Kylie Wise

Jordan Rodu

STAT 4559

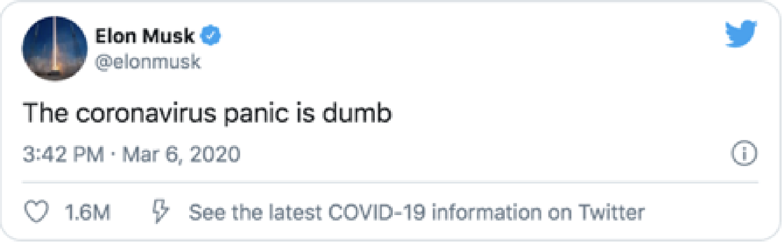
9 December 2020

Coronavirus Tweet Text Analysis

Introduction & Research Question

The Coronavirus has had several impacts on the way most people live their lives. From the small things, such as wearing a protective mask to the grocery store, to the larger things, such as losing a family member to the virus, many people have been struggling to adjust to a new way of living. As is expected, people feel strongly about the virus and its impact. Many are turning to twitter and other social media platforms in order to share thoughts and join a global discussion about the Coronavirus.

I was actually one of the over 13 million people infected with the Coronavirus this year. Luckily, I only experienced mild symptoms and have fully recovered. However, during my time with Coronavirus, without a twitter account, I was forced to lean on close friends to discuss, complain, and share my anxieties about the virus. Not having a twitter means that I have missed out on months of Coronavirus tweets. Because of this I was particularly drawn to the COVID-19 twitter datasets I came across and wanted to make it the focus of my project.



Elon Musk
@elonmusk

The coronavirus panic is dumb

3:42 PM · Mar 6, 2020

1.6M   See the latest COVID-19 information on Twitter

[1]

The research question that I am particularly interested in answering is how did people's attitudes towards the Coronavirus change as the virus evolved? In other words, how do people's attitudes from March 2020 compare to people's attitudes in December 2020. Or more specifically, I wanted to look at how attitudes and topics changed during monumental Coronavirus announcements. For example, how did people feel before there was an announcement of a vaccine versus after the announcement of a vaccine by pharmaceutical companies such as Pfizer and Moderna? Would we be able to see changes in attitudes, sentiment, or topics as the Coronavirus evolved?

[1] One of the over 628 Million Coronavirus tweets that live on twitter. Elon Musk shared his opinions on March 2020 about the Coronavirus, exemplifying how many people are expressing their feelings about the virus. https://www.tweetbinder.com/blog/covid-19-coronavirus-twitter/

My initial hypothesis is that there would be some sort of change in the topics or attitudes of people as the Coronavirus has evolved. Many of us have heard the adage, “If you can’t change it, change your attitude”. But do people really live by this?

The answer to these research questions are particularly interesting for two reasons. The first being that through exploring these research questions it will paint a detailed picture of people’s shifting attitudes towards the virus. At least for me, my attitudes and thoughts about the Coronavirus have shifted dramatically since earlier this year, so I would predict this to be true of many other people. The second reason I find this research question to be so interesting is that it perfectly exemplifies how people think and change when they are experiencing a crisis. By quantifying such a large group of people’s thoughts we may get a better look into how people, in general, deal with crisis, change, and uncertainty. Can we find patterns in how people respond to crises? If so, how can we learn from that to better understand people in preparation for the next global crisis?

