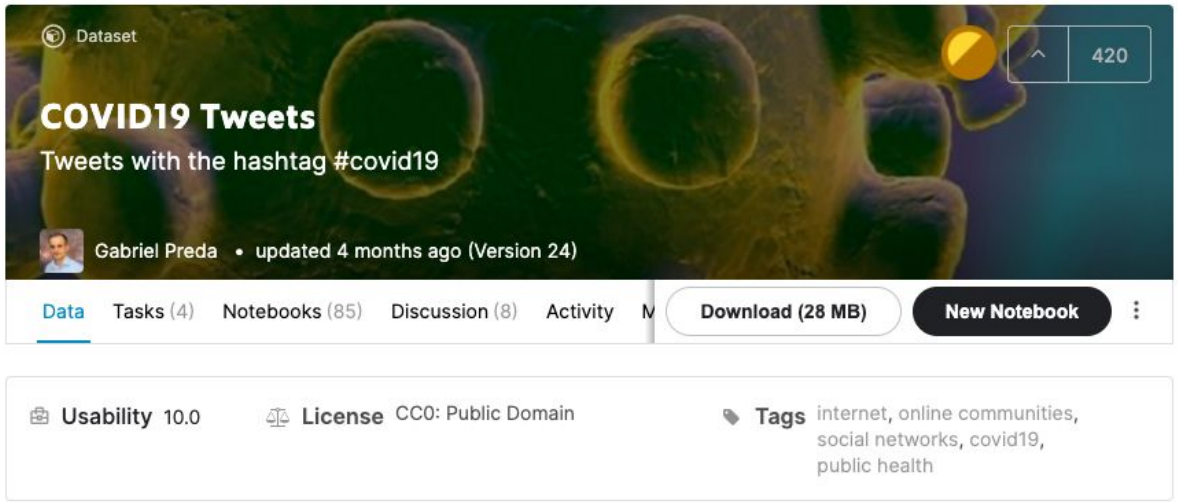
Topic Modeling & Text Mining of COVID-19 Tweets

By: Abby, Jon, and Shivani

Dataset Background & Hypotheses

Data Overview

- COVID-19 Tweets
- User information
  - Location
  - Date and time
  - Bio
  - Verified
  - Followers
  - Friends
- Body text of the tweet
- July 24th 2020 - August 30th 2020

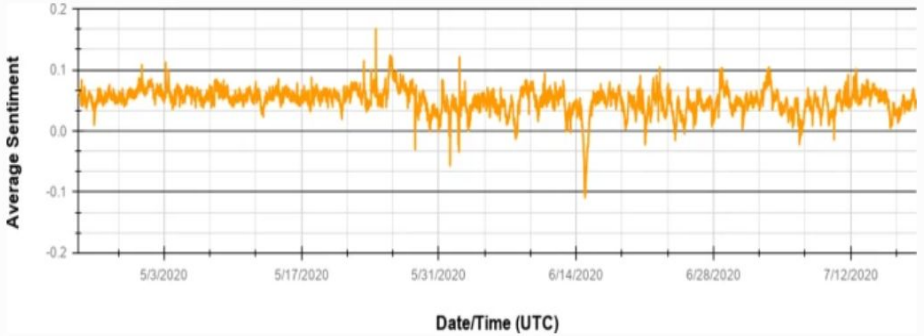


Previous Studies

Design and Analysis of a Large-Scale COVID-19 Tweets

- **Background:** Exploring the context of COVID tweets on various social media platforms and how people sharing their safety and updates on social media leads to a large repository of info
- **Sentiment Analysis:** there are many drops in the average sentiment over the analysis period. As seen in the figure below, there are 14 drops where the scores are negative.
- **Results:** only 141k tweets (0.045%) had “point” location data in the metadata, 7 significant drops found in sentiment analysis

Fig. 4



COVID-19 sentiment trend, since April 24, 2020 to July 17, 2020

Hypotheses

HEALTH AND SCIENCE

Trump says U.S. has reached deal with Moderna for 100 million doses of coronavirus vaccine

PUBLISHED TUE, AUG 11 2020•5:50 PM EDT | UPDATED WED, AUG 12 2020•7:37 AM EDT

Noah Higgins-Dunn
@HIGGINSDUNN

SHARE

Alternative Hypothesis 1: A two-samples t-test performed on afinn sentiment analysis values will show significantly more positive afinn sentiment values after August 11th as compared to before August 11th.

