HYUNWOO KIM kim00186 CSCI 5523 Project2

Problem 1[25 points]

The files for this problem is under Experiment 1 folder. Datasets to be used for experimentation: store transaction.csv. Jupyter notebook: Apriori analysis.ipynb. In this experiment, we give a dataset of a store with thousands of transactions of customers buying several items from the store. We will use the Apriori algorithm to find correlations between various items in the store. Answer the following question :

1. How many records are there in the dataset?

➔ 7501 records.

2. In a single transaction, what is the maximum number of items a customer has bought? We assume that each record is a separate transaction

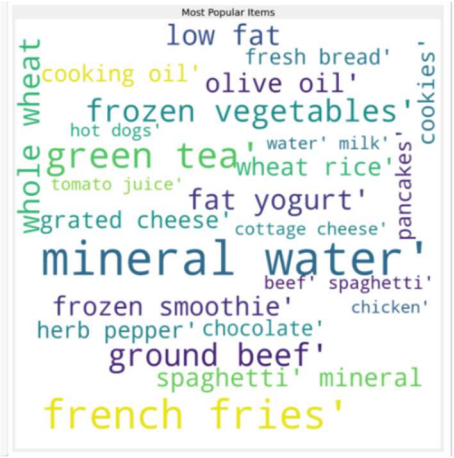
➔ 20 items.

3. Write any five transactions a customer has done.

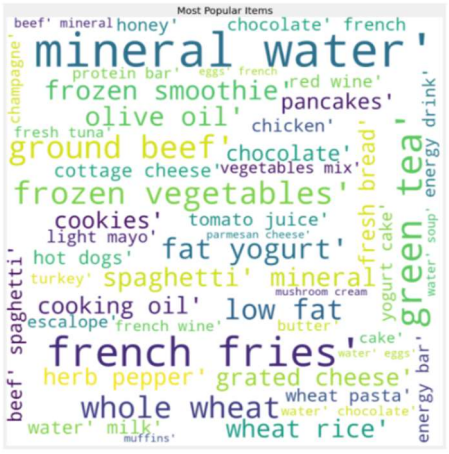
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
3578	burgers	shrimp	frozen vegetables	spaghetti	mineral water	chocolate	olive oil	french wine	pancakes	white wine	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2942	fresh tuna	ham	mineral water	avocado	meatballs	olive oil	almonds	eggs	french fries	frozen smoothie	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
318	spaghetti	mineral water	chicken	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
5606	shrimp	frozen vegetables	ground beef	spaghetti	mineral water	strong cheese	green beans	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
3813	low fat yogurt	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

4. We use the wordcloud to generate a stunning visualization format to highlight crucial textual data points and convey essential information. Generate and paste the wordcloud with max words set to 25 and 50. Briefly describe your understanding of the plot.

➔ Max words = 25



Max words = 50



In the order in which the customer purchases the most(frequency), each tag changed its shape, such as the color and thickness of the letters. Mineral water, French fries, and the other big letter could be seen that those big font items have a large number of customers has bought relatively among the items.

5. What are the top 5 most frequent items in the dataset?

➔ Mineral water, eggs, spaghetti, French fries, and chocolate.

6. Suppose we have the following transaction data: [['Apple', 'Beer', 'Rice', 'Chicken'], ['Apple', 'Beer', 'Rice'], ['Apple', 'Beer'], ['Apple', 'Bananas'], ['Milk', 'Beer', 'Rice', 'Chicken'], ['Milk', 'Beer', 'Rice'], ['Milk', 'Beer'], ['Apple', 'Bananas']]. Transform this input dataset into a one-hot encoded Boolean array. Hint: In the Jupyter notebook, we use TransactionEncoder to do the same.

```
In [47]: from mlxtend.preprocessing import TransactionEncoder

         dataset = [['Apple', 'Beer', 'Rice', 'Chicken'],
                    ['Apple', 'Beer', 'Rice'],
                    ['Apple', 'Beer'],
                    ['Apple', 'Bananas'],
                    ['Milk', 'Beer', 'Rice', 'Chicken'],
                    ['Milk', 'Beer', 'Rice'],
                    ['Milk', 'Beer'],
                    ['Apple', 'Bananas']]
```