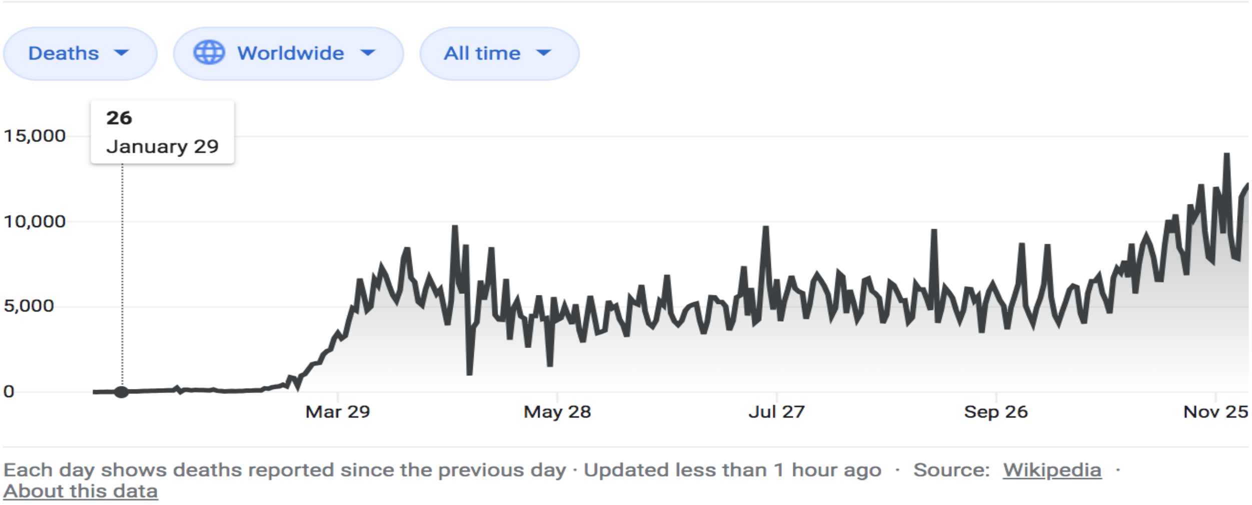
Research Methods

There are many ways of attempting to answer these research questions so I want to outline a few of them below. However, before I begin an explanation of my research method, I must outline how I thought about organizing, bucketing, and sectioning the tweets. I thought of three different ways to do this separation of tweets, each of which corresponds to a slightly different interpretation of the research question. My initial idea was to bucket the tweets based off of their month. To do this, I would use the data tweeted and group it by month. As a result of this there would be twelve documents, each representing one of the months. I could then do an observational study. I would use Latent Dirichlet Allocation topic modeling in order to look at topics for each month. I would select the number of k topics I was interested in then record the outcome. I would then overlay the topics for each month over a visualization such as the one below to look closer at how topics have changed as the months change and deaths increase (Image 1).

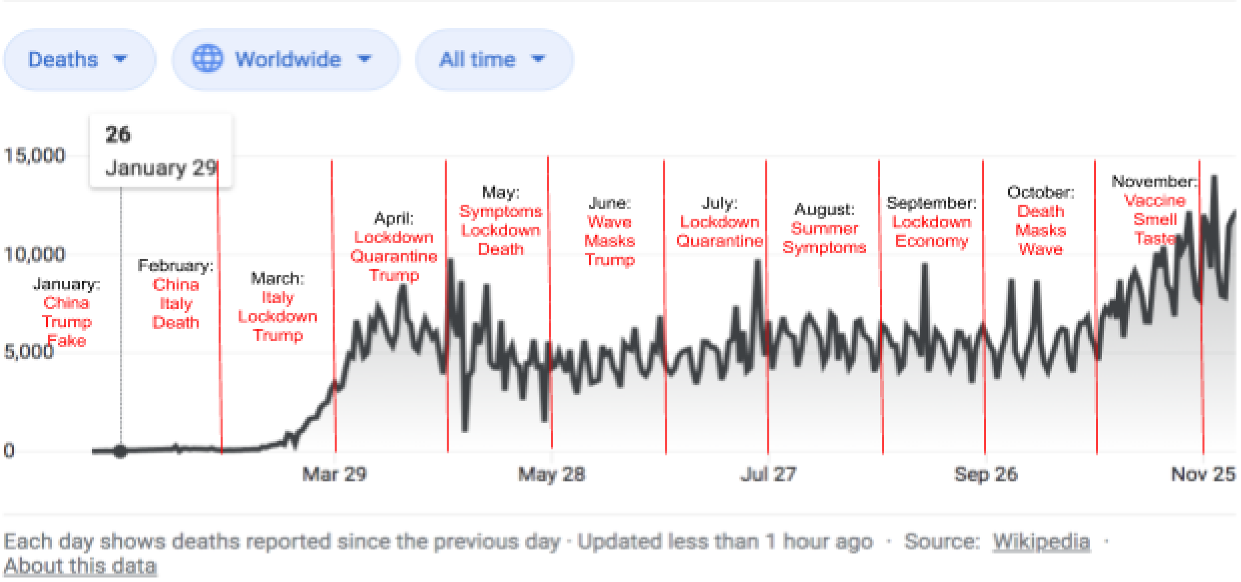
The reason I particularly like this research design is that it is very versatile. In other words, I could use the exact same layout to research how attitudes and thoughts about the Coronavirus may have changed from before the news of a vaccine to after the news of a vaccine. In this case instead of bucketing by 12 months I would bucket all of the tweets into two groups: before and after vaccine announcement. Similarly if I wanted to look more closely at people’s changing thoughts on government, authority, and policy regarding the Coronavirus I might look at before stay-at-home orders versus after stay-at-home orders by state. Each way of bucketing the tweets would produce different outcomes that would successfully show how Coronavirus topics and ideas evolved as the virus evolves.

The second way to answer this research question is by using sentiment analysis. Sentiment analysis helps determine whether the attitude of a tweet is positive, negative or neutral. This would be beneficial if we were particularly interested in looking at overall attitude shifts rather than topic shifts. To do this, I would use a lexicon sentiment analysis package in R or python such as “bing” or “nrc”. These packages have identified unigrams that are associated with negative or positive sentiment. I would then run the sentiment analysis using R to look at a distribution of sentiment. With this sentiment distribution I could do a t-test to compare sentiment before the Moderna and Pfizer vaccine announcement versus after. Similarly I could do a t-test of sentiment distributions before and after stay-at-home orders. Using these t-tests we could see if there is a statistically significant difference between sentiment before the vaccine or stay-at-home announcements versus after.

