Discussing the Covid19 Pandemic

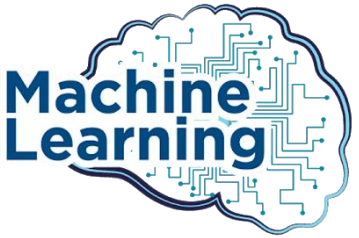
Capstone – Final Presentation

By:

Shreyas Chitransh

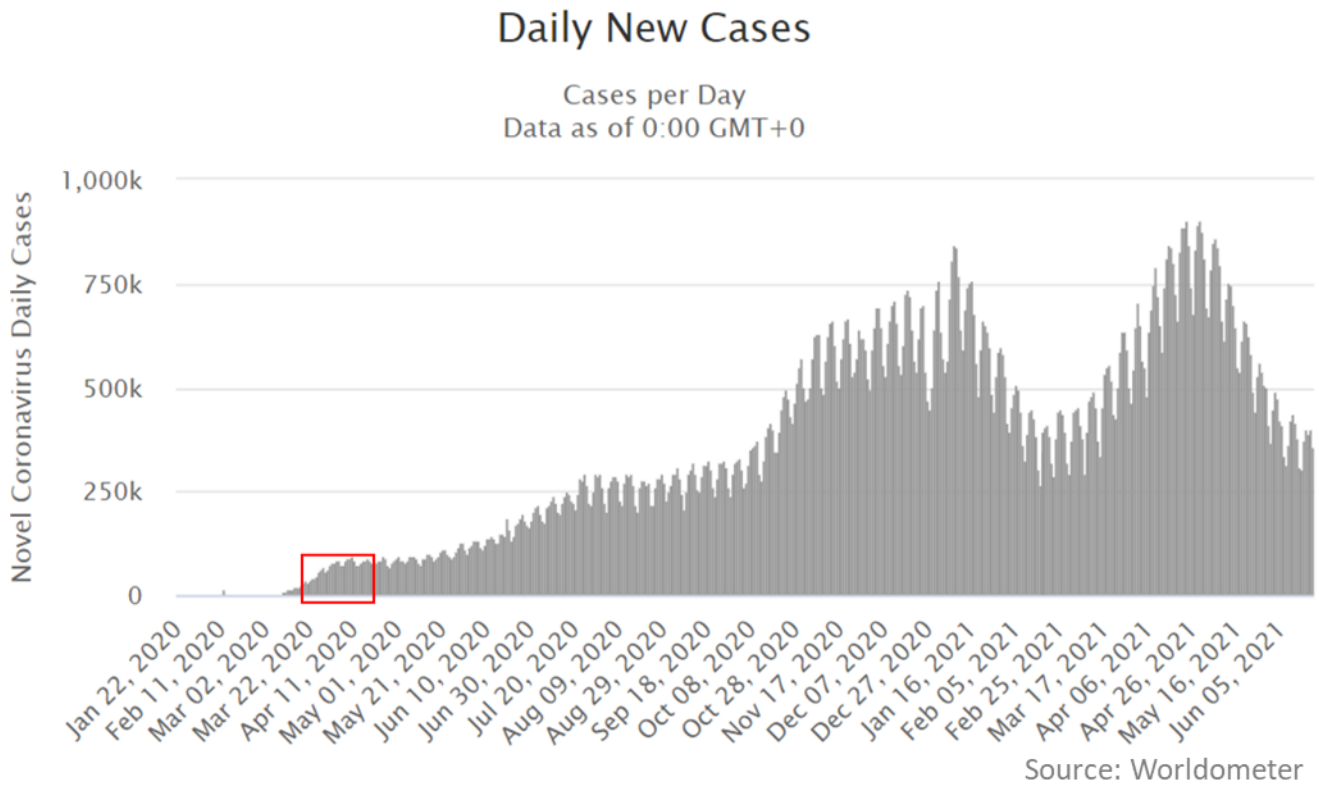
Agenda

- Background
  - Objectives
  - Methodology
- Process:
  - Data gathering and Cleaning
  - Model Optimization
- Results
- Conclusions & Further Work