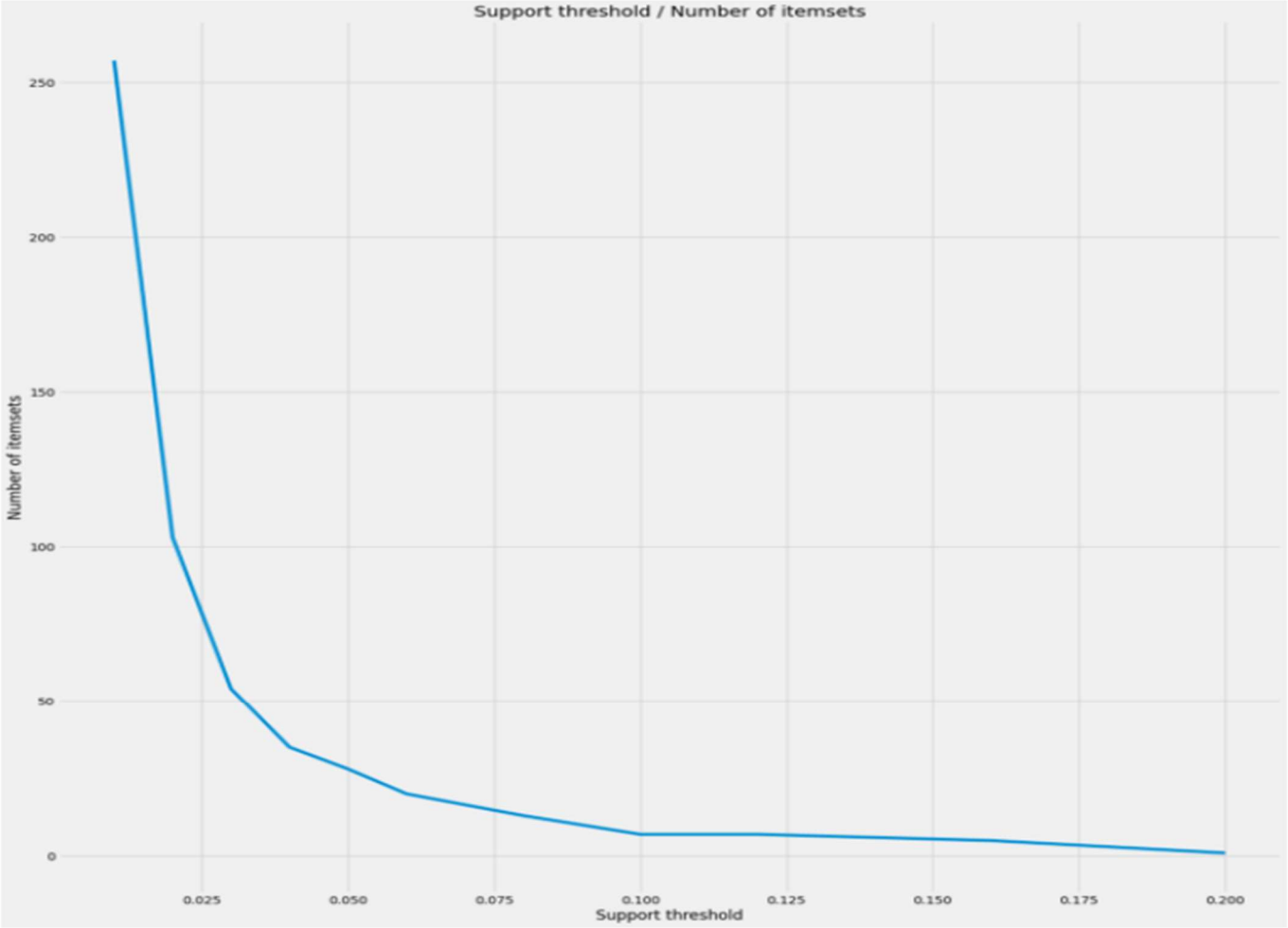
```
In [48]: te = TransactionEncoder()
         te_ary = te.fit(dataset).transform(dataset)
         te_ary

Out[48]: array([[ 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. In the input dataset, how many unique items are present?

➔ 121 unique items

8. Run Apriori to generate frequent itemsets at support thresholds of 1%, 2%, 3%, 4%, 5%, 6%, 8%,10%,12%,16% and 20%. In a single figure, for each threshold (X-axis), plot the number of itemsets (Y-axis). Comment on the general trends illustrated by the plots and the reason for the trend.

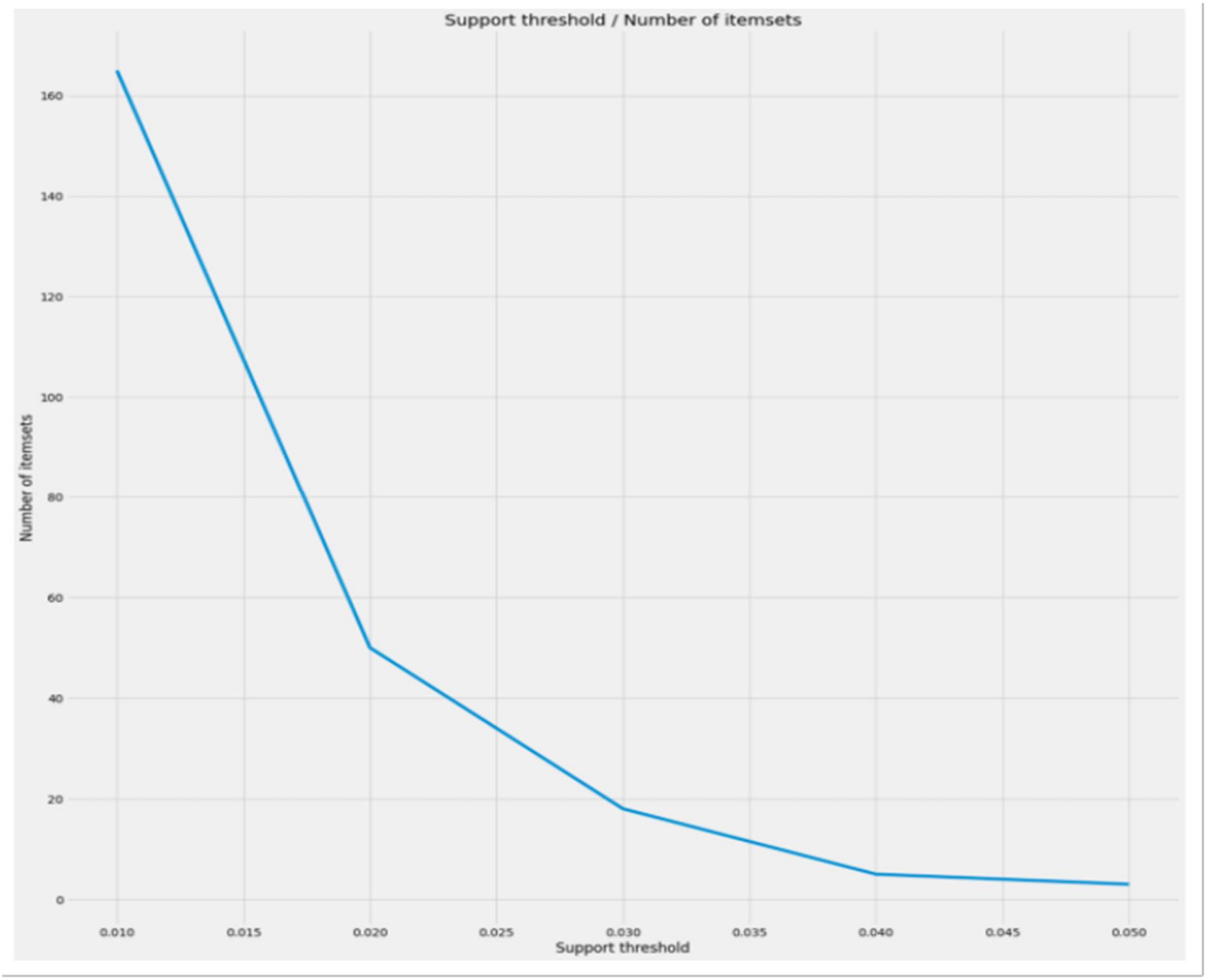


➔ As the supporting stress hold grows, the number of items set is decreasing. This is because the Apriori algorithm is to increase the number of support thresholds and continue to look for more frequently used item sets.

9. At support threshold 1%, we see frequent itemset of size 3 along with size 2 and 1. However, at the support threshold of 2%, we observe itemsets of size 1 and 2 only. Why do you think this is so?

➔ This is because support threshold determines the minimum support count. To be more specific, when at support threshold 1%, if the item frequency is not over 1%, then it does not count as a frequent. There is no frequent itemset of size 3 when at support threshold of 2 %.

10. Run Apriori to generate frequent itemsets of length 2 at support thresholds of 1%, 2%, 3%, 4% and 5%. In a single figure, for each threshold (X-axis), plot the number of itemsets of length 2 (Y-axis). Comment on the general trends illustrated by the plots and the reason for the trend.



➔ As the supporting stress hold grows, the number of items set is decreasing. This is because the Apriori algorithm is to increase the number of support thresholds and continue to look for more frequently used item sets.

11. For the following itemset, write down its corresponding support value:

- Mineral Water = 0.238
- Chocolate = 0.1638
- Eggs = 0.1797
- Eggs, Mineral Water = 0.0509
- Chocolate, Mineral Water = 0.05266

Problem 2 [25 points]

The files for this problem are under Experiment 2 folder. Datasets to be used for experimentation: Instacart_transaction.csv. Jupyter notebook: Instacart association.ipynb. Instacart, an online grocer, has graciously made some of their datasets accessible to the public (https://www.instacart.com/datasets/grocery-shopping-2017).

In this experiment, we will use Apriori algorithm to find correlations between the different items in the store. Answer the following question:

1. Given following transactions:

• order 1: apple, egg, milk

• order 2: carrot, milk

• order 3: apple, egg, carrot

• order 4: apple, egg

• order 5: apple, carrot

Using the Apriori algorithm, write down the pair of items having a minimum threshold of 3.

Briefly describe your steps.

➔ {apple, egg}

