Sarah Batta, Arushi Srivastava, Stephanie Cao, Madelyn Dempsey

Empirical Analysis

Empirical Analysis
o Summary Report that includes:
§ Hypotheses that you proved/disproved
§ Analysis results
§ Graphs supporting your analysis
o Any code that you wrote or used for the analysis
o The data set that you used (Please include the entire data set and a link to where you got the dataset from.)

Dataset: https://www.kaggle.com/gpreda/all-covid19-vaccines-tweets

This dataset that we used from kaggle called “All COVID-19 Vaccines Tweets” by Gabriel Preda is a set of recent tweets about the COVID-19 vaccines used worldwide. For each tweet in the csv file, there is data for the tweet’s id, the user’s username, location, the actual tweet content, the date the user’s profile was created, their number of followers, number of favorites, the number of retweets, whether or not they are verified, and the date the tweet was created (the file will be included in the zip for this project). Our empirical analysis aims to analyze the relationship between twitter hashtags in our dataset using our getRecommendations function. I was interested in the potential relationship between vaccine effectiveness and degree in the graph of hashtags. My thought process was that if a vaccine was more effective then more people would be talking about it so there would likely be more hashtags used alongside the more effective ones since they would be more widely known about and used.

Hypothesis: Hashtags related to more effective vaccine brands will have a higher degree in the hashtag graph than less effective vaccines.

Results:

pfizer:
Degree : 216
moderna:
Degree: 247
johnsonandjohnson:
Degree: 23
novavax:
Degree: 2
vaxzevria:
Degree: 0

sinovac:

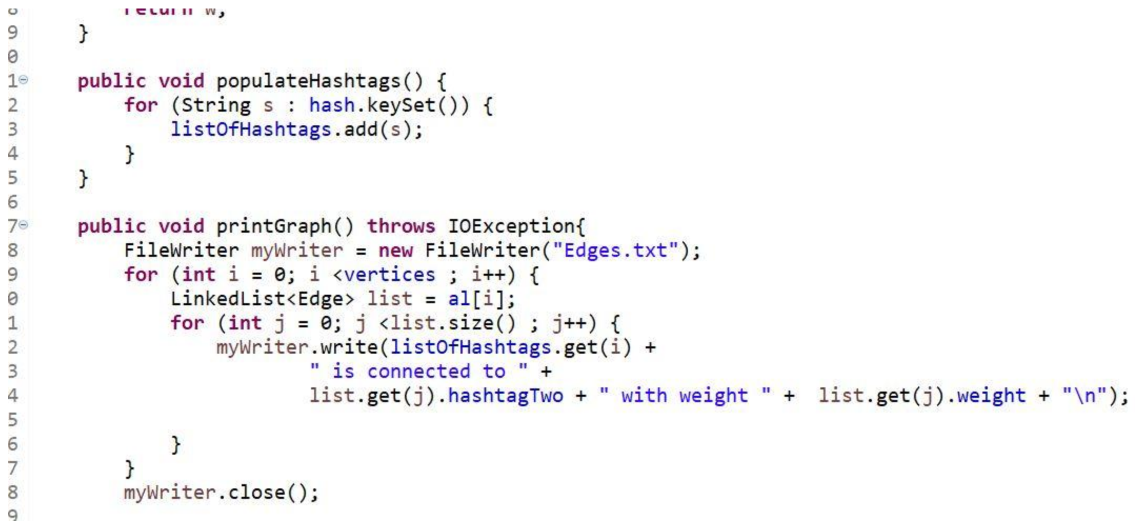
Degree: 62

Analysis:

Since we had issues with actually getting results for my first empirical analysis attempt, I decided to utilize the information we collected about the related hashtags in the COVID-19 dataset. We created methods to find all the different hashtags in the file as well as a recommendation system that has the user choose a hashtag to look at, and then the program returns the top 3 hashtags connected to that hashtag with the greatest edge weights. Using this application, I thought it would be interesting to keep the same theme from my previous experiment where I compared different vaccine brands.