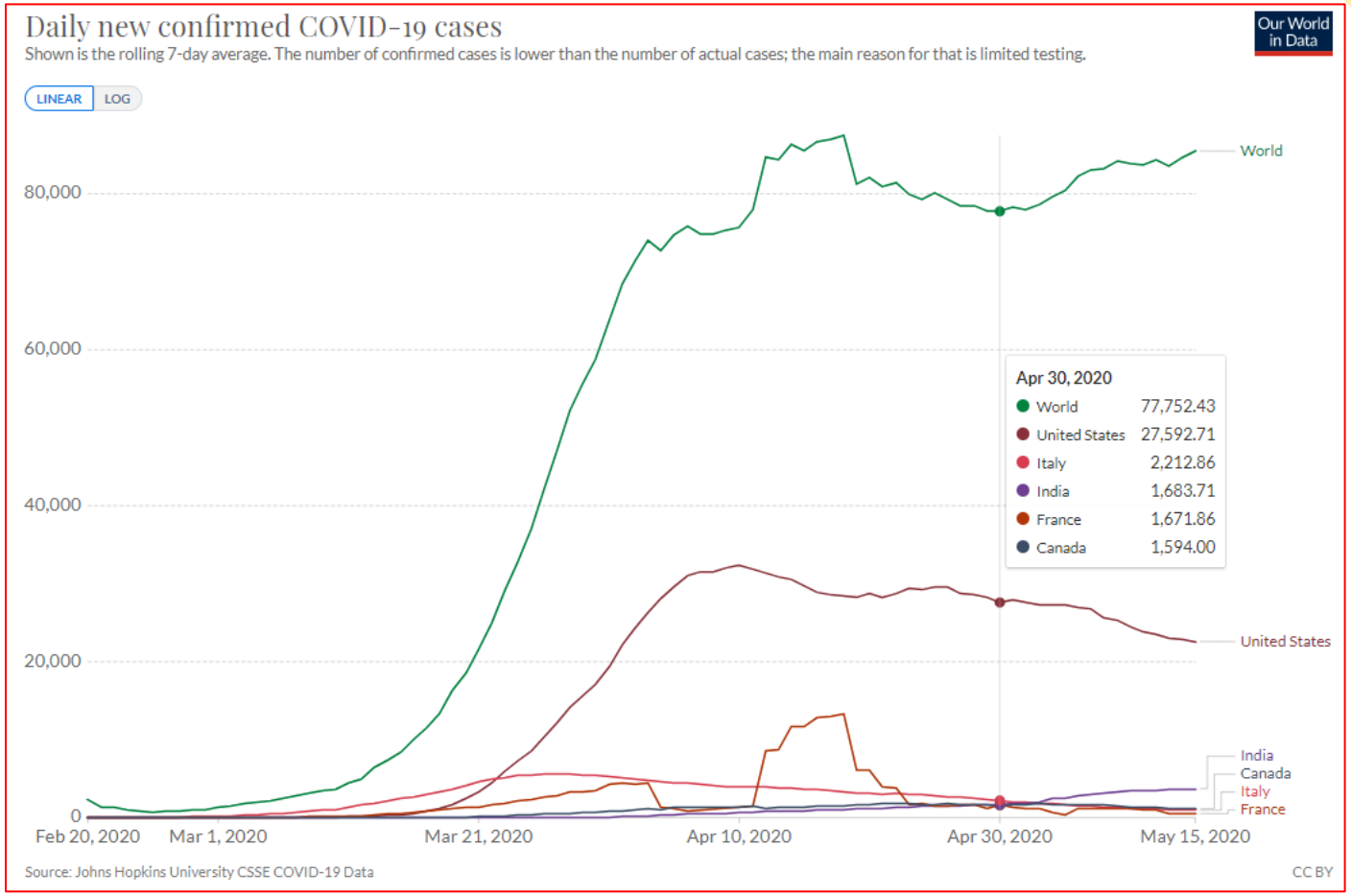
Background

- Covid-19 culminated in fear and grief but also some of the best solidarity amongst the global citizens;
- Goal:
  - To automatically identify the underlying topics being discussed with context to Covid19;
- Methodology:
  - Analyze Twitter data from the first global surge of Covid19 cases using Unsupervised ML;
  - Topic Modeling - Latent Dirichlet Allocation (LDA)
  - 28th March – 30th April 2020



Background

- Covid-19 culminated in fear and grief but also some of the best solidarity amongst the global citizens;
- Goal:
  - To automatically identify the underlying topics being discussed with context to Covid19;
- Methodology:
  - Analyze Twitter data from the first global surge of Covid19 cases using Unsupervised ML;
  - Topic Modeling - Latent Dirichlet Allocation (LDA)
  - 28th March – 30th April 2020



