Covid Lab Report

Jiayue Meng, Xubin Lou

Abstract— **This project analyzed tweets collected every day for several months during the COVID-19 by identifying topics of each tweet, aggregating the data set and calculating the cosine similarity values by state and region. Parallel coordinates plots were used to illustrate average topic similarity values for each state. Also, through running the clustering metric Calinski-Harabasz score, we found the ideal number of clusters for k-means clustering and spectral clustering. Colored scatterplots were applied to show the results of the two clustering methods.**

I. INTRODUCTION

In this lab, we want to study the association between the location and word types distribution in texts.

We expected to derive a clean data set with no duplicate rows, only one state in the location column, and extra attributes called day and region. We also expected to clean the text of each tweet to become plain text with no hashtags, punctuation, stop words, links, emojis, words that are 3 letters longer or shorter, and the first two characters from the tweet if they are "'b". What's more, we expected to identify the topics people are talking about in tweets and calculate average normalized cosine similarity scores for each topic in each dictionary according to state and region. Then we wanted to find the best number of clusters between 2 to 20 for k-mean clustering and spectral clustering. Finally we tried to visualized the results of average topic similarity values and clustering through parallel coordinates plots and scatter plots.

We found that the location information such as states and region did not affect the distribution of words in text significantly. The distributions of words in texts are so similar regardless of location. That means the location is not a significant factor affecting the word types about Covid-19 in the text. By using Calinski-Harabasz score, parallel coordinates plot, and PCA, we found k-means clustering created better results than the spectral clustering.

II. DATA

The data set we worked with is a large collection of tweets originating from the United States. The data set contains tweets collected for several months during the COVID-19. There are 2000000 observations in the data set with 10 attributes including users' unique identifier, unique identifier of tweets, date and time of tweets, screen names, content of tweets, hashtags, locations, number of followers, friends, and tweets of users. The 10 attributes are userid, statusid, createdat, screenname, text, hashtags, location, followerscount, friendscount, and statusescount.

Some additional dictionaries were also used in the project: disinfectants, vaccine, medicine, and isolation. These dictionaries are related to the COVID-19 on different aspects.

III. RESULTS

We cleaned the original data set by dropping duplicate rows with identical statusid and tweets that have no state name or have more than one state names in the location column. After dropping, the number of observation in the updated data set is 499716, and the percentage of the data lost in the cleaning process is 0.000568. The elapsed time for part 1 is 149.7632279396057 seconds.

Then we extracted date from createdat column by transforming the column to datetime and gave it a new column called day. We also extracted region from the location column by creating a dictionary and mapping the state column to our dictionary. We gave it a new column called region. We also found the regions associated with different states in the usstates.xlsx file. The elapsed time for the second section is 1.5876948833465576 seconds.

Importing the package re and using regular expression, we defined a function called tweetcleaner to remove the first two characters from the tweet (which is 'b), all hashtags, punctuation, stop words, all words that are 3 letters long or shorter, all links, and all emojis. Stop words are the English words which does not add much meaning to a sentence. They can safely be ignored without sacrificing the meaning of the sentence. The elapsed time for the third section is 1.477220058441162 seconds.

Lemmatizer function links words with similar meanings to one word. It groups together the different inflected forms of a word so they can be analyzed as a single item. Lemmatizer function converted original text to lemmatized text. The elapsed time for this section is 505.8753080368042 seconds.

In Figure 1 and 2, I observe that the texts during covid-19 period generally include more medicine words and vaccine words, no matter in which region or state. Because the average of the cosine similarity values in the column medicinecosinenormal and vaccinecosinenormal are generally higher than values in the other two columns, which indicates that there are more text containing words in medicine dictionary and vaccine dictionary. The elapsed time for the sixth section is 8.391955852508545 seconds.

According to the Figure 3, by observing the columns states, kmeans.labels, region, and clustering.labels, we found there's no obvious relation between the clustering and locations. Compared the k-means and spectral clustering, k-mean gives relatively better results since it has a small pattern among states by observation, while spectral clustering seems totally random without any pattern.

Figure 4 and Figure 5 are Calinski-Harabasz score for K-means clustering and Calinski-Harabasz score for spectral

clustering. K-means clustering gives better results since it generally has higher Calinski-Harabasz scores. The result is the same as what we identified in Q7. And we found when cluster number n=2, Calinski-Harabasz score is the highest.

According to Figure 6 Parallel coordinates plots for state, we can see that there are similar patterns but no identical patterns. The disinfectant cosine number generally around 0.002 to 0.003; isolation cosine normal is around 0.0045 to 0.006; medicine cosine normal is around 0.012 to 0.016; vaccine cosine normal is around 0.0115 to 0.016. Since they are not in the identical pattern, it means stats still make some difference in the distribution of word types.