First, we need to count the number of times each item occurs {apple} 4, {egg} 3, {milk} 2, {carrot} 3 and delete {milk} 2 because it does not meet minimum threshold of 3.

Second, build item sets of size 2 using the remaining items from the first step. {apple, egg} 3, {apple, carrot}: 2, {milk, carrot}: 1 and delete those do not meet minimum threshold of 3.

Then, {apple, egg} this item set is only the set in the remining items from second step.

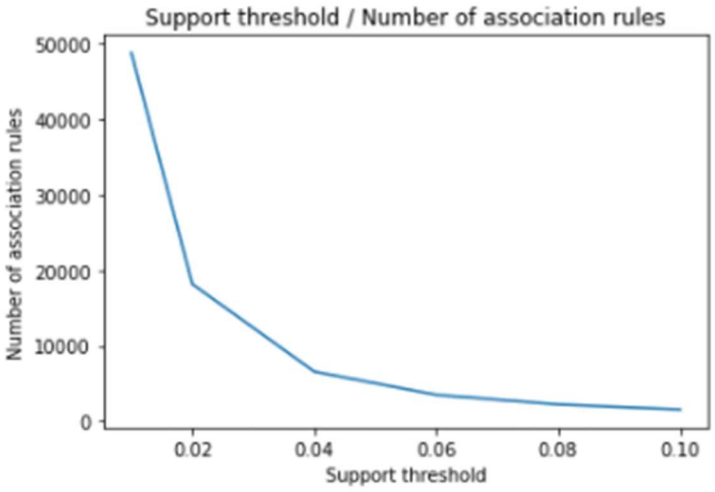
2. How many unique orders and unique items are there in the dataset? Are unique order same as number of records in the dataset? On average, per order, how many items does a customer order?

➔ 3214874 unique orders and 49677 unique items.

➔ 32434489 records and 3214874 orders not same.

➔ 10.1 items per order on average. = Dimensions / unique orders

3. Run Apriori to generate pairs of itemset at support thresholds of 1%, 2%, 4%, 6%, 8%, and 10%. In a single figure, for each threshold (X-axis), plot the number of association rules (Y-axis). Comment on the trend of algorithm runtime at different thresholds.



➔ Algorithm runtime decreased by the support thresholds increase. This is granted by that if the support thresholds increase, the number of frequent itemset will be decreased by the Apriori algorithm.

4. Run Apriori at support thresholds of 1%, 2%, 4%, 6%, 8% and 10%. For each threshold, write a pair of association rules (you can choose any) along with its key metrics (i.e. freqAB, supportAB, freqA,supportA, freqB, support, con_denceAtoB, con_denceBtoA, lift). As a data scientist for the retailer giant, after observing the association rules, what would you do to increase the sales.

➔ 1%

	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
1	Grain Free Chicken Formula Cat Food	Grain Free Turkey Formula Cat Food	318	0.010553	1809	0.060033	879	0.029170	0.175788	0.361775	6.026229
3	Organic Fruit Yogurt Smoothie Mixed Berry	Apple Blueberry Fruit Yogurt Smoothie	349	0.011582	1518	0.050376	1249	0.041449	0.229908	0.279424	5.546732

➔ 2%

	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
2	Unsweetened Blackberry Water	Raspberry Essence Water	660	0.022184	3108	0.104468	2025	0.068066	0.212355	0.325926	3.119850
3	Organic Fiber & Protein Pear Blueberry & Spina...	Fiber & Protein Organic Pears, Raspberries, Bu...	606	0.020369	2782	0.093511	2167	0.072839	0.217829	0.279649	2.990560

➔ 4%

	itemA	itemB	freqAB	supportAB	freqA	supportA	freqB	supportB	confidenceAtoB	confidenceBtoA	
0	Milk Strawberry Yogurt	Milk Blueberry Yogurt	1640	0.056431	5708	0.196408	4710	0.162068	0.287316	0.348195	1.772
1	Almond Milk Peach Yogurt	Almond Milk Blueberry Yogurt	1289	0.044354	4703	0.161827	4710	0.162068	0.274080	0.273673	1.691
2	Almond Milk Strawberry Yogurt	Almond Milk Peach Yogurt	1376	0.047347	5708	0.196408	4703	0.161827	0.241065	0.292579	1.489

➔ 6%

	itemA	itemB	freqAB	supportAB	freqA	supportA	freqB	supportB	confidenceAtoB	confidenceBtoA	lif