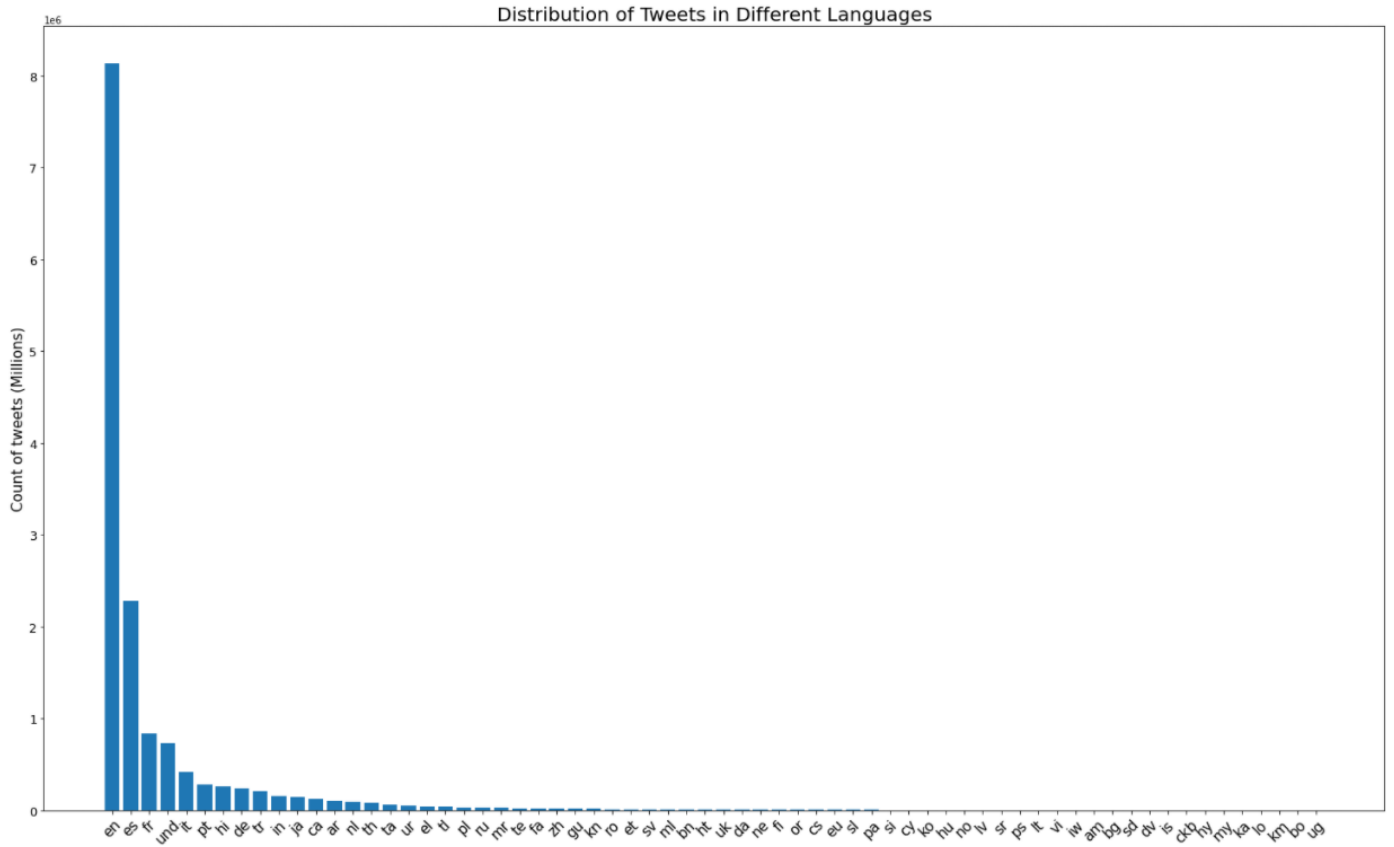
Process - Data Gathering

- Kaggle dataset >4GB of Tweets spanning 28th March – 30th April 2020;
  - CSV Files
- >14M Rows and 22 Columns
  - Rows : Separate Tweets
  - Columns : Tweet and User Metadata
- Metadata:
  - Tweet:
    - Status ID, Date, Time, Favorites, Retweets , Autodetected language etc.
  - User:
    - User ID, Screen Name, Account Creation Date etc.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 14607013 entries, 0 to 14607012
Data columns (total 22 columns):
 #   Column                 Dtype
---  ------                 -----
 0   status_id              int64
 1   user_id                int64
 2   created_at             object
 3   screen_name            object
 4   text                   object
 5   source                 object
 6   reply_to_status_id     float64
 7   reply_to_user_id       float64
 8   reply_to_screen_name   object
 9   is_quote               bool
 10  is_retweet             bool
 11  favourites_count       int64
 12  retweet_count          int64
 13  country_code           object
 14  place_full_name        object
 15  place_type             object
 16  followers_count        int64
 17  friends_count          int64
 18  account_lang           float64
 19  account_created_at     object
 20  verified               bool
 21  lang                   object
dtypes: bool(3), float64(3), int64(6), object(10)
memory usage: 2.1+ GB
```

Data Clean-Up

- Clean up and Preprocessing most important steps for Unsupervised NLP
- Cleaning:
  - Removed Duplicates
  - Removed columns containing null values (over 80% null)
  - For remaining data, removed rows containing null values (884 rows, <0.05%)
  - Removed rows containing non-English Tweets
  - Final English Tweets --> 8.1M



Analysis Dataset – Improve Manageability

Original Dataset

14.6M Rows with 22 Columns

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 14607013 entries, 0 to 14607012
Data columns (total 22 columns):
 #   Column                Dtype
---  ------                -----
 0   status_id             int64
 1   user_id               int64
 2   created_at            object
 3   screen_name           object
 4   text                  object
 5   source                object
 6   reply_to_status_id    float64
 7   reply_to_user_id      float64
 8   reply_to_screen_name  object
 9   is_quote              bool
 10  is_retweet            bool
 11  favourites_count      int64
 12  retweet_count         int64
 13  country_code          object
 14  place_full_name       object
 15  place_type            object
 16  followers_count       int64
 17  friends_count         int64
 18  account_lang          float64
 19  account_created_at    object
 20  verified              bool
 21  lang                  object
dtypes: bool(3), float64(3), int64(6), object(10)
memory usage: 2.1+ GB
```

