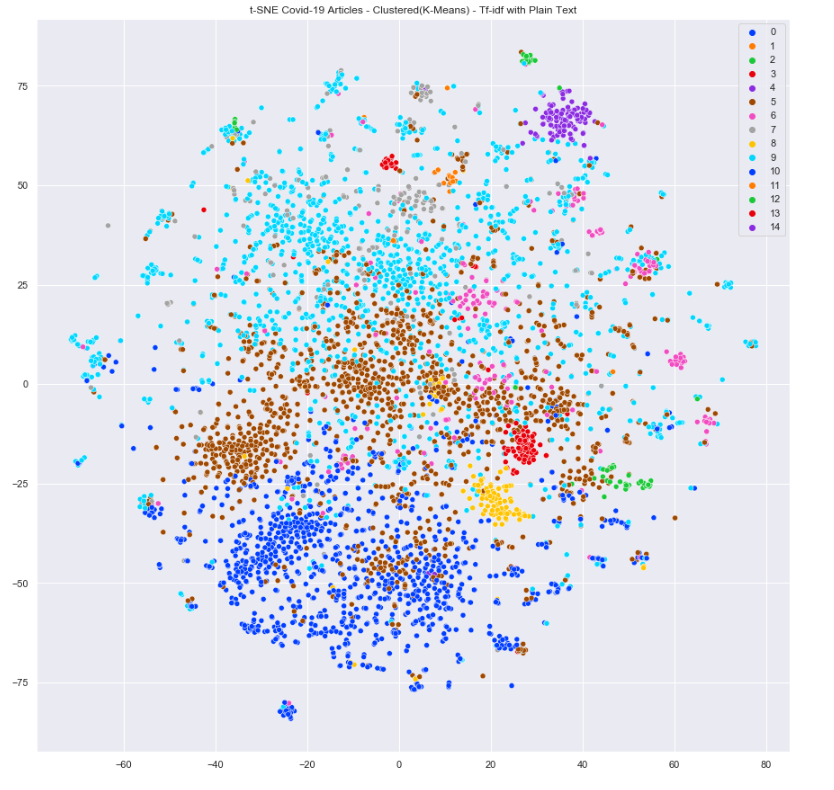
Feature size = 4096, k = 10

7.



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





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

8.