0	Total 0% Raspberry Yogurt	Fat Free Blueberry Yogurt	1731	0.060884	12118	0.426225	7151	0.251521	0.142845	0.242064	0.567926
3	Blackberry Cucumber Sparkling Water	Kiwi Sandia Sparkling Water	1863	0.065527	11132	0.391544	9097	0.319968	0.167355	0.204793	0.523039
4	Grapefruit Sparkling Water	Lemon Sparkling Water	2164	0.076114	14528	0.510991	9318	0.327741	0.148954	0.232239	0.454487

➔ 8%

	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
1	Icelandic Style Skyr Blueberry Non-fat Yogurt	Nonfat Icelandic Style Strawberry Yogurt	2699	0.096877	19213	0.689624	10636	0.381764	0.140478	0.253761	0.367970
2	Icelandic Style Skyr Blueberry Non-fat Yogurt	Non Fat Raspberry Yogurt	3802	0.136467	19213	0.689624	16340	0.586501	0.197887	0.232681	0.337402

➔ 10%

	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
3	Non Fat Raspberry Yogurt	Icelandic Style Skyr Blueberry Non-fat Yogurt	3422	0.125228	16327	0.597488	19200	0.702626	0.209591	0.178229	0.298297
4	Vanilla Skyr Nonfat Yogurt	Icelandic Style Skyr Blueberry Non-fat Yogurt	3483	0.127461	18070	0.661273	19200	0.702626	0.192750	0.181406	0.274329

➔ We can see that the top associations which customers have bought such items together such as Strawberry Chia, Cottage Cheese with Blueberry Acai Cottage Cheese, Chicken Cat Food with Turkey Cat Food when support threshold is 1%. Shortly, in an association rule, if we find the positive relationships between items, that will be great to increase sales by making item A and B be a package item that must purchasable if a client need to buy a separate A and B. This is the purpose of using the Apriori algorithm to produce the most frequent itemset among the whole itemset.

Problem 3 [20 points]

The files for this problem is under Experiment 3 folder. Datasets to be used for experimentation: 2d data, chameleon, elliptical, and vertebrate. Jupyter notebook: cluster analysis.ipynb. Cluster analysis seeks to partition the input data into groups of closely related instances so that instances that belong to the same cluster are more similar to each other than to instances that belong to other clusters. In this experiment, we provide examples of using different clustering techniques provided by the scikit-learn library package.

Answer the following question :

1. In the notebook, k-mean clustering assign users to two clusters i.e., cluster one has a higher rating for action movies, and cluster two has higher ratings for horror movies.

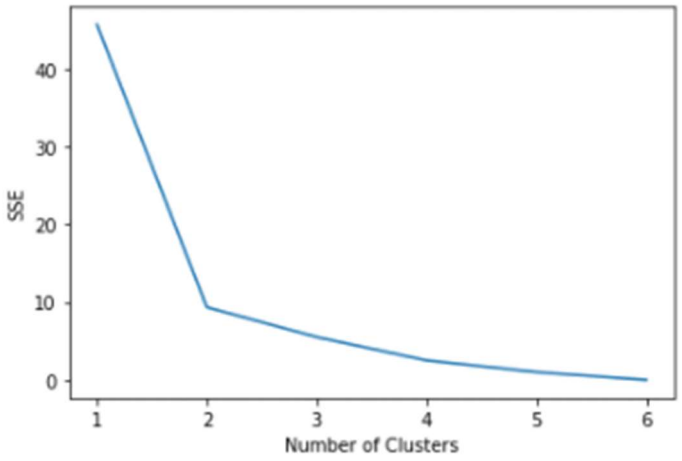
Given the cluster centroid, assign the following users to their respective cluster assignment:

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

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

User	Cluster ID
Paul	1
Adel	0
Kevin	0
Jessi	0