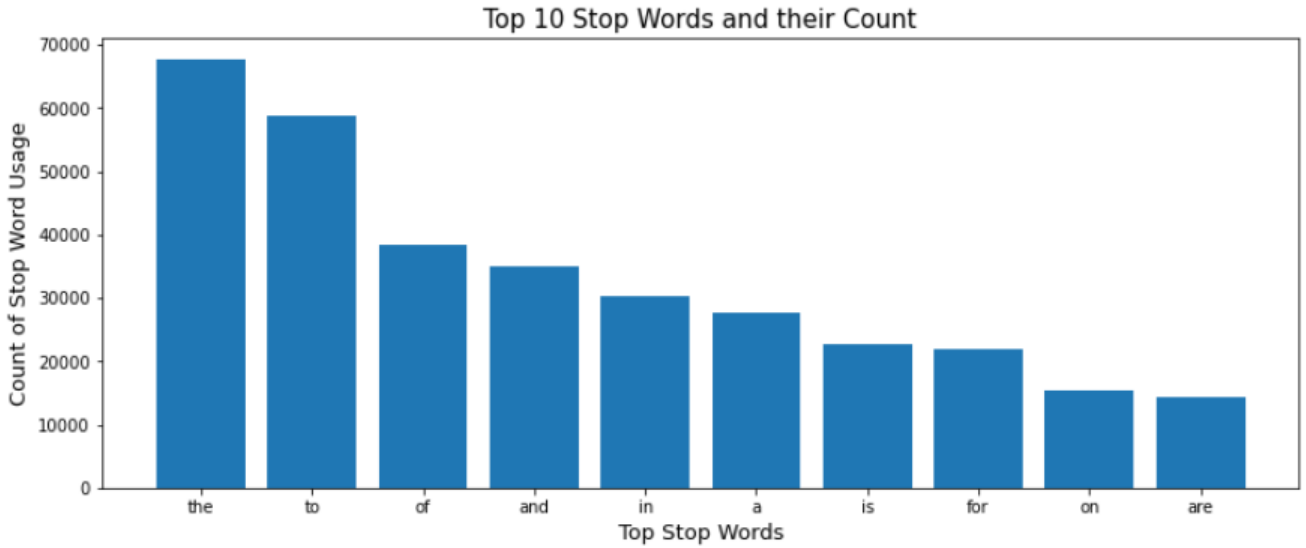
---> 8.1M English Tweets with 9 Columns --->