Either method would be effective in answering the research question. This research might gain insight on how people, particularly twitter users, have responded to the Coronavirus and on a larger scale cope with crises.



Each day shows deaths reported since the previous day · Updated less than 1 hour ago · Source: Wikipedia · About this data



Each day shows deaths reported since the previous day · Updated less than 1 hour ago · Source: Wikipedia · About this data

Ideal Data

In this section I will describe my dream dataset to answer my research questions. In a perfect world I would be able to survey everyone in the world each day of 2020 about how they felt about the pandemic. This way I would know how all people felt about the Coronavirus and not only how twitter users felt. This robust population would allow me to perfectly assess how the world population's feelings changed as the global pandemic evolved. Of course this widespread data collection is far from realistic so I want to conceptualize a more realistic idealized dataset.

To begin collecting my ideal dataset I would go to twitter. I would know how to navigate the twitter API to collect a dataset with all tweets of mention of the Coronavirus, COVID19 and similar terms into one collection. The tweets would ideally be in chronological order and the tweets would be labeled with information on users, their location, and the time and date of the tweet.

The data I originally intended to use (see here)[2] from a link that Professor Rodu provided almost all of these qualifications for my ideal dataset. Unfortunately the data didn't yield any results when I loaded it into R. This was particularly frustrating because I know the information needed to answer at least a version of my research questions would be in the data, however I was unable to reach it. As a novice R user, I could not access the data in a way in which I could actually view any of the contents of any of the tweets. I played around with this data for hours, hoping that I would figure out how to debug my loading process to get the data to work, but unfortunately had no luck. After a lot of research, I concluded that I think the problem is the tweets weren't hydrated. What this means, is that to get the information I wanted it would be necessary to hydrate the tweets in order to get the complete details such as time and date of the post. [3] Unfortunately, I do not know how to hydrate tweets at all so I was unable to utilize this database.

Actual Data

Unfortunately, there were a lot of constraints, particularly relating to data that resulted in me not being able to exactly answer my exact research questions. However, the purpose of this section is to show the data I instead used to answer a variation of my research question. This was a dataset that I found on Kaggle (see here)[4]. This dataset included tweets pertaining to the Coronavirus based on certain hashtags including: #coronavirus, #covid19, #coronavirusPandemic. This particular data excludes retweets for size purposes, but includes a field for whether it is

[2] Top Twitter Datasets for NLP and Machine Learning: https://lionbridge.ai/datasets/top-20-twitter-datasets-for-natural-language-processing-and-machine-learning/

[3] Hydrate Definition: https://stackoverflow.com/questions/34191022/what-does-hydrate-mean-on-twitter/34192633#:~:text=HYDRATE%20%3D%20get%20complete%20details%20(i.e.,call%20to%20get%20these%20fields.

[4] Kaggle Data used for analysis: https://www.kaggle.com/smid80/coronavirus-covid19-tweets

retweeted or not. I think the size of this dataset was much more manageable as compared to the first data set I tried which is why I was able to manipulate this data more easily in R. The columns of the dataset are specified below.

Data Dictionary

User_Name	Creator of the tweet; char
User_Location	Location of user at time of tweet if specified; char
User_Description	Description of user as indicated on twitter profile; char
User_Created	Date the user created their account; date
User_Followers	Number of followers the user has; int
User_Friends	Number of friends the user has; int
User_Verified	Indication of user verification or lack thereof; boolean
Text	Content of the tweet; char
Hashtags	Hashtags included in the tweet; char
Source	Indication of platform/device used to tweet the message (Twitter for Iphone, Twitter web app, etc); char
is_retweet	Indication of retweets or lack thereof; boolean

This data, although extensive in other ways, is lacking some important features that it would require to answer my research question. The first of that feature being a time or date of the tweet. The whole purpose of the research is to see how attitudes or topics change over time. Because these tweets lack timing information it is basically impossible to lay any of the sentiment or LDA findings over the actual pinpoints in Coronavirus announcements. Despite this setback, we can still pull relevant information from our tweets that may answer particular parts of my research question. Luckily, the tweets are in chronological order. Thus it is still possible to look at general change over time, for example the very beginning of the pandemic versus today.