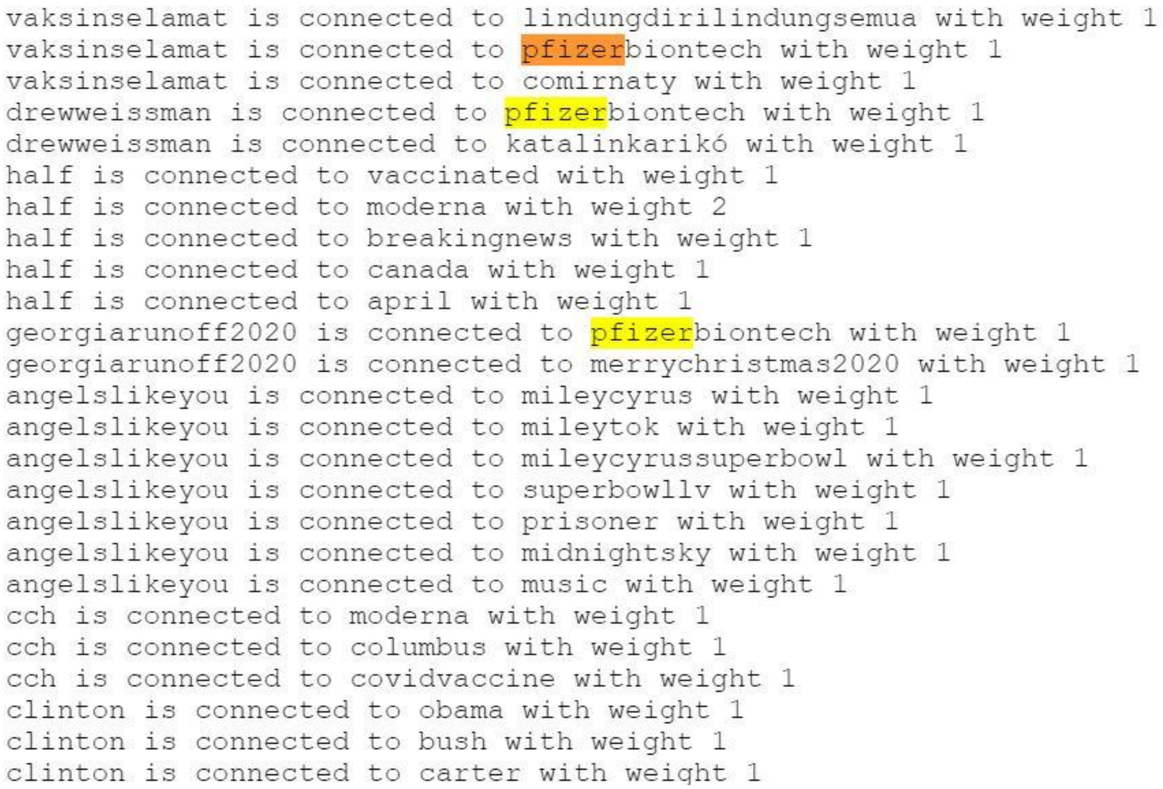
To analyze the different vaccines in this context, I thought it would be beneficial to look at the graph that we create in the WeightedGraph class. In this class, we are able to create a graph in which the vertices represent the hashtags we found in the tweets from the dataset, and there is an edge between two hashtags if they are used in the same tweet. This is also a weighted graph where we add 1 to the weight for every tweet that contains both of the hashtag endpoints. To make the final graph visualization, we make use of adjacency lists where each node has a linked list of edges. The edge object keeps track of other hashtags it is connected to and their weight. We also later added an adjacency matrix of hashtags where there is a 0 if there is no edge between the hashtags and a non-zero number if there was an edge between them.

