Covid Lab

Dang Pham and Kairuo Yan

***Abstract*— In this lab, we built functions to calculate cosine similarity score of each COVID-related tweet to understand what topics are people talking about. We have learnt about natural language processing, cosine similarity score, and Kolmogorov-Smirnov test. However, my supporting visuals have serious drawbacks such that they do not bring great values or show any identifiable patterns.**

I. INTRODUCTION

Coronavirus disease 2019 (COVID-19) is an infectious disease caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). It was first identified in December 2019 in Wuhan, Hubei, China, and has resulted in an ongoing pandemic. As of March 10, 2021, more than 117 million cases have been reported across 188 countries and territories with more than 2.5 million deaths. In this lab, we used natural language processing method to analyze COVID-related tweets using pre-defined dictionaries in various topics. By calculating cosine similarity of each tweet and topic dictionaries, we can see what topic the tweet are mentioning. One of our main findings was that North Dakota is the state the concern the most about this disease in the United State while the state tweet the least about COVID is New York. Another finding in our lab is that discussions related to disinfectant, isolation, medicine, and vaccine increase gradually overitme in every regions of the U.S.

II. DATA

For this lab, we used data collected from a large collection of tweets originating from the United States: tweets_us_states.csv. We also used four topic dictionaries in the form of Excel tables: disinfectants, vaccine, medicine, and isolation. There are a wide range of columns with information relating to the tweets. However, we only used 'text' and 'location' columns for our data manipulation and analysis.

There are 265203 observations and 9 variables from the tweets_us_states.csv file. In this dataset, we can find various information relating to the tweet, such as user_id, status_id, when was it created (created_at), tweet's text, list of hashtags, location, followers_count, friends_count, statuses_count.

We are only interested in three variables: created_at, text, and location. From the 'created_at' variable, we extracted the date each tweet was posted, and for the 'location' variable, we kept observations that only have one U.S.' state. With the 'text' variable, we lemmatized the content of the tweets to generate a cleaner version of this text. (I think we need to explain the lemmatized function here). From this cleaned text, we ran cosine similarity calculation function compares the similarity of the text with each topic dictionary. This returns similarity scores, then we added them into the original dateset as 4 new columns: disinfectant_cosine, isolation_cosine, medicine_cosine, vaccine_cosine.

III. RESULTS

There are three sections of findings we explored in our lab. In the first section, we explored the total computation time, and in the second section, we explored the similarity scores for tweet topics in states and regions over time. In the third section, we compare the distributions of topics scores for different regions in a pairwise manner between regions.

For the total computation time in our lab, we found the part where we were lemmatizing the words in tweets and returning the cleaned and lemmatized tweet take the longest time, which is about 1034 seconds. Cleaning the original dataset takes the second-longest time, which is about 972 seconds. Aggregating the Data in Panel Format takes the least of time, which is about 0.5 seconds.

For the similarity scores for tweet topics in states and regions over time, we first did a parallel coordinate plot using average topic similarity values for each state. We obtained four main findings. The three states with the highest vaccine similarity values are South Dakota, Nebraska, and North Dakota, and the values are about 0.707, 0.174, and 0.147. The three states with the lowest vaccine similarity values are that New York, Florida, and Taxis, and the values are about 0.027, 0.026, and 0.025. The three states with the highest disinfectant similarity values are Alabama, Rhode Island, and Idaho, and the values are about 0.182, 0.121, and 0.105. The three states with the lowest disinfectant similarity values are Georgia, Washington, and Louisiana, and values are about 0.020, 0.019, and 0.012.

The three states with the highest medicine similarity values are North Dakota, Idaho, and Alabama, and the values are 0.096, 0.093, and 0.092. The three states with the lowest medicine similarity values are California, Washington, and New York, and the values are 0.020, 0.018, and 0.015.

The three states with the highest isolation similarity values are Vermont, North Dakota, and Mississippi, and the values are 0.196, 0.147, and 0.131. The three states with the lowest isolation similarity values are Washington, New York, and Alabama, and the values are 0.010, 0.008, and 0.

Then we did a line plot for the discussion of four topics for each state over time. We have four findings. The change in discussion for the vaccine is fluctuated and remained low from April to June, except for South Dakota, which has a drastic increase in April. In mid-May, there's an increase in discussion for North Dakota. However, except for these two peaks, all states' discussions are relatively

low before June 2020. And after June 1st, there is more discussion on the vaccine, and more states have an increase in discussion over time. There's a significant and long-lasting increase in disinfectant discussion for Alabama in June. The second peak discussion of disinfectant happened in July in Rhode Island, and the third peak discussion occurred at the beginning of June in Hawaii. The general disinfectant discussions increase in states over time. The changes in discussion for disinfectant are small and relatively stable. However, some peak discussions happened in some specific states. The changes for medicine discussions increase and are immensely fluctuated for many states over time. The isolation discussion changes are relatively less fluctuated and remained low-level similarity scores, except for a few states, such as Hawaii, Vermont, and Minnesota. Most isolation-related discussions in states happened in May and June.