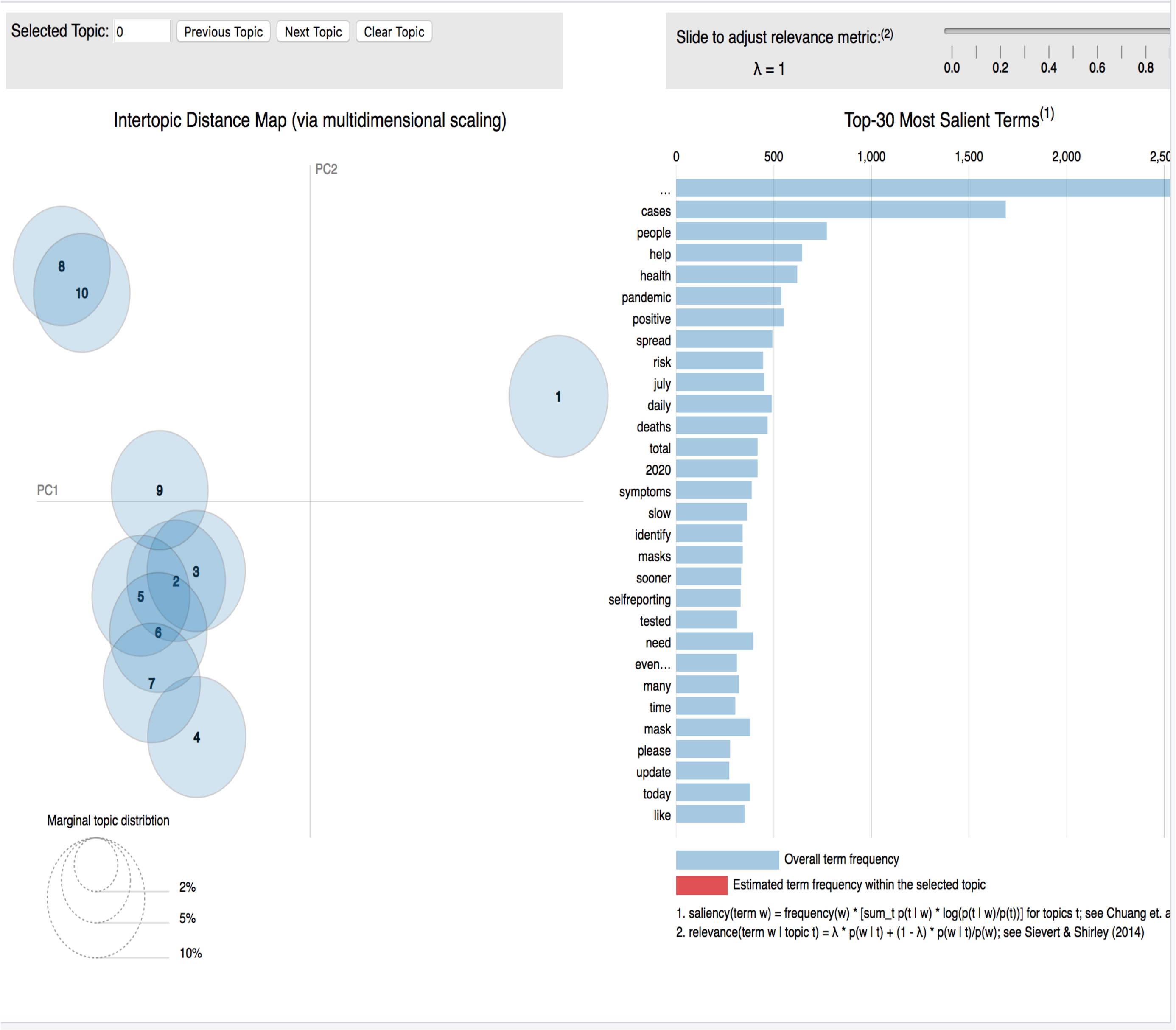
Methodology, Results, and Future Studies

Unfortunately, as I've previously mentioned, I spent a large portion of my time on this project working with a dataset that I never actually was able to derive insights from. Despite this being a large setback for the amount and quality of results I was able to find, it was a great learning experience on working with bigger datasets in R. It also exemplified exactly what Professor Rodu has been emphasizing all semester:

that data cleaning is 90% of the work. Luckily, I was able to get my second dataset to work and was able to derive some results.

Because I wanted to still look at the change in topics and attitude over time without the time data actually available, I relied on the chronological organization of the dataset to break my data up into two groups. The first being the first 50,000 tweets in the set and the second being the last 50,000 tweets in the dataset. I began by cleaning it and removing punctuation, stopwords, numbers, symbols, etc. Then I separated out the first 50,000 tweets to create a document term matrix where I could view the frequency of tokens used in these tweets. I have included an image of the document term matrix in the appendix for reference. Using the document term matrix I ran a Latent Dirichlet Allocation analysis. I then visualized the results in a graphic below. I repeated this process for the last 50,000.