In order to address my hypothesis, I needed to see the graph visualization in order to see how connected hashtags containing the names of different vaccines were to the rest of the graph. My idea was that hashtags with a greater degree would be more interconnected in the graph and therefore have more potential hashtags to recommend to twitter uses and thus would become even more popular and widely used on Twitter. I thought that there might be a relationship between vaccine effectiveness and hashtag degree because more effective vaccines would probably be more widely known and talked about, so they would probably be used in conjunction with more varied hashtags. Thus, I needed to analyze hashtags related to different vaccine brands and see how many and to which other hashtags they were connected to. To do this, I collaborated with Steph to make a method called printGraph() in our WeightedGraph class using the implementation she created for the weighted graph recommendation system. This method printed out every edge and the vertices connecting them along with their weight. Then we wrote it to a txt file for easy access (this file is provided in the zip of our code). The method we wrote is provided below:



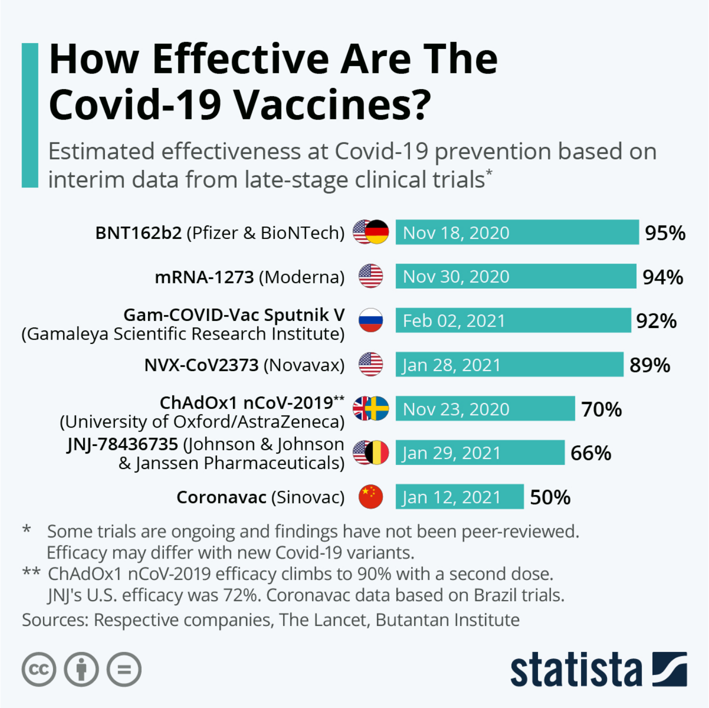
```
        return w;
    }

    public void populateHashtags() {
        for (String s : hash.keySet()) {
            listOfHashtags.add(s);
        }
    }

    public void printGraph() throws IOException{
        FileWriter myWriter = new FileWriter("Edges.txt");
        for (int i = 0; i <vertices ; i++) {
            LinkedList<Edge> list = al[i];
            for (int j = 0; j <list.size() ; j++) {
                myWriter.write(listOfHashtags.get(i) +
                        " is connected to " +
                        list.get(j).hashtagTwo + " with weight " +  list.get(j).weight + "\n");

            }
        }
        myWriter.close();
```

Using this code I wrote, I was able to see exactly how every hashtag was related to the others and by what weight for example:



```
vaksinselamat is connected to lindungdirilindungsemua with weight 1
vaksinselamat is connected to pfizerbiontech with weight 1
vaksinselamat is connected to comirnaty with weight 1
drewweissman is connected to pfizerbiontech with weight 1
drewweissman is connected to katalinkarikó with weight 1
half is connected to vaccinated with weight 1
half is connected to moderna with weight 2
half is connected to breakingnews with weight 1
half is connected to canada with weight 1
half is connected to april with weight 1
georgiarunoff2020 is connected to pfizerbiontech with weight 1
georgiarunoff2020 is connected to merrychristmas2020 with weight 1
angelslikeyou is connected to mileycyrus with weight 1
angelslikeyou is connected to mileytok with weight 1
angelslikeyou is connected to mileycyrussuperbowl with weight 1
angelslikeyou is connected to superbowllv with weight 1
angelslikeyou is connected to prisoner with weight 1
angelslikeyou is connected to midnightsky with weight 1
angelslikeyou is connected to music with weight 1
cch is connected to moderna with weight 1
cch is connected to columbus with weight 1
cch is connected to covidvaccine with weight 1
clinton is connected to obama with weight 1
clinton is connected to bush with weight 1
clinton is connected to carter with weight 1
```