We also found the average topic similarity for each region. Northeast has the highest discussion on the vaccine, Midwest has the most elevated discussion on disinfectant and isolation, and the South has the most increased discussion on medicine. Besides, we explored the change in discussion overtime in four topics, and we found that the overall trend for vaccine discussion increases overtime. The discussion on disinfectant is at the peak in April and decreases dramatically after. The discussion in medicine immensely fluctuates over time, and it doesn't show a noticeable trend. The discussion on isolation is more concentrated in June, especially in Midwest.

For the last section, we compared the distributions of topic scores for different regions in a pairwise manner between regions. We found the p-value for all regions is 1. Therefore, there's not a difference in the distributions of topic scores between regions.

The cosine normal value explains the outcome the most because it measures the similarity between the topic list and the comment list using cosine values. We normalized it so all the values could be remained in the range from 0 to 1. Using this value, we could compare the discussion on four topics over time for each state and region.

We could collect data in a more oversized time frame. For example, if we could have data from 2020- 2021, we can see how the amount of discussion changes for covid19, especially on the four topics.

We could collect information for each state and discuss the reason for variances in the discussions on covid-19. We were also trying to explore the relationship between variance and other factors—for example, the number of cases, the death rate, or the public attitudes.

There could be selection bias since not everyone tweets online, so the sample size might not be randomized.