2. A fundamental step for any unsupervised algorithm is to determine the optimal number of clusters into which the data can be clustered. The Elbow Method is one of the most popular methods to find the optimal value of k. For the toy example of the movie rating dataset, what is the optimal value of K? Briefly explain your reasoning.



The performance of the K-means cluster depends on the choice of the initial center point, which requires a

performance metric to perform the test while continuously changing the center point. We are not looking for the smallest SSE here using the Elbow method. SSE is sum of squared error which represent for each point, the error is the distance to the nearest cluster center.

If you look at the graph, you can see that the "elbow" appears when we have 2 clusters. Thus, the optimal value of K is 2 to 6. After K is equals to 6, SSE expected to be stable.

3. 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. As a data scientist, given the class of the original dataset, which clustering algorithm makes more sense. Explain your reasoning.

According to each hierarchical clustering algorithms, each defines Inter-Cluster Similarity:

Min – We can see the proximity of two clusters is based on the two closest points in the different clusters

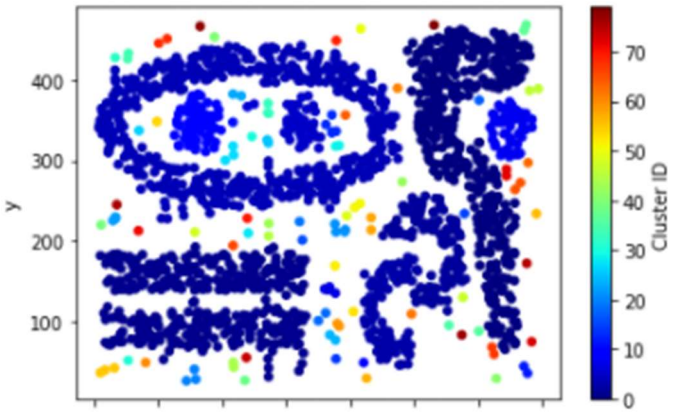
Max - We can see the proximity of two clusters is based on the two most distant points in the different clusters

Group average - We can see the proximity of two clusters is the average of pairwise proximity between points in the two clusters.

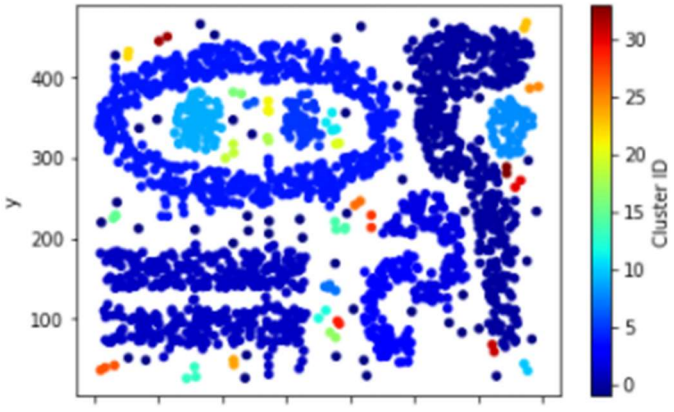
- ➔ Group average is proper in this case because it calculates the average of pairwise proximity between clusters. The figure demonstrates more clearly how make them clusters than Min and Max method.

4. For DBSCAN, how many clusters are formed when the minimum number of points (min samples) to set to 1, 2, 3, 4, and 5, respectively. For each instance, copy and paste the plot of the clusters.

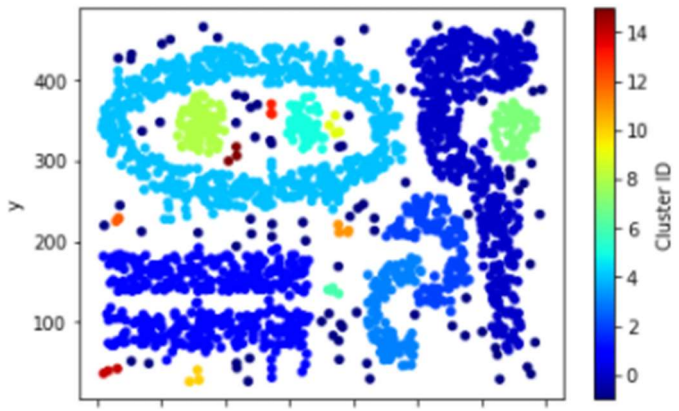
Minimum number of point = 1, 80 clusters



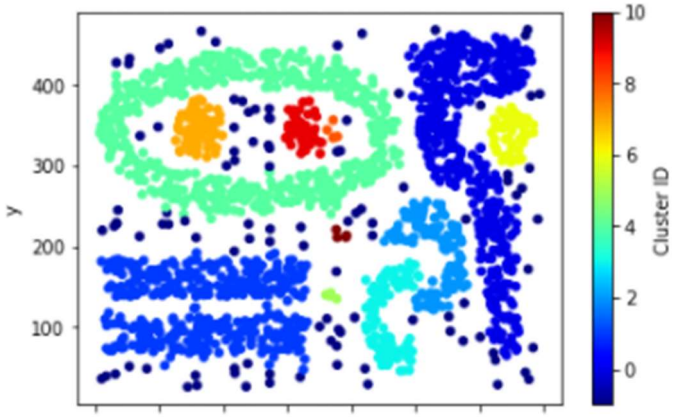
Minimum number of points = 2, 34 clusters



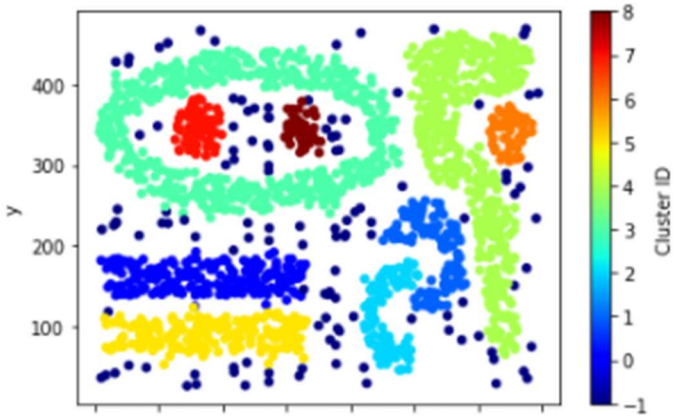
Minimum number of points = 3, 16 clusters



Minimum number of points = 4, 11 clusters

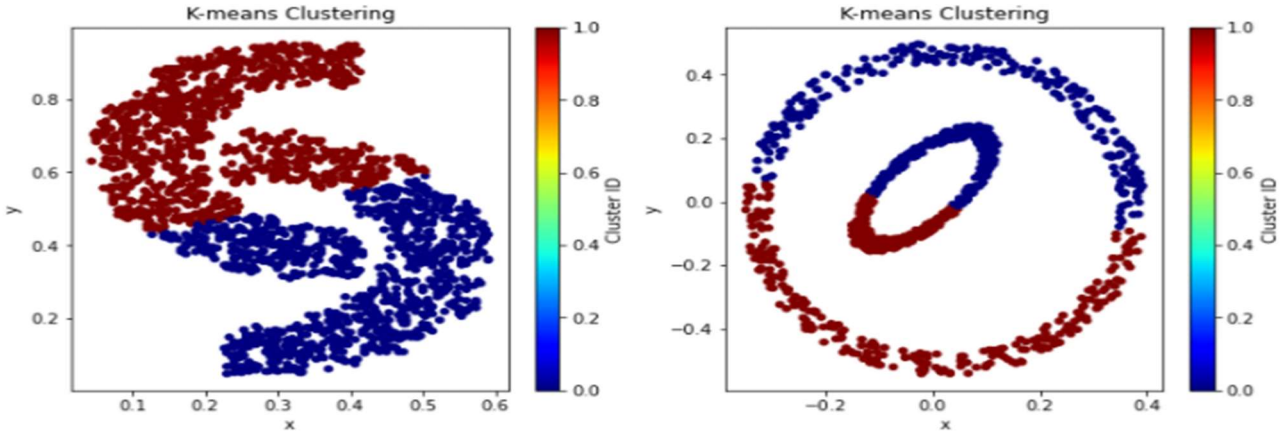


Minimum number of points = 5, 9 clusters



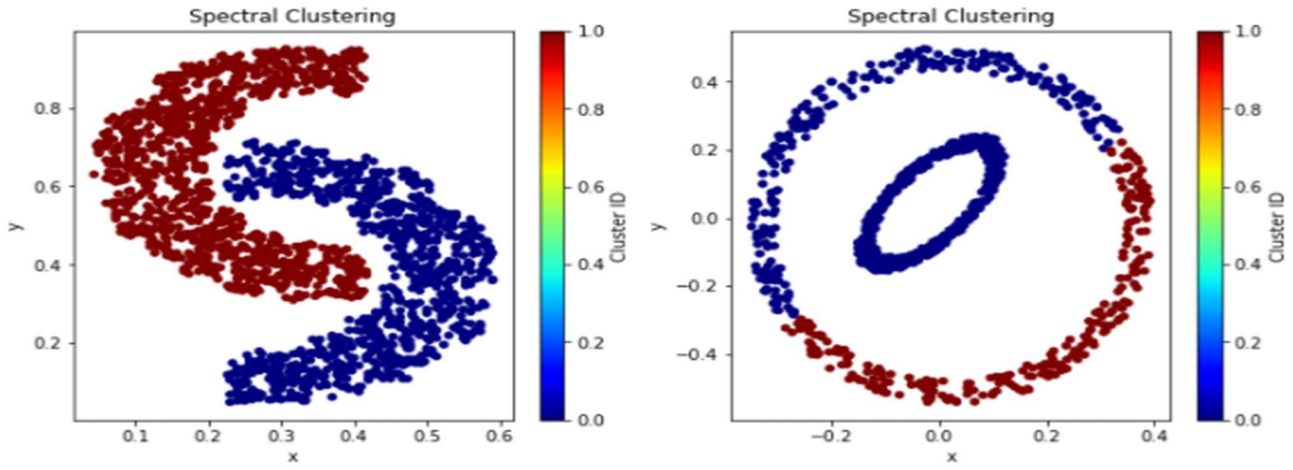
5. For elliptical and 2D data, we apply k-means and spectral clustering with the number of clusters(k) = 2. Repeat the same set of clustering for k = 4 and copy and paste the clusters formed. Which cluster method performs well when k =2 and k = 4.

K = 2, K-means Clustering

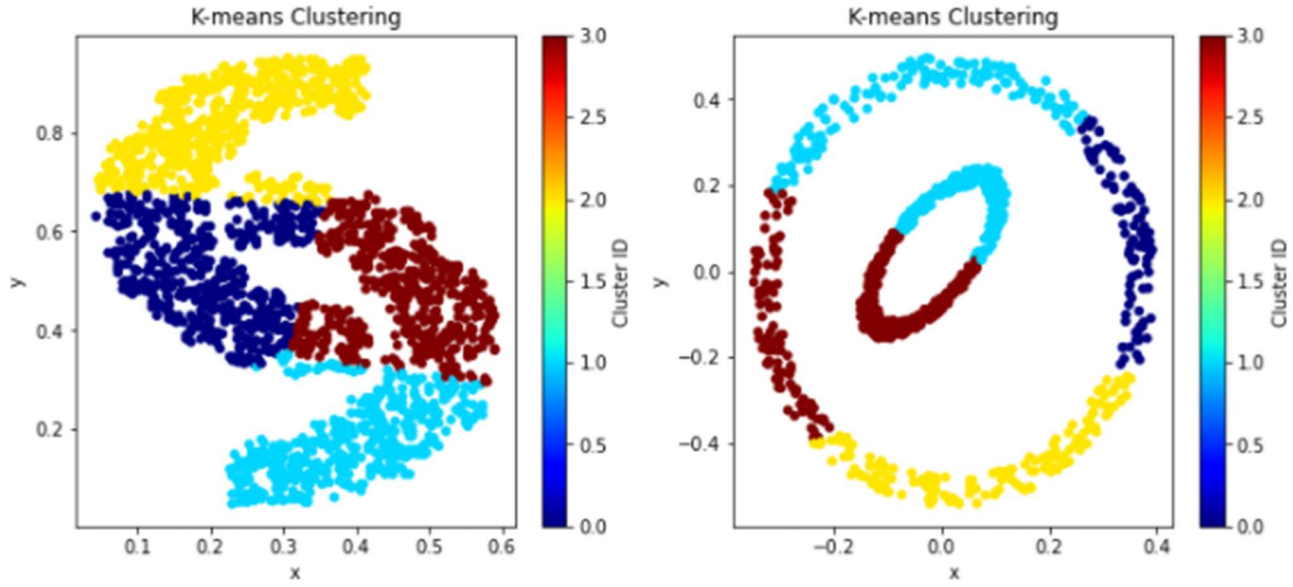


K=2, We can't use the K-means clustering which is not the convex set and when the K-means' shape is globular, it seems well distinguished, but the Spectral clustering shows more distinguishable result in comparison by the center ring shows 2 colors in K-means, but Spectral shows only 1 colored cluster ID.