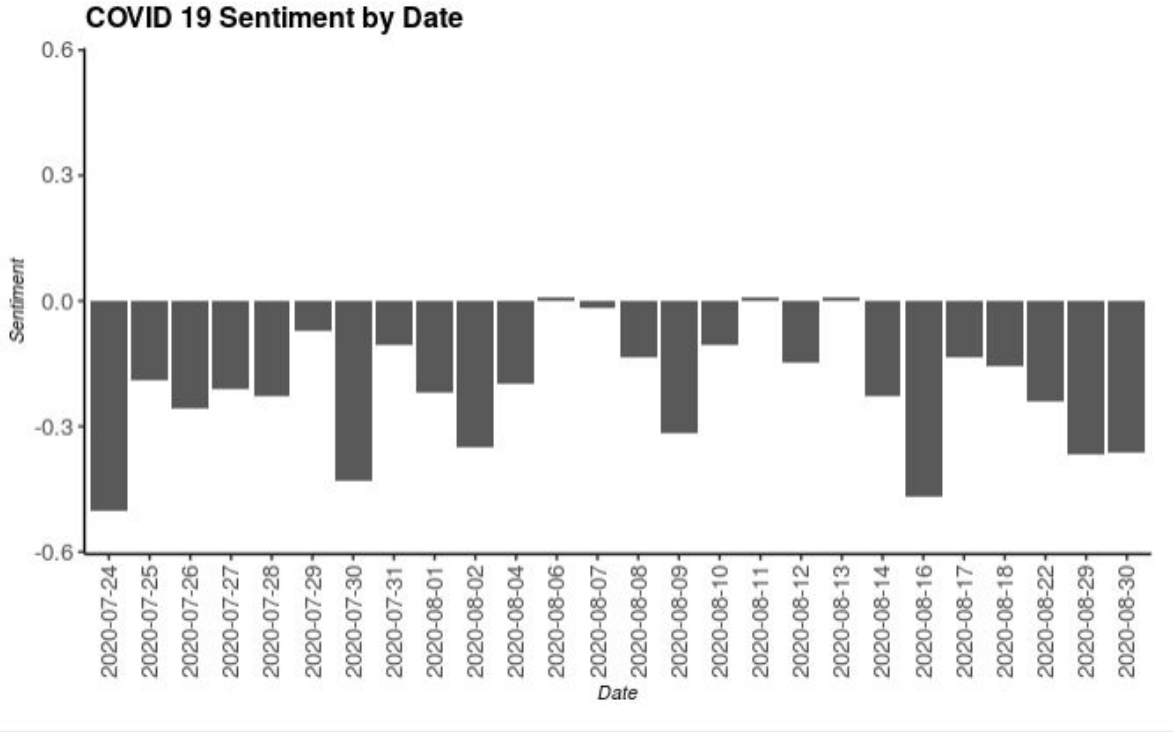
Null Hypothesis 1: An two-samples t-test performed on afinn sentiment analysis values will show no significant difference in afinn sentiment values after August 11th as compared to before August 11th.

Alternative Hypothesis 2: If we partition the data into tweets before and after August 11th the topics created will differ in the two topic models.