Analysis Dataset

81.3K Rows with 9 Columns

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 81333 entries, 0 to 81332
Data columns (total 9 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   created_at        81333 non-null  datetime64[ns, UTC]
 1   screen_name       81333 non-null  object
 2   text              81333 non-null  object
 3   is_retweet        81333 non-null  bool
 4   favourites_count  81333 non-null  int64
 5   retweet_count     81333 non-null  int64
 6   followers_count   81333 non-null  int64
 7   friends_count     81333 non-null  int64
 8   verified          81333 non-null  bool
dtypes: bool(2), datetime64[ns, UTC](1), int64(4), object(2)
memory usage: 4.5+ MB
```

Preprocessing (Crucial Step) & EDA

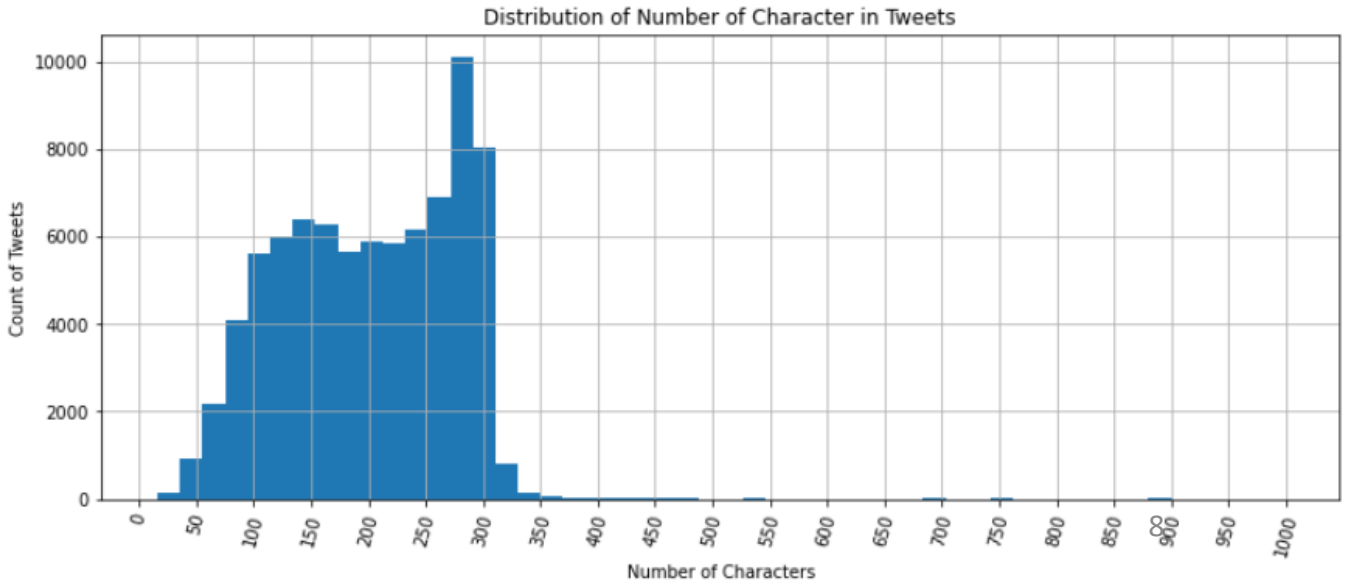


Preprocessing:

- Removing URLs, mentions, symbols and stopwords;
- Tokenization and Lemmatization;
- Changed with evolving models.

EDA:

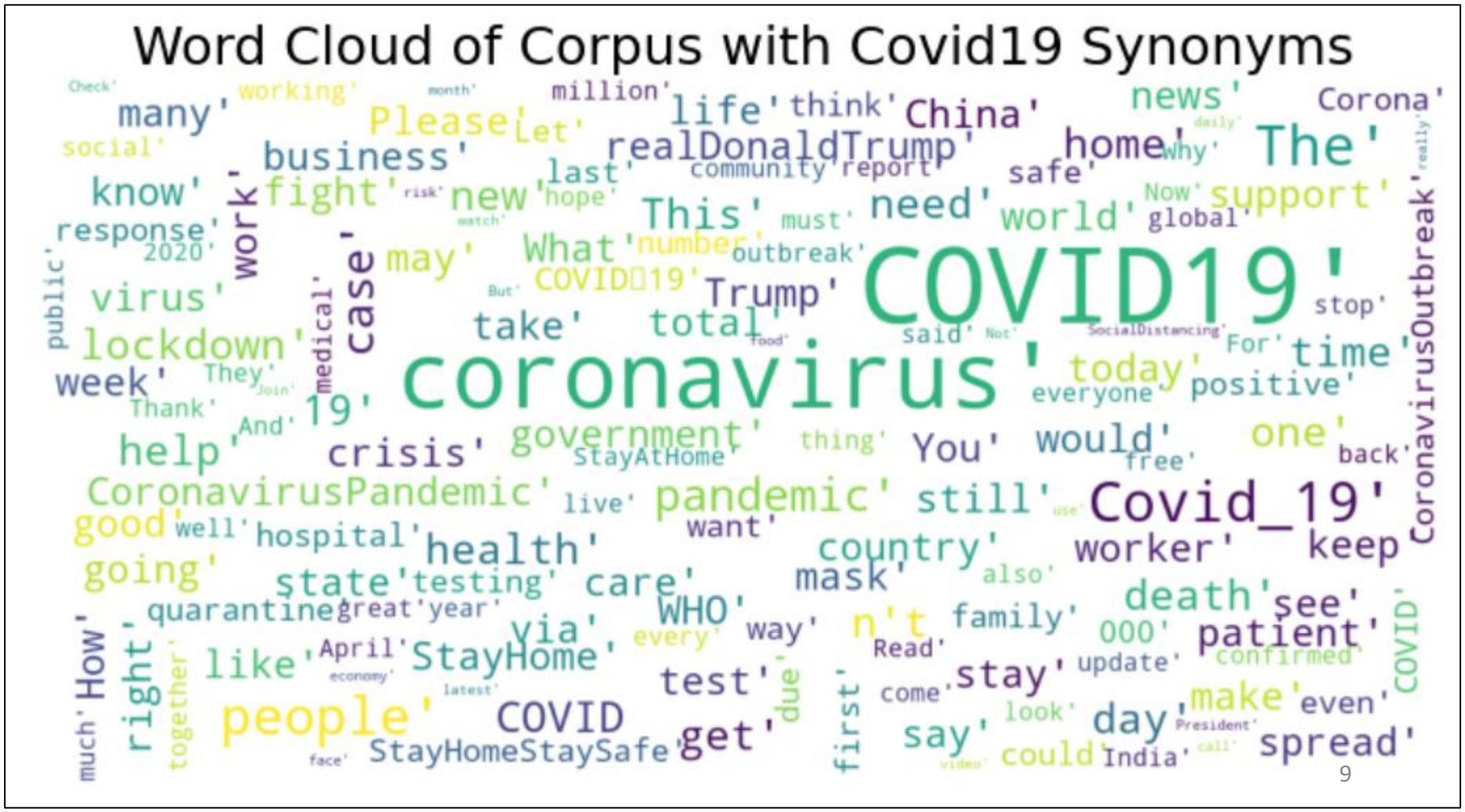
- Identifying data trends to improve analysis methodology;
- Changed with evolving models.



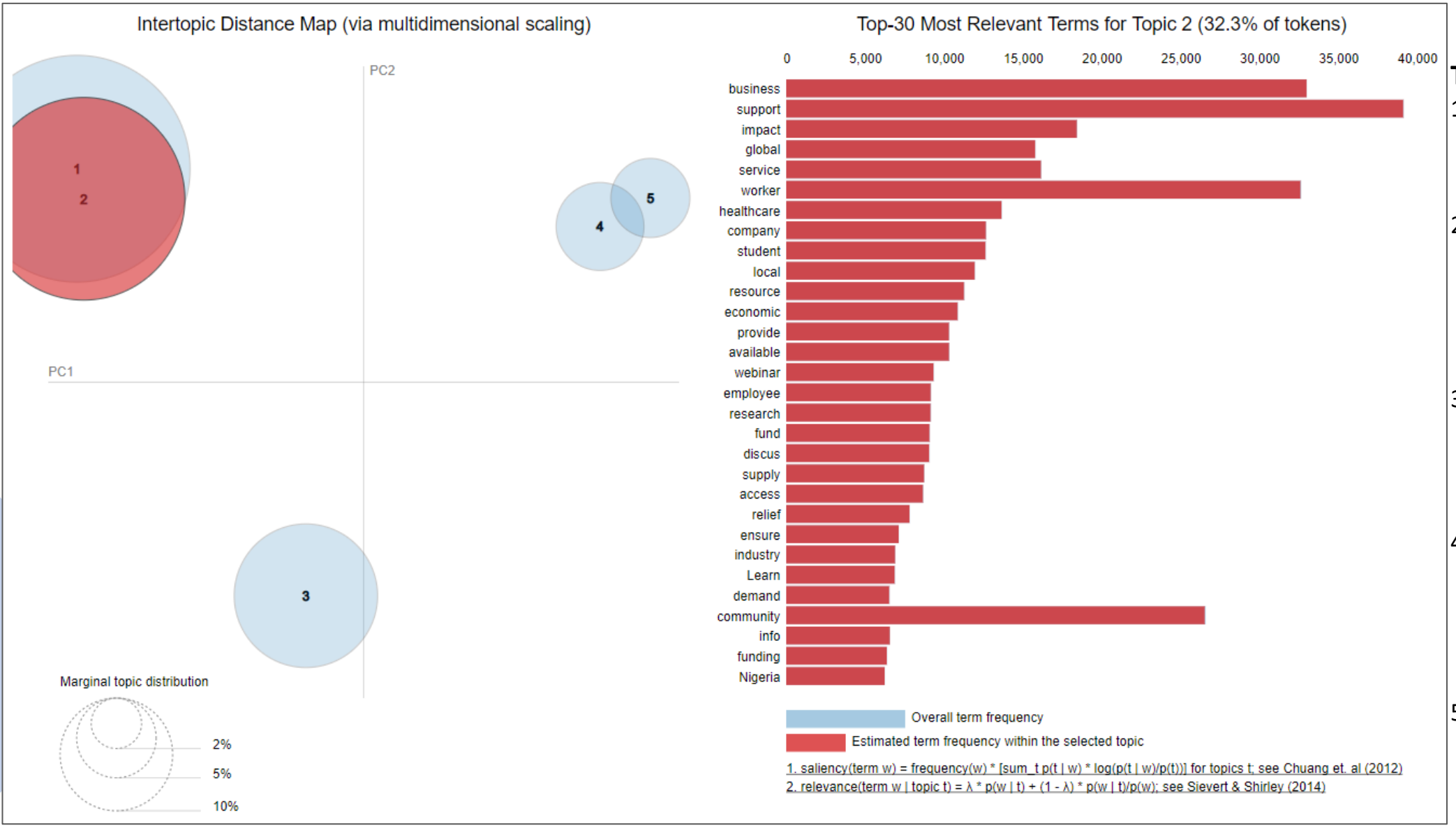
Corpus : Model 1

Basic Preprocessing

Base Corpus: 2.2M
Preprocessed Corpus: 1.5M



LDA Model 1 - 5 Topics & 35% Coherency



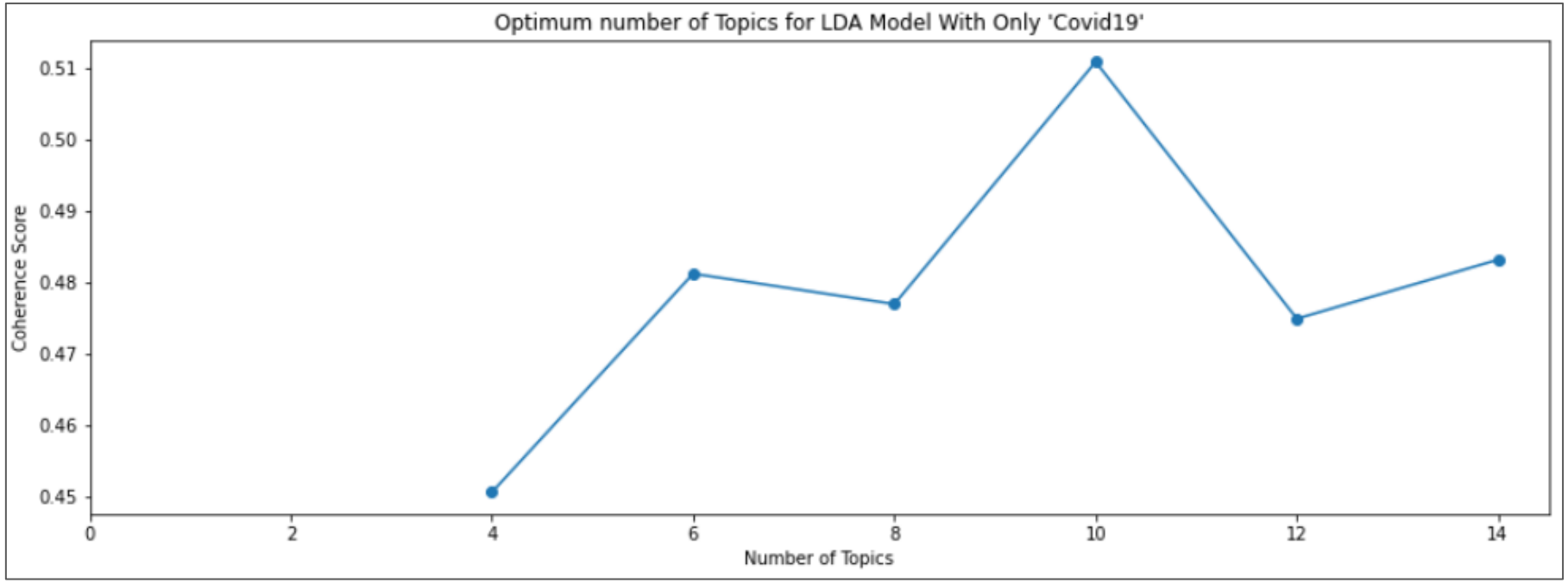
Topics:

1. American Politics:
   - 'Trump' and 'realDonaldTrump'
2. Business/Economic Impact:
   - 'business', 'support', 'impact', 'global' and 'economic'