Spectral Clustering

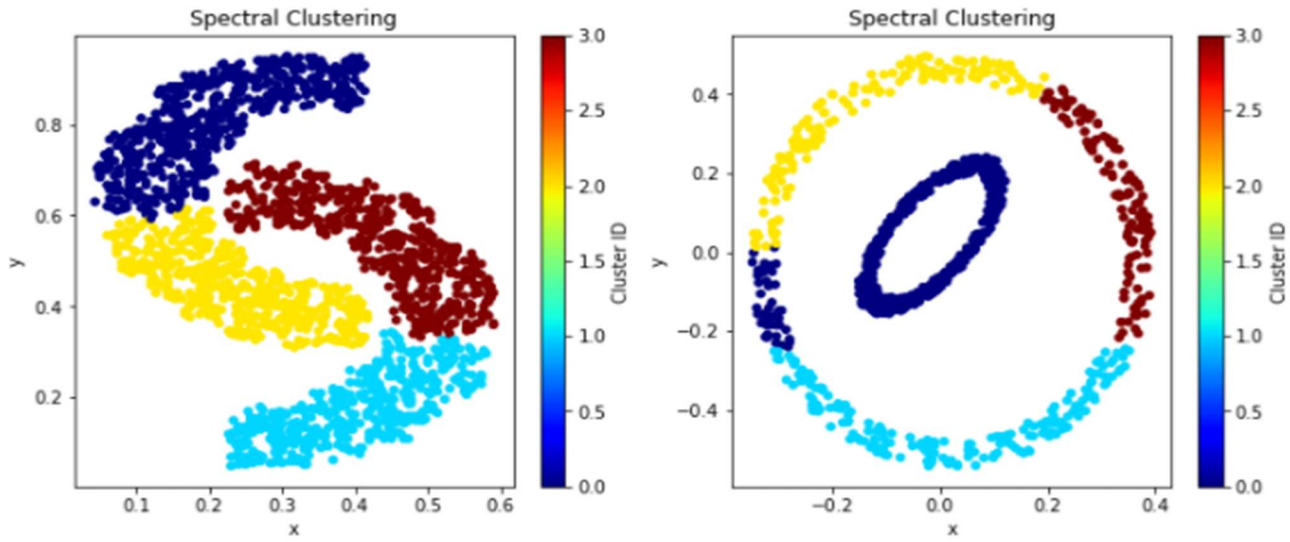


K = 4, K-means Clustering



K=4, We can't use the K-means clustering which is not the convex set and when K-means' shape is globular, it seems well distinguished, but the Spectral clustering shows more distinguished result in comparison by the all colors are all over the place, but Spectral clustering's colors are divided exactly.

Spectral Clustering



Problem 4 [30 points]

The files for this problem is under Experiment 4 folder. Jupyter notebook: covid-19- research-challenge.ipynb. In this experiment, given the large amount of academic literature surrounding COVID-19, you will help overloaded scientists to keep up with the research happening all around the globe for faster development of the vaccine. Given that we have recently studied clustering, can we cluster similar research articles together to make it easier for health professionals to find relevant research articles? Clustering can be used to create a tool to identify related articles, given a target article. Dataset Description: In response to the COVID-19 pandemic, the White House and a coalition of leading research groups have prepared the COVID-19 Open Research Dataset (CORD-19). CORD-19 is a resource of over 29,000 scholarly articles, including over 13,000 with full text, about COVID-19, SARS-CoV-2, and related coronaviruses. This freely available dataset is provided to the global research community to apply recent advances in natural language processing and other AI techniques to generate new insights in support of the ongoing fight against this infectious disease. There is a growing urgency for these approaches because of the rapid acceleration in modern coronavirus literature, making it difficult for the medical research community to keep up.

1. After we handle duplicates, what is the count, mean, standard deviation minimum, and maximum values for abstract word count and body word count?

]:

	abstract_word_count	body_word_count
count	24584.000000	24584.000000
mean	216.446673	4435.475106
std	137.065117	3657.421423
min	1.000000	23.000000
25%	147.000000	2711.000000
50%	200.000000	3809.500000
75%	255.000000	5431.000000
max	3694.000000	232431.000000

2. Briefly describe the data pre-processing steps done in the notebook for cleaning up the text.

In order to improve any clustering or classification efforts, we dropped Null values in the datasets, limited number of articles to speed up computation, removed punctuation from each text, and converted each text to lower case. Eliminate useless or unnecessary information to overcome runtime and achieve accurate results.

3. For clustering, to create a feature vector, on what part of the article did we focus?

For clustering, to create a feature vector, we focused on the only body_text of the articles.