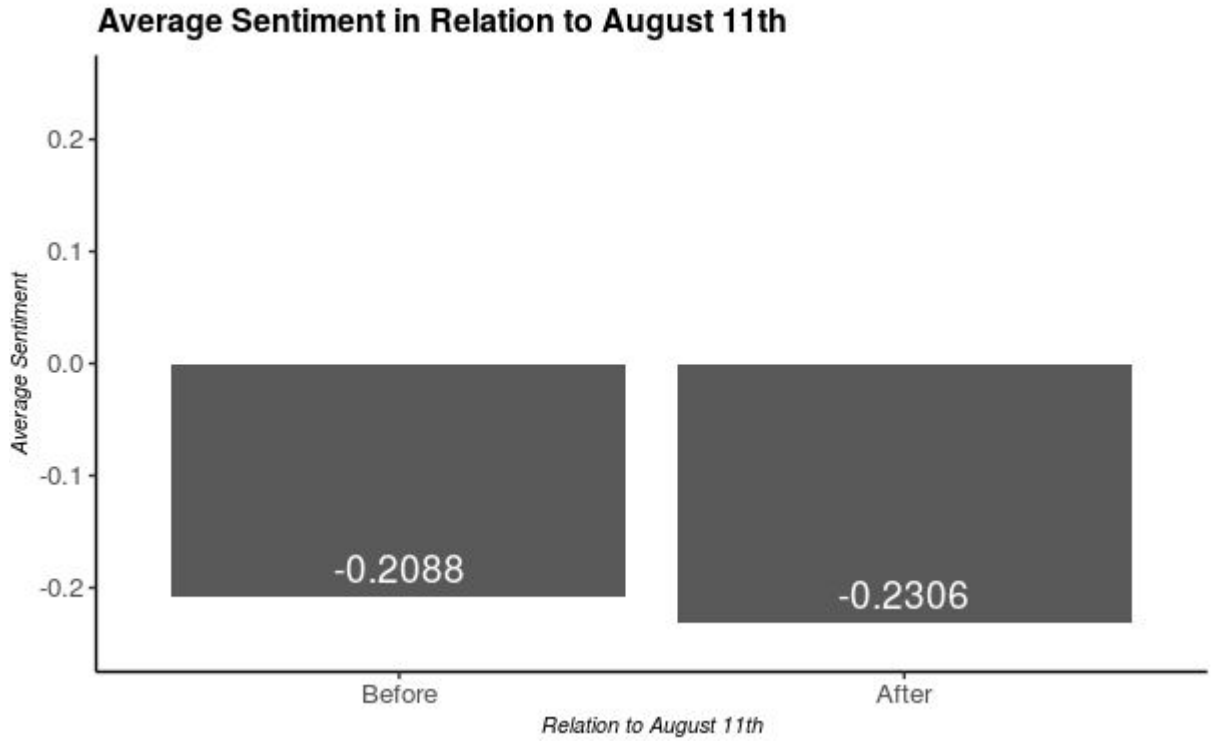
Null Hypothesis 2: If we partition the data into tweets before and after August 11th the topics created will not differ in the two topic models.

EDA Highlights

What was the average sentiment per day?



What was the average sentiment before and after August 11th?



Word Clouds before and after August 11th

Before

After

Model Construction

Unpaired two-samples t-test

Conditions:

- Normal distribution (Shapiro-Wilk Normality Test)
  - Before Aug. 11th data:
    - P-value: 0.8104
  - After Aug. 11th data:
    - P-value: 0.8218
  - Data not significantly different from normal distribution
- Variances of two groups equal (F-test)
  - P-value = 0.84
  - Failed to reject null → variances are equal

Results:

- Failed to reject null hypothesis that average afinn sentiment value before August 11th is equivalent to the average afinn sentiment value after August 11th.
  - P-value = 0.6878

Topic Modeling Background

- Unsupervised machine learning
- Two components
  - The probability that a word belongs to a document
  - The probability that each document will belong to a topic
- Considered Mixture model
- Parameter k → number of topics a model will output
  - Can be tuned, but large numbers of topics are difficult to interpret
- Topics are composed of words, not given labels
- LDA
  - Latent Dirichlet Allocation