In Figure 7 Parallel coordinates plots for region, there are nearly identical patterns. The disinfectant cosine number generally around 0.0022; isolation cosine normal is around 0.005; medicine cosine normal is around 0.014; vaccine cosine normal is around 0.0135 to 0.014. Since there is nearly identical patterns among different states, it means states do not affect the words type distribution. Thus, we get the k-means clustering

According to the scatterplots Figure 8 and 9, there's a left to right pattern in the k-means clustering scatter plot, while there's no obvious pattern in spectral clustering scatter plot. This also checks the outcome I got in Q7, since it also shows that the k-means clustering create better result than the spectral clustering.

Variables disinfectant cosine normal,isolation cosine normal, medicine cosine normal, and vaccine cosine normal seem the variables we got to explain the outcomes the most because they showed the associations between dictionaries and locations about COVID-19-related words in tweets in a clear any numerical way, which is one of the main goals of this lab.

We could improve this analysis by trying more clustering methods and more possible parameters. So that we will have better insight about which clustering method perform better and why. We can also improve this analysis by including more words in dictionaries and providing more dictionaries containing different types of words. Some words may not covered in these four dictionaries, so some tweets that are related to COVID-19 might be omitted.

There's causal inference problem selection bias. Since all texts are collected from tweeter, it can't include the information of certain kinds of people who do not use tweeter for socialization and communication. There might be a range of aged people who will not use tweeter. Also, in different regions, people's habits are also not the same. If the people in a certain region generally don't get used to using tweeter, the information collected from the tweets will not represent the general persons' real condition to covid-19.

Below are the figures:

Figure1: state topic score data

Figure2: region topic score data

Figure 3: K-means and Spectral Clustering labels

Figure 4: Calinski-Harabasz score for K-means Clustering

Figure 5: Calinski-Harabasz score for Spectral clustering

Figure 6: Parallel coordinates plots for state

Figure 7: Parallel coordinates plots for region

Figure 8: Scatterplot for K-means clustering

Figure 9: Scatterplot for spectral clustering

	state	disinfectant_cosine_normal	isolation_cosine_normal	medicine_cosine_normal	vaccine_cosine_normal
0	alabama	0.002194	0.005417	0.014093	0.014238
1	alaska	0.002073	0.004883	0.013072	0.013477
2	arizona	0.002190	0.004967	0.013306	0.013177
3	arkansas	0.002100	0.006103	0.013392	0.012304
4	california	0.002336	0.004720	0.013124	0.012918
5	colorado	0.002200	0.004900	0.014075	0.014349
6	connecticut	0.002125	0.005051	0.014756	0.016227
7	delaware	0.002309	0.004461	0.012756	0.013295
8	florida	0.002118	0.004837	0.013087	0.012396
9	georgia	0.002312	0.004588	0.014064	0.013792
10	hawaii	0.002152	0.006105	0.012230	0.011406
11	idaho	0.002202	0.005641	0.015817	0.014398
12	illinois	0.002201	0.005052	0.013358	0.013760
13	indiana	0.002150	0.004717	0.014404	0.014268
14	iowa	0.002346	0.005292	0.013693	0.011933
15	kansas	0.002375	0.005167	0.013418	0.013267
16	kentucky	0.002249	0.005415	0.013974	0.013293
17	louisiana	0.002180	0.005595	0.012731	0.013078
18	maine	0.002234	0.004822	0.012353	0.012539
19	maryland	0.002371	0.005572	0.015310	0.014835
20	massachusetts	0.002267	0.004832	0.014687	0.013615

Fig. 1. State

	region	disinfectant_cosine_normal	isolation_cosine_normal	medicine_cosine_normal	vaccine_cosine_normal
0	Midwest	0.002229	0.005085	0.013848	0.013373
1	Northeast	0.002222	0.005043	0.013656	0.013929
2	South	0.002196	0.005025	0.013732	0.013542
3	West	0.002183	0.004989	0.013732	0.014189

Fig. 2. region

state	disinfectant_cosine_normal	isolation_cosine_normal	medicine_cosine_normal	vaccine_cosine_normal	kmeans.labels_	region	clustering.labels_
alabama	0.002194	0.005417	0.014093	0.014238	1	South	3
alaska	0.002073	0.004883	0.013072	0.013477	2	West	0
arizona	0.002190	0.004967	0.013306	0.013177	2	West	0
arkansas	0.002100	0.006103	0.013392	0.012304	0	South	1
california	0.002336	0.004720	0.013124	0.012918	2	West	2
colorado	0.002200	0.004900	0.014075	0.014349	1	West	3
connecticut	0.002125	0.005051	0.014756	0.016227	3	Northeast	3
delaware	0.002309	0.004461	0.012756	0.013295	2	South	0
florida	0.002118	0.004837	0.013087	0.012396	0	South	2
georgia	0.002312	0.004588	0.014064	0.013792	1	South	2
hawaii	0.002152	0.006105	0.012230	0.011406	0	West	1
idaho	0.002202	0.005641	0.015817	0.014398	3	West	3
illinois	0.002201	0.005052	0.013358	0.013760	2	Midwest	0
indiana	0.002150	0.004717	0.014404	0.014268	1	Midwest	2
iowa	0.002346	0.005292	0.013693	0.011933	0	Midwest	1
kansas	0.002375	0.005167	0.013418	0.013267	2	Midwest	0
kentucky	0.002249	0.005415	0.013974	0.013293	2	South	1
louisiana	0.002180	0.005595	0.012731	0.013078	2	South	0
maine	0.002234	0.004822	0.012353	0.012539	0	Northeast	0
maryland	0.002371	0.005572	0.015310	0.014835	3	Northeast	3