First 50,000 Tweet Results:

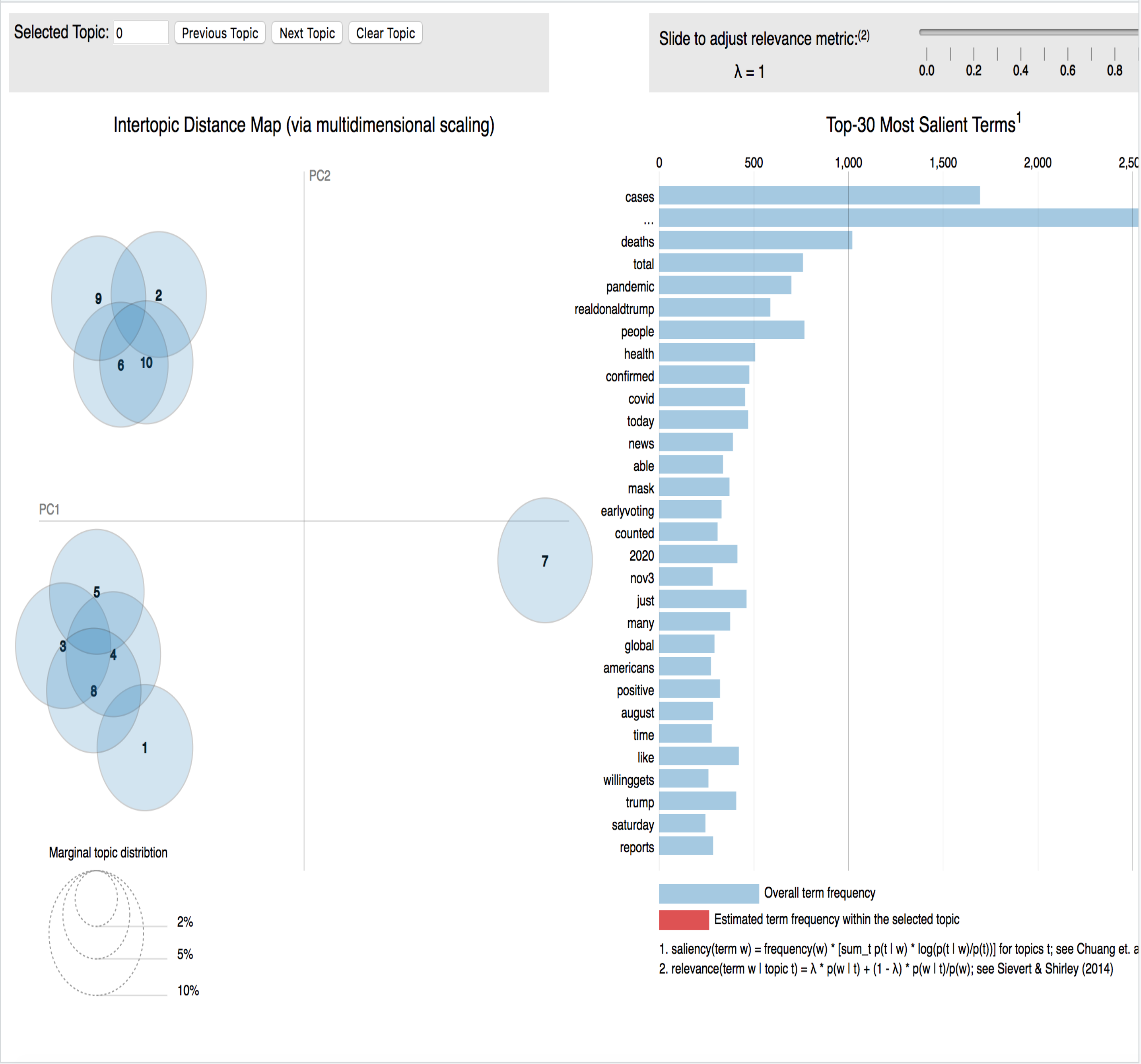


Selected Topic: 0

Slide to adjust relevance metric:(2) λ = 1

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. a
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Last 50,000 Tweet Results:



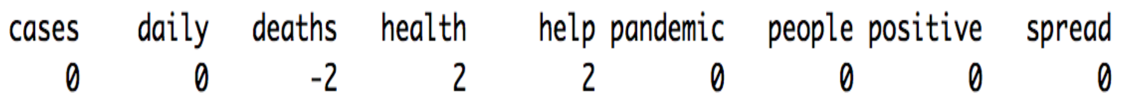
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. a

2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Although this is a rather interesting representation of the tokens and topics, it is not clear cut from the information whether people's attitudes, topics, or thoughts about the Coronavirus definitively changed. However, we still can draw some conclusions. For example, the distribution of topics shifted dramatically from the first to the second and many of the top tokens also changed. Which shows that there was some sort of change in topics and ideas, whether that be due to the coronavirus or not.

The second brief analysis I did was look at the top ten words and do a sentiment analysis. The sentiment analysis I chose to do depended on AFINN lexicon. “The AFINN lexicon assigns words with a score that runs between -5 and 5, with negative scores indicating negative sentiment and positive scores indicating positive sentiment.”[5] From this analysis we saw that the top ten tokens of the first 50,000 had an average score of .2, meaning that it was slightly positive. Whereas the last 50,000 had an average score of 0, which means the sentiment was neutral. This sentiment analysis is extremely elementary, thus I will not spend too much time interpreting the results of it.