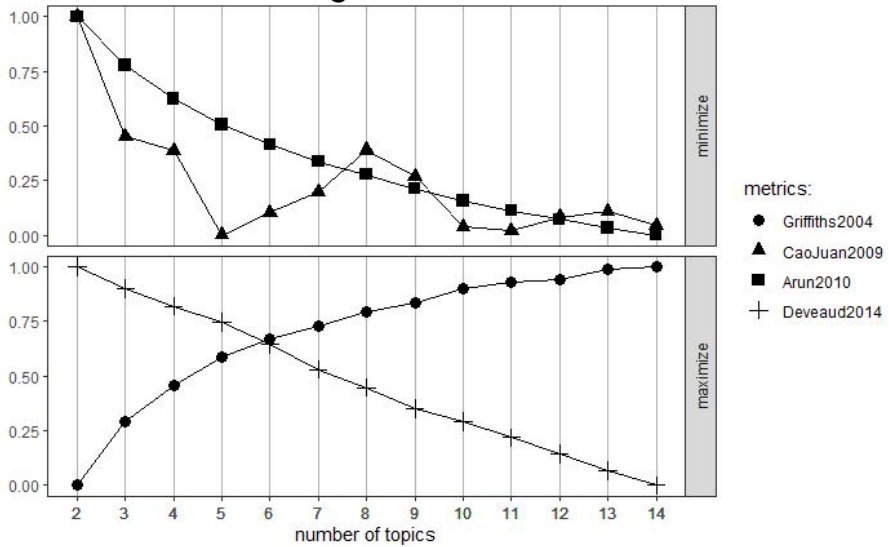
Metrics for optimal number of topics used in ldatuning

- **Metrics to minimize:**
  - CaoJuan2009
  - Arun2010
- **Metrics to maximize:**
  - Griffiths2004
  - Deveaud2014