Fig. 3. Clustering

```
[42.60453059920297,
 36.4168461751497,
 34.274322354067145,
 32.21569586191919,
 32.42323267641285,
 31.61701687656642,
 30.9896381220678,
 30.254808271988978,
 31.89527586792317,
 30.75229048406643,
 31.276601889442667,
 30.609340633654014,
 31.728297500057273,
 32.78500674187649,
 32.00292627442344,
 32.21126741902102,
 33.33463224412626,
 33.82653307969675,
 32.05220738299071]
```

Fig. 4. kmeans

```
[41.86125455975408,
 24.58896027558781,
 14.793770776736158,
 14.774022775978027,
 12.089086346464788,
 9.168731105167495,
 8.212379828064504,
 5.539946609454138,
 4.051911058619384,
 3.6075790861928723,
 3.771146156466814,
 3.8912442720790645,
 3.117457697236752,
 2.767238419614901,
 2.38283350190769,
 3.791165063603663,
 3.258186898671288,
 3.6284555689288616,
 2.233190268227704]
```

Fig. 5. spectrual

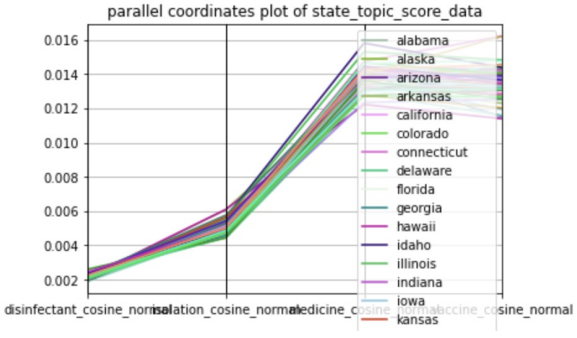

Fig. 6. state

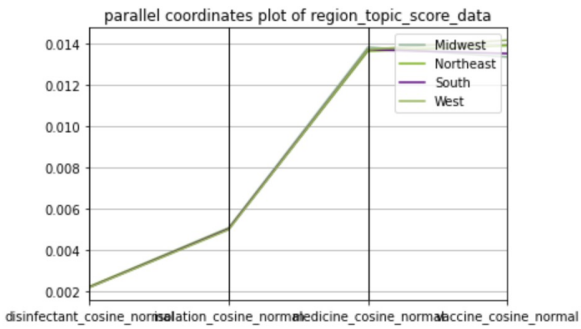

Fig. 7. region

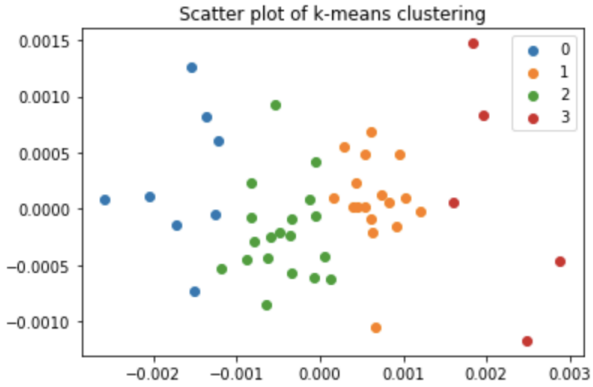

Fig. 8. kmeans

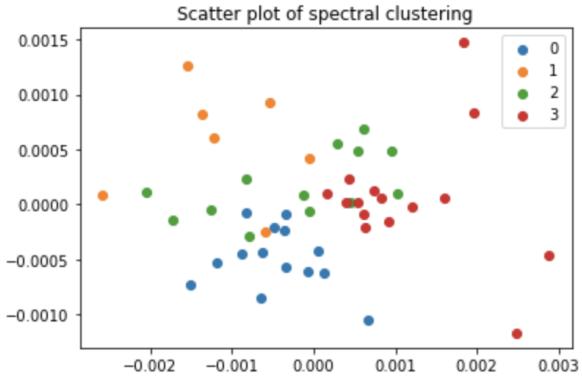

Fig. 9. spectral