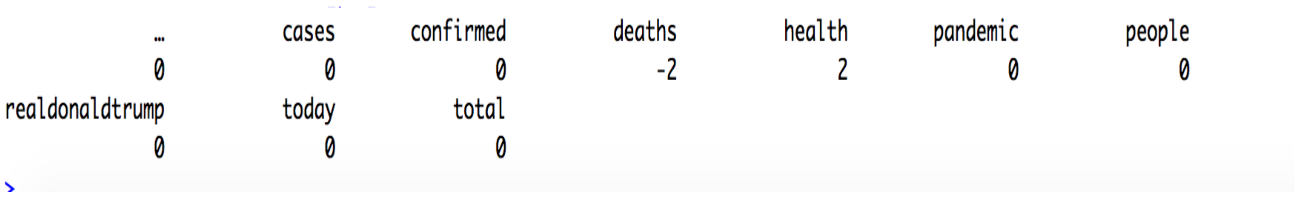
Sentiment Analysis: Using First 50,000 Tweets & AFINN Sentiment Scoring



```
   cases     daily    deaths    health      help pandemic    people positive    spread
       0         0        -2         2         2         0         0         0         0
```

Overall Sentiment: +2, positive

Sentiment Analysis: Last 50,000 Tweets (December 5)



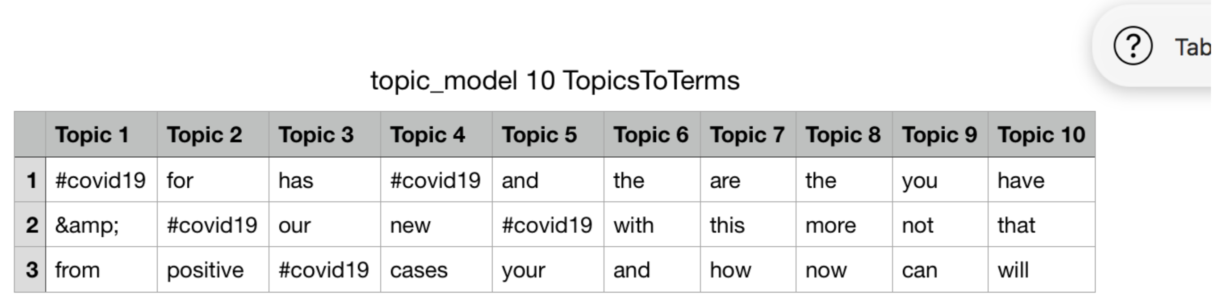
```
              …            cases        confirmed          deaths          health        pandemic          people
              0                0                0              -2               2               0               0
realdonaldtrump            today            total
              0                0                0
>
```

Overall Sentiment: 0, Neutral

These analyses are just touching the surface of valuable insights. As for future studies, it would be interesting to compare the sentiment ebbs and flows of the Coronavirus pandemic to other wide-scaled crises to see if we can pick up patterns on how people’s attitudes change as crises unfold. I would also love to bring an economic component to this analysis and somehow compare individuals of a higher socioeconomic status versus individuals of a lower one. One way to go about this using this particular dataset may be to look at verified users versus unverified users. The Coronavirus has undoubtedly disproportionately negatively impacted people of lower socioeconomic status, thus it is always beneficial to consider these factors when doing analyses.

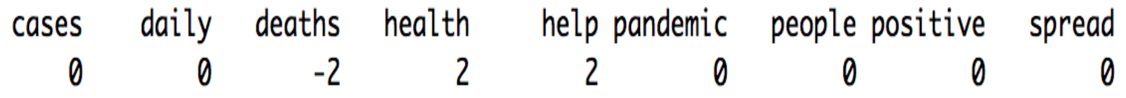
[5] https://www.tidytextmining.com/sentiment.html

Appendix:

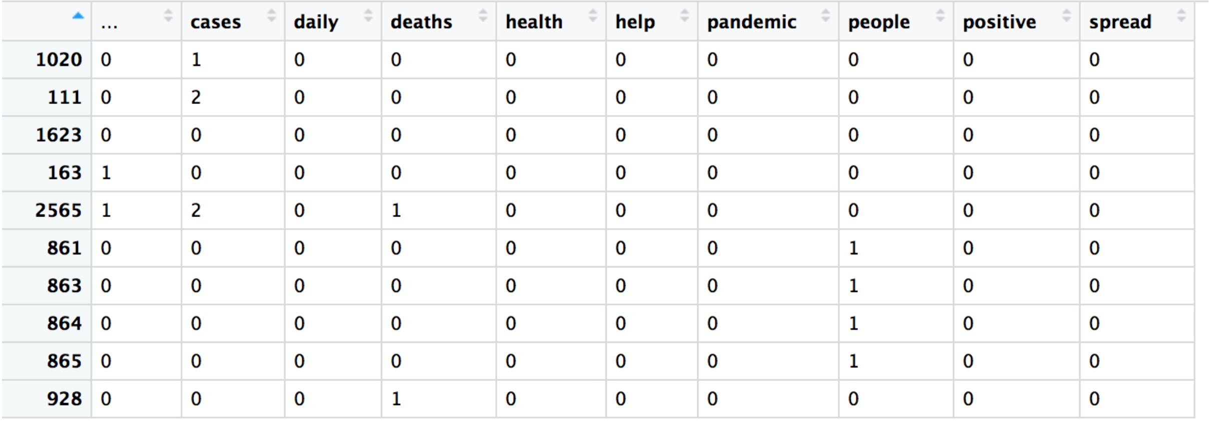


topic_model 10 TopicsToTerms

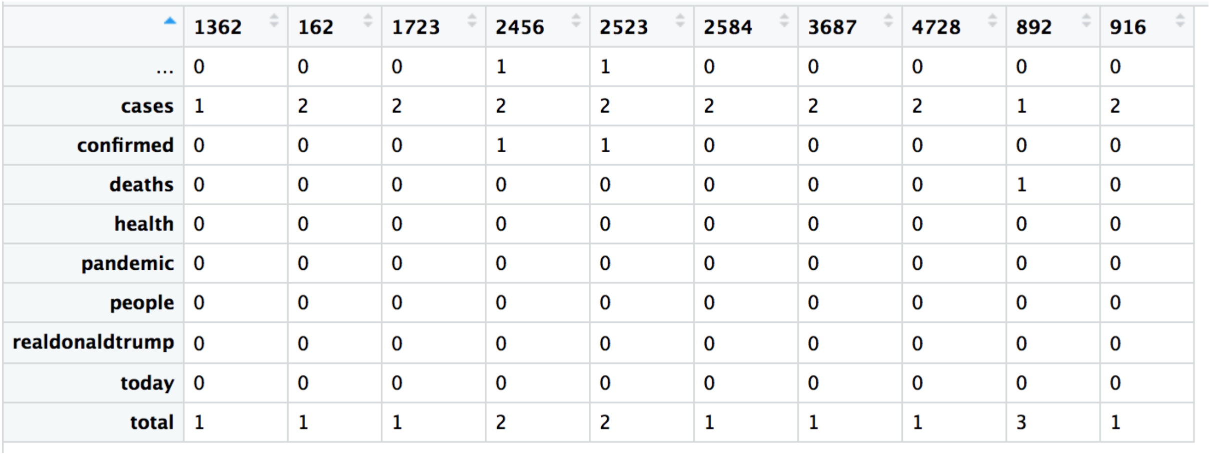
	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
1	#covid19	for	has	#covid19	and	the	are	the	you	have
2	&amp;	#covid19	our	new	#covid19	with	this	more	not	that
3	from	positive	#covid19	cases	your	and	how	now	can	will



```
   cases    daily   deaths   health     help pandemic   people positive   spread 
       0        0       -2        2        2        0        0        0        0 
```



	...	cases	daily	deaths	health	help	pandemic	people	positive	spread
1020	0	1	0	0	0	0	0	0	0	0
111	0	2	0	0	0	0	0	0	0	0
1623	0	0	0	0	0	0	0	0	0	0
163	1	0	0	0	0	0	0	0	0	0
2565	1	2	0	1	0	0	0	0	0	0
861	0	0	0	0	0	0	0	1	0	0
863	0	0	0	0	0	0	0	1	0	0
864	0	0	0	0	0	0	0	1	0	0
865	0	0	0	0	0	0	0	1	0	0
928	0	0	0	1	0	0	0	0	0	0