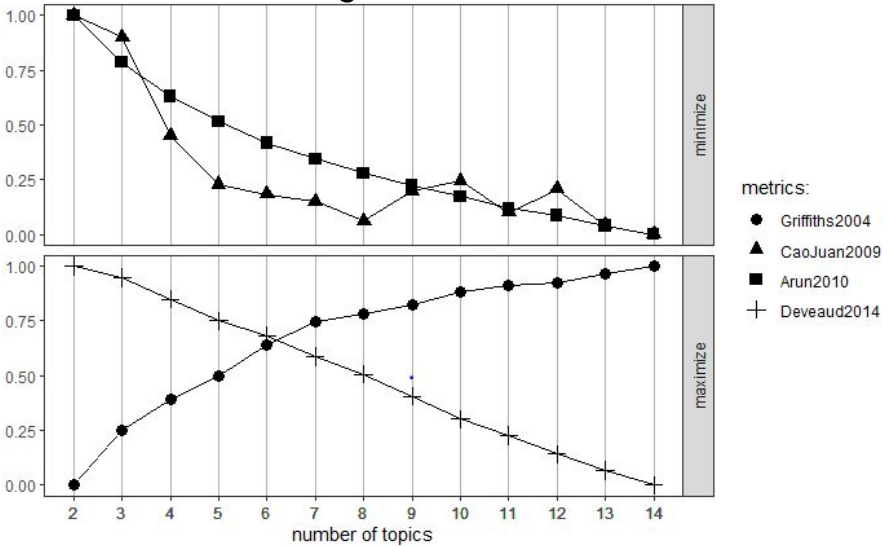
Finding the optimal number of topics

- Used ldatuning package
- Metrics are labelled with their corresponding papers

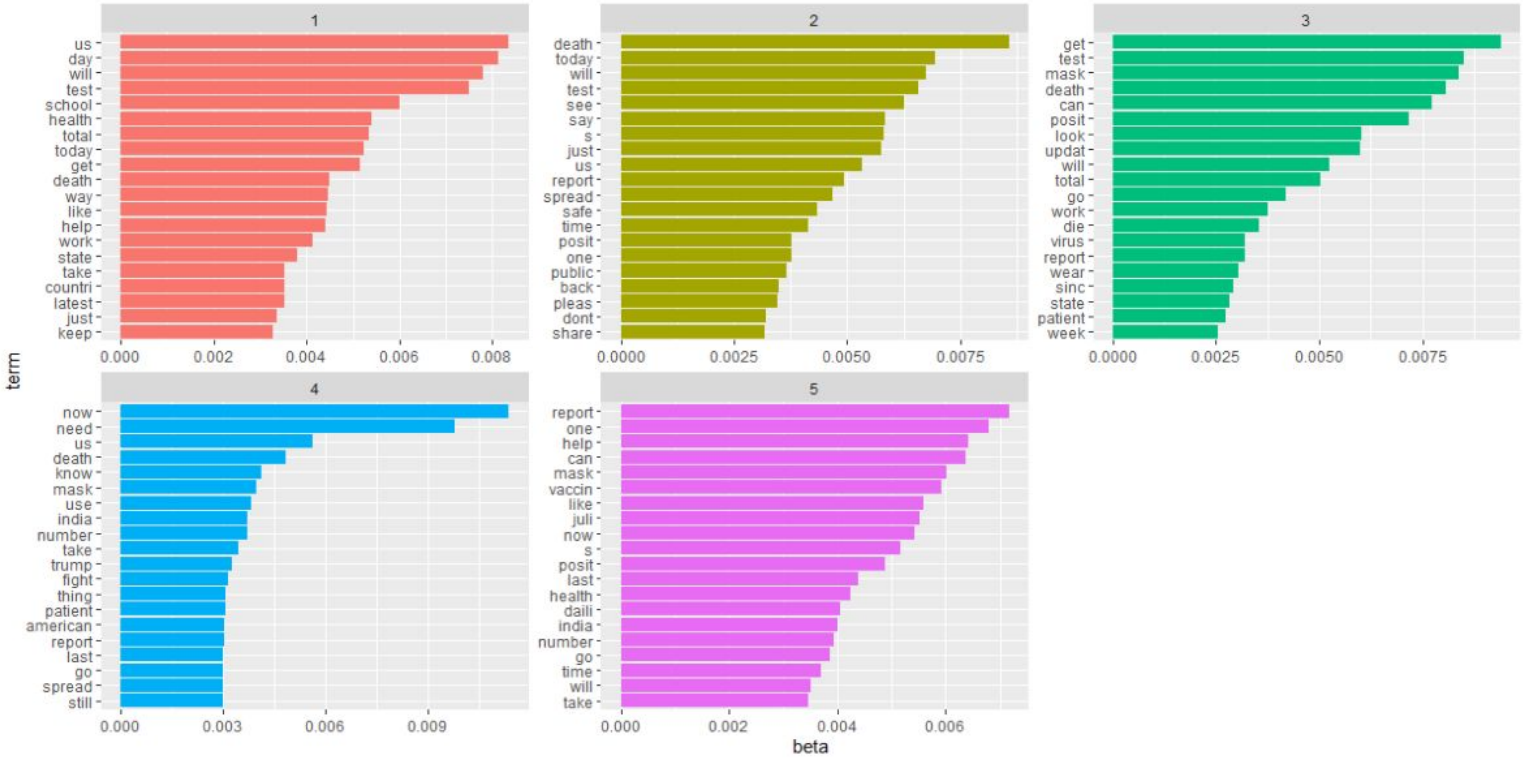
Before August 11th



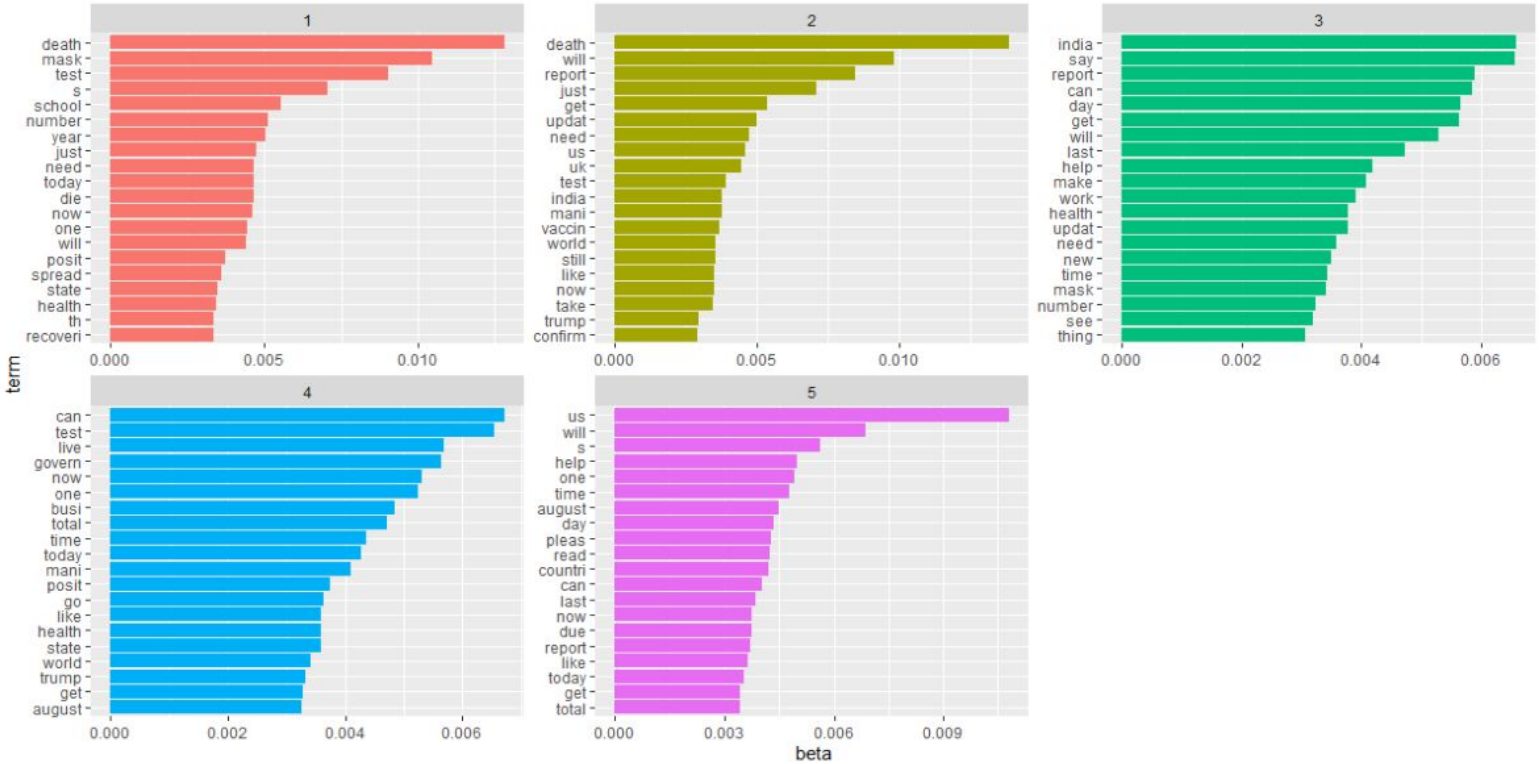
After August 11th



Final Topic Model with 5 Topics: Before August 11th



Final Topic Model with 5 Topics: After August 11th



Insights

Before August 11th

- Topic 1: A lot to due with Institutions
- Topic 2: A lot to due with Public Health
- Topic 3: Reports/Updates
- Topic 4: Politics
- Topic 5: Miscellaneous