IV. VISUALIZATION

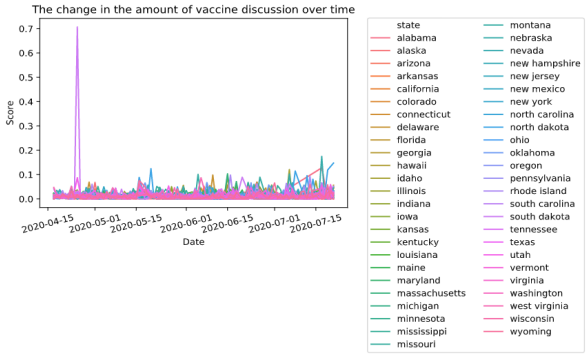

Fig. 1. Vaccine

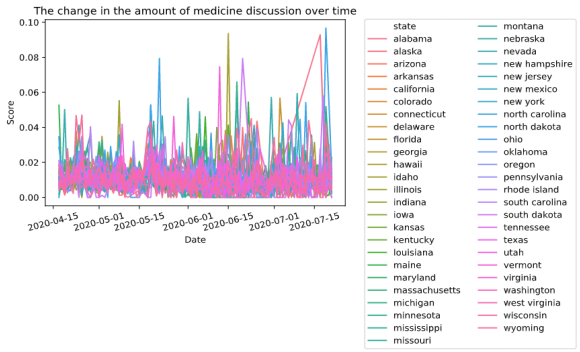

Fig. 2. Medicine

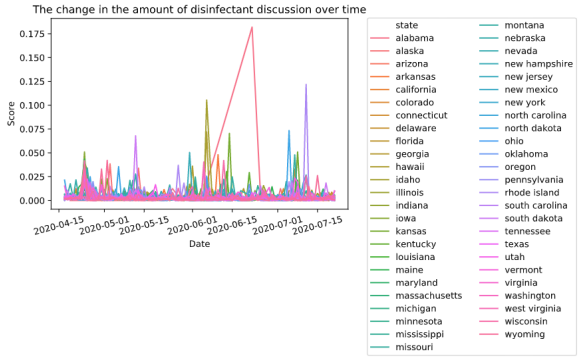

Fig. 3. Disinfection

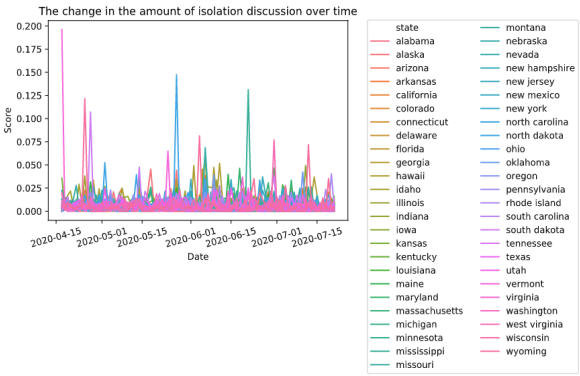

Fig. 4. Isolation

Fig. 5. The change in the amount of discussion in different topic over time

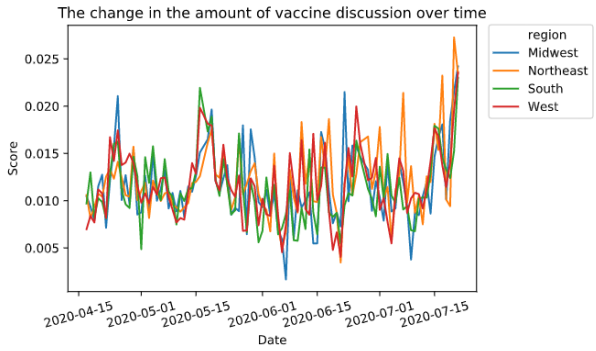

Fig. 6. Vaccine

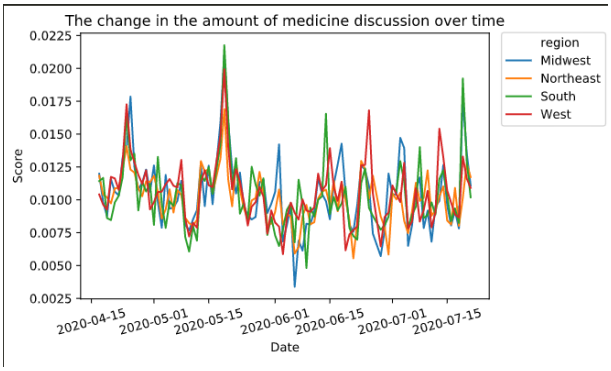

Fig. 8. Medicine

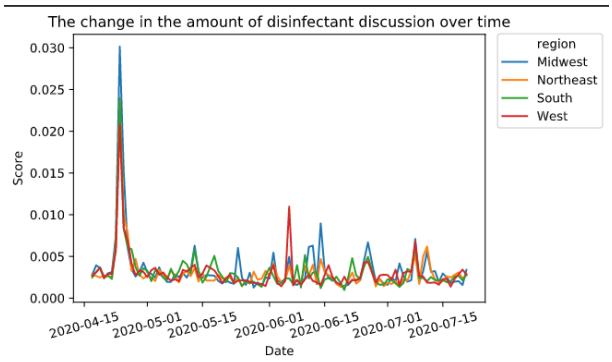

Fig. 7. Disinfectant

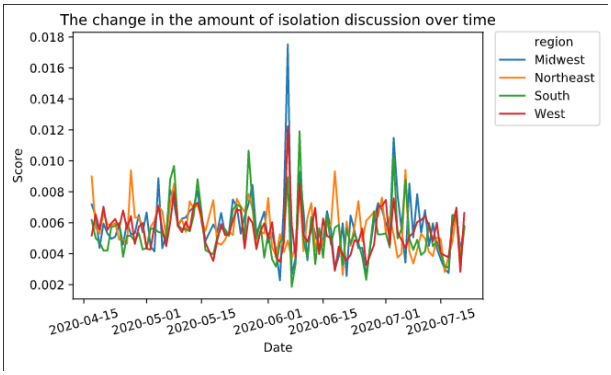

Fig. 9. Isolation

	Midwest	Northeast	South	West
Midwest	1	1	1	1
Northeast	1	1	1	1
South	1	1	1	1
West	1	1	1	1

Fig. 10. Pairwise topic score in different regions

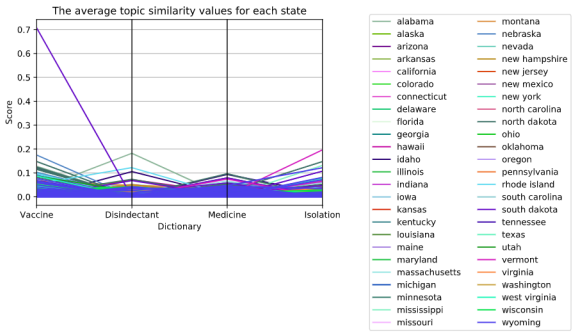

Fig. 11. Average topic similarity values for each state

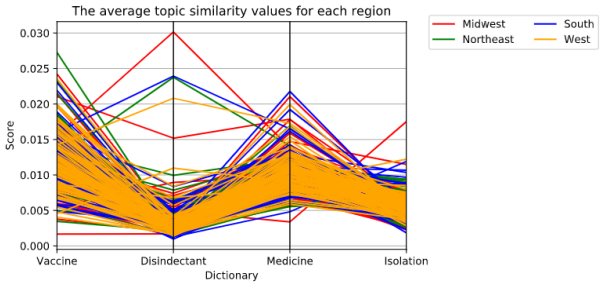

Fig. 12. Average topic similarity values for each region