	1362	162	1723	2456	2523	2584	3687	4728	892	916
...	0	0	0	1	1	0	0	0	0	0
cases	1	2	2	2	2	2	2	2	1	2
confirmed	0	0	0	1	1	0	0	0	0	0
deaths	0	0	0	0	0	0	0	0	1	0
health	0	0	0	0	0	0	0	0	0	0
pandemic	0	0	0	0	0	0	0	0	0	0
people	0	0	0	0	0	0	0	0	0	0
realdonaldtrump	0	0	0	0	0	0	0	0	0	0
today	0	0	0	0	0	0	0	0	0	0
total	1	1	1	2	2	1	1	1	3	1

```
# Project 1 Snippets of Code
# Kylie Wise - Jordan Rodu

install.packages("readtext")
install.packages("tm")
install.packages("NLP")
install.packages("quanteda")
install.packages("topicmodels")
install.packages("stopwords")
install.packages("partition")
install.packages("tidytext")
library(readtext)
library(tm)
library(NLP)
library(tidyverse)
library(quanteda)
library(topicmodels)
library(stopwords)
library(partition)
library(tidytext)
library(ggplot2)
library(dplyr)

library(tidyverse)
library(lubridate)

setwd("~/Desktop/Text Analysis/dataverse_files")


# Retried this from the first time that I did a similar thing in homework 2. Didnt work
when I tried it in homework two, however
# it worked this time, which is nice. However opted with a different strategy in the
second part just so that it would align
# with the homework I did in project 2 to make the entire proccess much smoother.

all_files <- list.files("~/Desktop/Text Analysis/dataverse_files", pattern = "*.txt",
full.names = TRUE, recursive = TRUE)
texts1 <- sapply(all_files, function(x) readLines(x, encoding = "UTF-8"))
texts1 <- lapply(texts, function(x) paste(x, collapse = " " ))

# This is the option I ended up using to load the data (see explanation above to as
why I did this).
texts_final1 <- readtext(all_files, ignore_missing_files = FALSE, docvarsfrom =
"filepaths", dvsep = "/")
texts_final_as_string <- as.String(as.character(texts_final1))

# In this section of code I am cleaning up the tweets, hopefully (if I'm doing it
correctly) I'm removing all numbers,
# punctuation, symbols, and urls. I understand many of these things may be
important to an analysis, particularly of tweets
```

```
# however with my limited knowledge of R it is important that I am using the most
simplified version of what I want to do to avoid errors
# and making things more complicated than they need to be.
corpus_list <- Corpus(texts_final1$text, texts_final1$doc_id)
token_list <- tokens(corpus_list, remove_numbers = TRUE, remove_punct = TRUE,
                  remove_symbols = TRUE, remove_separators = TRUE, remove_url =
TRUE)
tokens_lower <- tokens_tolower(token_list)

# Here I am going to go in and take out some stop words. Especially because I am
going to be doing some topic modeling
# I think it will be important to remove some of these words so we can actually get a
feel for what is genuinely being discussed.
# I know have done the analysis and am going back to add additional stop words
that came up.

sw <- stopwords(language = "en", source = "nltk") # stopwards from pythons nltk
package
token_list_wosw <- tokens_remove(tokens_lower, sw)
additional_stop <-
c("also","could","year","people","said","first","last","people","would","one",
                  "time","years","would", "best","get","made","make",

"like","many","new","next","told","two","three","world","take","way","set","back","adde
d","bbc",
                  "says","well","good","may","going","number","still") # went back and did
this after I looked at my dtm


texts_final_as_vector <- VectorSource(token_list_wosw)
texts_final_as_corpus <- VCorpus(texts_final_as_vector)
dt_matrix <- DocumentTermMatrix(texts_final_as_corpus)

dtm <- inspect(dt_matrix)
rowTotals <- apply(dtm , 1, sum) #Find the sum of words in each Document
dtm.new   <- dtm[rowTotals> 0, ]


k <- 10
LDA_attempt <- LDA(dtm, k =k, control = list(seed = 10000000))
# found this next part online and copied and pasted it basically
ap_topics <- tidy(LDA_attempt, matrix = "beta")
ap_top_terms <- ap_topics %>%
  group_by(topic) %>%
  top_n(10, beta) %>%
  ungroup() %>%
  arrange(topic, -beta)

ap_top_terms %>%
  mutate(term = reorder_within(term, beta, topic)) %>%
  ggplot(aes(term, beta, fill = factor(topic))) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~ topic, scales = "free") +
  coord_flip() +
  scale_x_reordered()


# Second Attempt Project Final
install.packages("tm")
install.packages("topicmodels")
install.packages("LDAvis")
install.packages("servr")
install.packages("stringi")
library(tm)
library(topicmodels)
library(LDAvis)
library(servr)
library(dplyr)
library(stringi)
new_tweets <- read.csv("covid19_tweets.csv")
tweets <- new_tweets[0:3000,]

tweet_content <- iconv(tweets$content)
corpus <- Corpus(VectorSource(tweet_content))
dtm <- DocumentTermMatrix(corpus)
rowTotals<-apply(dtm,1,sum) #running this line takes time
empty.rows<-dtm[rowTotals==0,]$dimnames[1][[1]]
corpus<-corpus[-as.numeric(empty.rows)]
dtm <- DocumentTermMatrix(corpus)
inspect(dtm)


dtm.mx <- as.matrix(dtm)
frequency <- colSums(dtm.mx)
frequency <- sort(frequency, decreasing=TRUE)
frequency[1:25]
k <- 10
LDA_attempt <- LDA(dtm, k =k, control = list(seed = 10000000))
```

```
# found this next part online and copied and pasted it basically
ap_topics <- tidy(LDA_attempt, matrix = "beta")
ap_top_terms <- ap_topics %>%
  group_by(topic) %>%
  top_n(10, beta) %>%
  ungroup() %>%
  arrange(topic, -beta)

ap_top_terms %>%
  mutate(term = reorder_within(term, beta, topic)) %>%
  ggplot(aes(term, beta, fill = factor(topic))) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~ topic, scales = "free") +
  coord_flip() +
  scale_x_reordered()
```