After August 11th

- Topic 1: A lot to due with Institutions
- Topic 2: Global view of coronavirus
- Topic 3: Remote work
- Topic 4: Government/Politics
- Topic 5: Reports/Updates

This topic labels are open to interpretation, and were difficult to determine due to the commonalities between the words included

Topic Model Evaluation: Perplexity

- Perplexity
  - How surprised a model is of data it has not seen before
  - Best model assigns a high probability to the test set

$$PP(W) = \sqrt[N]{\frac{1}{P(w_1, w_2, \ldots, w_N)}}$$

- Before August 11th Topic Model Perplexity: 7870.432
- After August 11th Topic Model Perplexity: 7408.929

Perplexity was slightly better for the after August 11th model

Topic Model Evaluation: Kappa

- Measure of model giving the same score to the same data item
- Comparison of Topics 1-5 between before Aug. 11th and after Aug. 11th

Topic	Kappa
1	0.06
2	0.143
3	0.309
4	-0.29
5	0.559

Hypotheses

Alternative Hypothesis 1: A two-samples t-test performed on afinn sentiment analysis values will show significantly more positive afinn sentiment values after August 11th as compared to before August 11th.

- Failed to reject null hypothesis, thus alternative hypothesis 1 is not supported by our findings

Alternative Hypothesis 2: If we partition the data into tweets before and after August 11th the topics created will differ in the two topic models.

- Low kappa scores in between models for topics 1-4 indicate that the topics created are in disagreement with one another, however topic 5 kappa score indicates moderate agreement
    - Limitations of kappa score may make the score appear lower than actual

Limitations

- Only the tweets about covid with the #covid19 were captured by this analysis
- The Moderna vaccine purchase was not announced on twitter therefore an analysis of tweets might not fully capture the reaction
- **Demographic:** The demographics that Twitter captures may not be equally representative of individuals who are pro/anti-vaccination.
- **Discerning Topics:** The number of topics in the dataset are specified by the user, or based on some distribution such as Poisson by sampling, which is subjective and doesn't always show the true distribution of topics.
- **Static:** no evolution of topics over time
- **Capturing Correlations:** Dirichlet topic distribution can't capture correlations well
- Potential qualitative interpretation may have switched between models (limitations of kappa)
- Topic modelling is dependent on interpretations

Future Steps

- Analysis of covid tweets for a longer time period in order to determine if our results were statistically significant in a larger dataset
- Compare sentiment for a covid time period with a pre-covid twitter to see how sentiment differs
- Conduct a global sentiment analysis of COVID-19 tweets to see how responses differ country by country potentially based on the country's regulation of and response to the pandemic
- Sentiment analysis/topic modeling of COVID-19 related news articles before and after August 11th
- Use a more sophisticated sentiment measure that takes into account context rather than single words

Sources

- https://www.rdocumentation.org/packages/topicmodels/versions/0.2-11/topics/perplexity
- https://towardsdatascience.com/evaluate-topic-model-in-python-latent-dirichlet-allocation-lda-7d57484bb5d0
- https://s3.amazonaws.com/assets.datacamp.com/production/course_11368/slides/chapter4.pdf
- https://bookdown.org/Maxine/tidy-text-mining/latent-dirichlet-allocation.html
- https://medium.com/pew-research-center-decoded/making-sense-of-topic-models-953a5e42854e
- https://www.tidytextmining.com/topicmodeling.html
- https://www.statisticshowto.com/cohens-kappa-statistic/
- https://rpubs.com/nikita-moor/107657
- https://www.pnas.org/content/pnas/101/suppl_1/5228.full.pdf
- https://rstudio-pubs-static.s3.amazonaws.com/66739_c4422a1761bd4ee0b0bb8821d7780e12.html