4. What is N-gram in machine learning? Given the following word list: ['the', '2019', 'novel', 'coronavirus', 'sarscov2', 'identified', 'as', 'the', 'cause', 'of'], what is its 2-gram?

➔ N-gram is a sequence of N words and 2-gram is a two-word sequence of words.

['the', '2019', 'novel', 'coronavirus', 'sarscov2', 'identified', 'as', 'the', 'cause', 'of'] in 2-gram is below

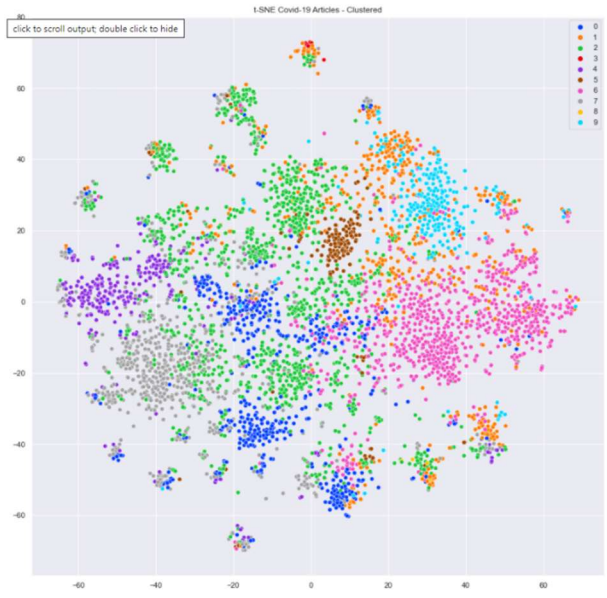
= "the2019", "2019novel", "novelcoronavirus", "'coronavirus'sarscov2", "'sarscov2identified", "'identifiedas", "asthe", :thecause", and "causeof".

5. What does HashingVectorizer do? What is the feature size of HashingVector that we used in our analysis?

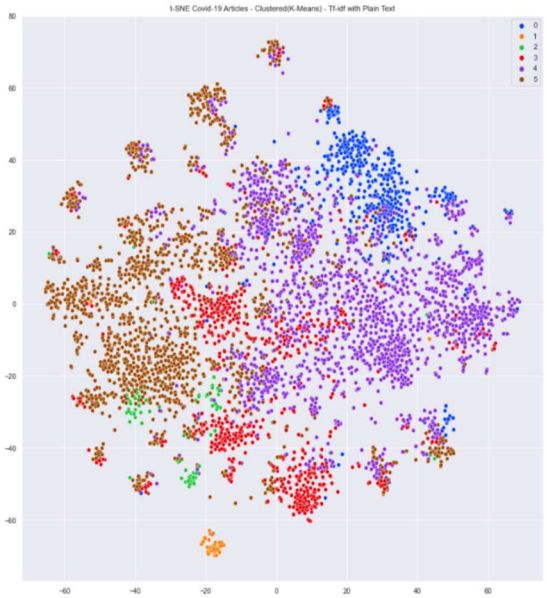
➔ We used HashingVectorizer to limit the features size to 2**12 (4096) to speed up the computation and reduce memory reduction.

6. We have randomly chosen 10 clusters using k-means clustering, vectorized using HashingVector, which makes some sense if we plot the t-SNE plot as articles from the same cluster are near each other, forming groups. However, there are still overlaps. Can you improve this by changing the cluster size or choosing a different feature size? Give the size of the cluster and the feature size that makes more sense for you. Copy and paste the corresponding t-SNE plot.

➔ max features as 2**12 and k = 10



max features as 2**10 and k = 6

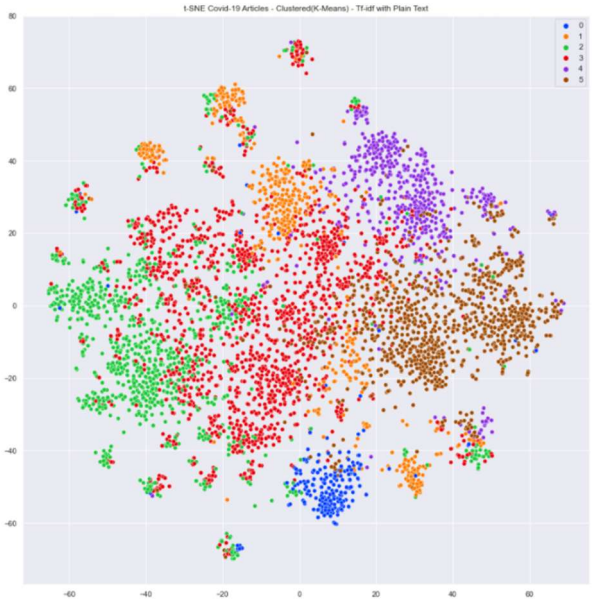


➔ Original plot shows overlapped clusters' color or ID, not easy to classify the clusters well. Even though they look like similar, but some cluster ID does not have overlapped cluster on Improved plot which looks way better to classify the clusters and stay such a distance from other clusters.

7. We have randomly chosen 10 clusters using k-means clustering, vectorized using tf-idf, and we can see clusters more clearly. Can you improve this by changing the cluster size or changing the max features value of TfIdfVectorizer? Give the size of the cluster and the max features value that makes more sense for you. Copy and paste the corresponding t-SNE plot.

➔ **Original, max feature as 2**12, k =10**

Improved, max feature as 28 and k = 6**



➔ Original plot shows overlapped clusters' color or ID, not easy to classify the clusters well. Even though the original plot shows clear clusters, but Improved plot looks way better to classify the clusters and stay such a distance from other clusters.

8. In the interactive t-SNE with 20 clusters, can you do a manual analysis of each cluster to see what articles cluster 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 moreover choose to display points of one cluster only in the plot. Also, name the clusters that include articles involving the social and economic impacts of the coronavirus?

Source - https://maksimekin.github.io/COVID19-Literature-Clustering/plots/t-sne_covid-19_interactive.html

➔ It is not easy to do a manual analysis of each cluster to see what articles cluster together. Instead of having to manually search for related work, every publication is connected to a larger topic cluster.