Once I got these results, I found that there were 216 edges to the pfizer hashtag, 247 edges to the moderna hashtag, 23 to johnsonandjohnson, 2 to novavax, 0 to vaxzevria, and 62 to sinovac. Based on my hypothesis, I would have expected that pfizer would have the greatest degree then moderna, novavax, johnsonandjohnson, vaxzevria (60%), then sinovac. However, the results indicated that moderna had the greatest number of edges followed closely by pfizer, and then sinovac was the third highest even though it was the least effective.


How Effective Are The Covid-19 Vaccines?

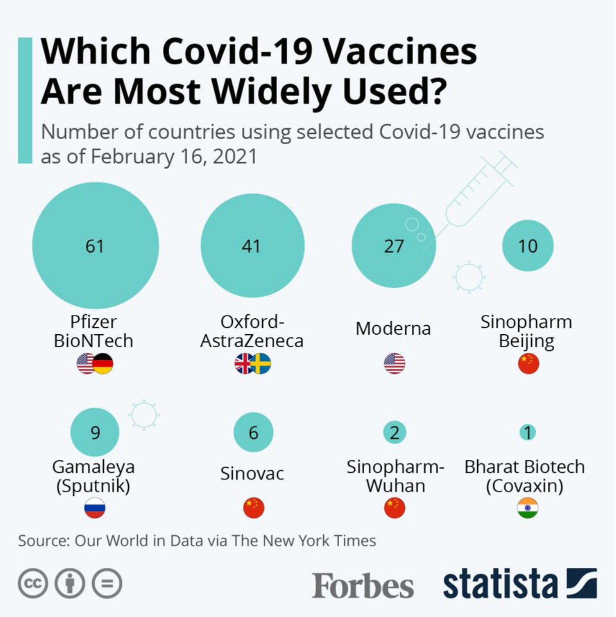
Estimated effectiveness at Covid-19 prevention based on interim data from late-stage clinical trials*

* Some trials are ongoing and findings have not been peer-reviewed. Efficacy may differ with new Covid-19 variants.

** ChAdOx1 nCoV-2019 efficacy climbs to 90% with a second dose. JNJ's U.S. efficacy was 72%. Coronavac data based on Brazil trials.

Sources: Respective companies, The Lancet, Butantan Institute

This indicated that there could be stronger factors at play that are increasing the use of the less effective vaccines in conjunction with other hashtags. For instance, the availability of vaccine brands in different countries could be skewing these results because our dataset considers tweets from all over the world and not just in the United States. Thus, less effective vaccines could have been more connected in the graph because tweets from people in other countries would likely tweet more about the vaccines they have access to. For instance, in our dataset the only vaccine that shared an edge with the hashtag “china” was sinovac. Sinovac is one of the most commonly used vaccines in China, so this could have made the sinovac hashtag more used on twitter than other, more effective brands if more people in china or talking about china were tweeting about vaccinations. In terms of the discrepancy between Moderna and Pfizer, we see in the graph below that Moderna is only widely used in America, and we know that Americans make up the largest group on twitter with an audience of 69.3 million.


Which Covid-19 Vaccines Are Most Widely Used?

Number of countries using selected Covid-19 vaccines as of February 16, 2021

Source: Our World in Data via The New York Times (https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/) The fact that so many Americans are present on twitter could be the factor impacting this deviance from the hypothesis because Americans would likely be the ones talking about the Moderna vaccine on Twitter because it is the only country using it, so there may be more tweets about it in comparison to the other vaccines.

Graph Links:

https://www.forbes.com/sites/niallmccarthy/2021/02/16/which-covid-19-vaccines-are-most-widely-used-infographic/?sh=372fb65d77d9

https://www.statista.com/chart/23510/estimated-effectiveness-of-covid-19-vaccine-candidates/