C-1 MERS

1. Enhanced protection in mice induced by immunization with inactivated whole viruses compare to spike protein of middle east respiratory syndrome coronavirus

2. Host-directed therapies for improving poor treatment outcomes associated with the middle east respiratory syndrome coronavirus infections

3. Taming the beast: Hospital management of a nosocomial middle east respiratory syndrome outbreak

4. Development of Dual TaqMan Based One-Step rRT-PCR Assay Panel for Rapid and Accurate Diagnostic Test of MERS-CoV: A Novel Human Coronavirus, Ahead of Hajj Pilgrimage

5. Effect of isolation practice on the transmission of middle east respiratory syndrome coronavirus among hemodialysis patients: A 2-year prospective cohort study

C-2 CHILDREN INFECTIONS

1. Impact of viral infections in children with community-acquired pneumonia: results of a study of 17 respiratory viruses

2. Pathogens Causing Respiratory Tract Infections in Children Less Than 5 Years of Age in Senegal

3. Molecular monitoring of causative viruses in child acute respiratory infection in endemo-epidemic situations in Shanghai

4. Human metapneumovirus in patients hospitalized with acute respiratory infections: A meta-analysis

5. Elucidation and Clinical Role of Emerging Viral Respiratory Tract Infections in Children

C-3 COVID19

1. BARICITINIB - A JANUASE KINASE INHIBITOR - NOT AN IDEAL OPTION FOR MANAGEMENT OF COVID 19

2. What Should Gastroenterologists and Patients Know About COVID-19?

3. Clinical trials on drug repositioning for COVID-19 treatment

4. Application of refined management in the prevention and control of coronavirus disease 2019 epidemic in non-isolated areas of a general hospital

5. COVID-19: towards controlling of a pandemic

C-12 PREDICTION of COIVD19

1. Production of IFN-β during Listeria monocytogenes Infection Is Restricted to Monocyte/Macrophage Lineage

2. Coupled effects of local movement and global interaction on contagion

3. Statistical learning techniques applied to epidemiology: a simulated case-control comparison study with logistic regression

4. Prediction of COVID-19 Outbreak in China and Optimal Return Date for University Students Based on Propagation Dynamics

5. Anticipating epidemic transitions with imperfect data

C-17 PCR

1. Evaluation of a specialized filter-paper matrix for transportation of extended bovine semen to screen for bovine herpesvirus-1 by real-time PCR

2. Practical experience of high throughput real time PCR in the routine diagnostic virology setting

3. Advances in Real-Time PCR: Application to Clinical Laboratory Diagnostics

4. Use and Evaluation of Molecular Diagnostics for Pneumonia Etiology Studies

5. Identification of Upper Respiratory Tract Pathogens Using Electrochemical Detection on an Oligonucleotide Microarray