3. International Outlook:
   - 'case', 'India', 'total', 'confirmed', 'Spain' and 'France'
4. Social Media with American Politics:
   - 'Youtube', ', 'MAGA', 'IngrahamAngle, 'Joe', 'PressBriefing'
5. Random:
   - 'support', 'last', 'you'

LDA Model Optimization

2 Intermediate Models (2 & 3) Created with Successive Improvements

- Key Preprocessing Changes (Based on Learning):
  - Covid19 Synonym Unification;
  - All tokens turned lowercase;
  - Balance between number of Topics and Coherency Score;

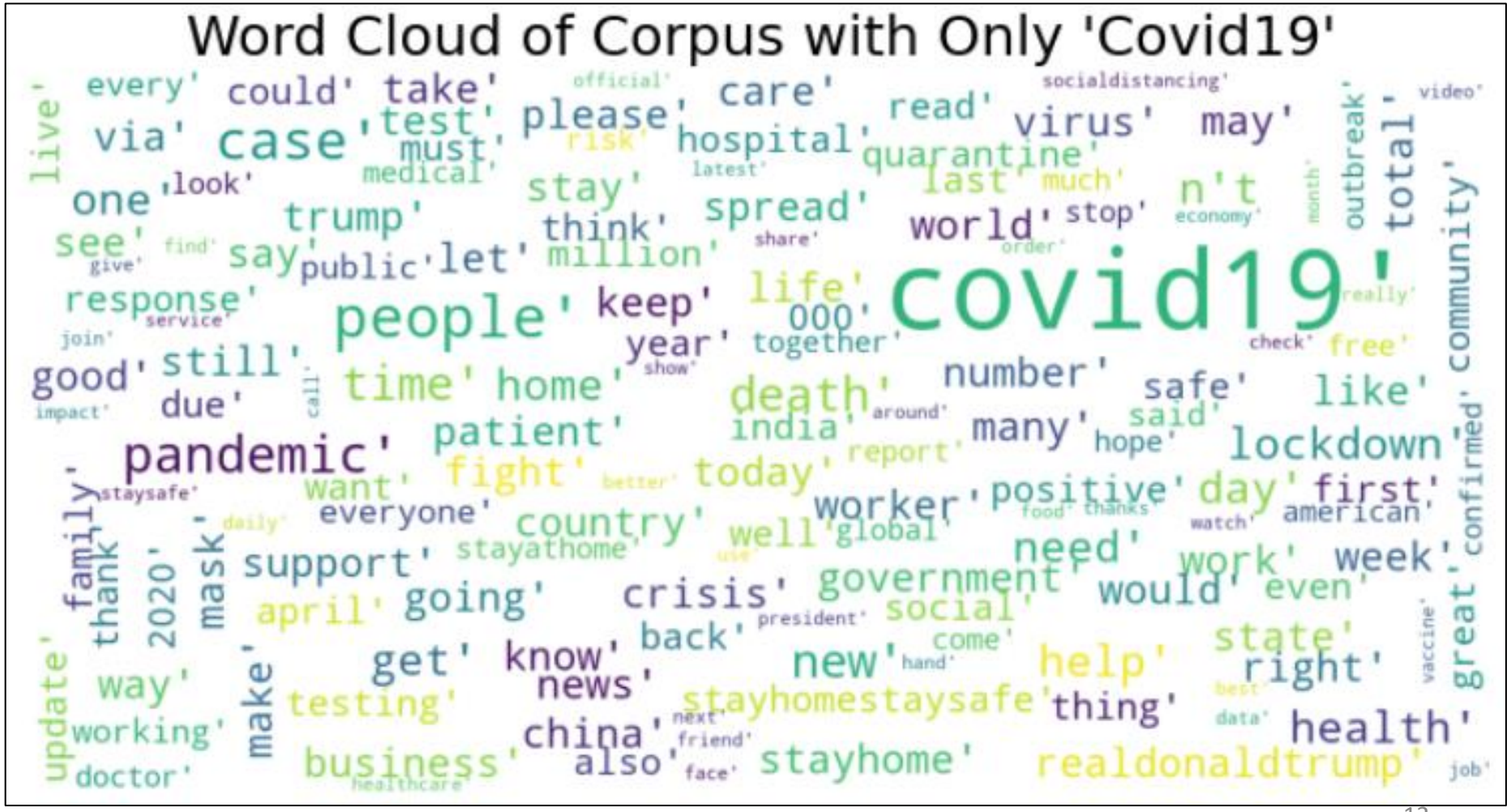


Final Corpus (Model 4)

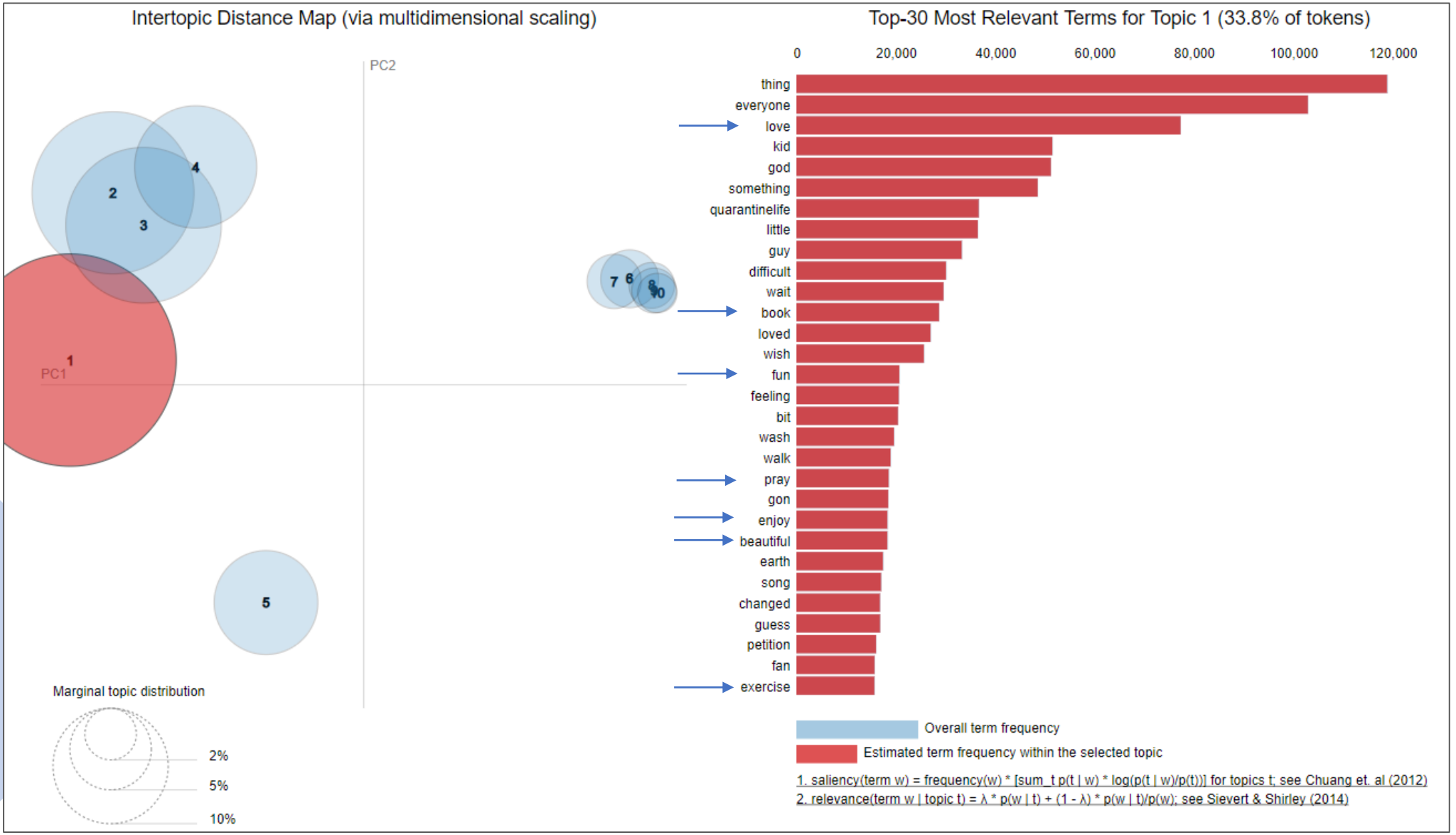
Additional Preprocessing

Base Corpus: 2.2M
Preprocessed Corpus: 1.3M

- All lowercase
- Fewer stopwords
- Emphasis of ‘covid19 ‘



Results: LDA Model 4 - 10 Topics & 51% Coherency



Topics:

1. Quarantine Silver Lining:
   - 'loved', 'beautiful', 'feeling', 'enjoy' and 'fun';
2. Business and Financial Impacts:
   - 'business', 'impact', 'industry' and 'financial';
3. Covid19 Tracking;
4. Covid19 Medical Advances;
5. American Republican Politics;
6. Indian Politics w/ Randomness;
7. Democratic and Canadian Politics;
8. Random:
   - 'taiwan', 'housing', 'tired', 'togetherathome','imf' and 'novadairy';
9. Negative Lockdown Aspect;
10. American Conspiracy Theories:
    - 'agenda', 'qanon', 'immigration', 'wwg1wga' and 'russian';

Conclusions & Potential Further Work

Conclusions

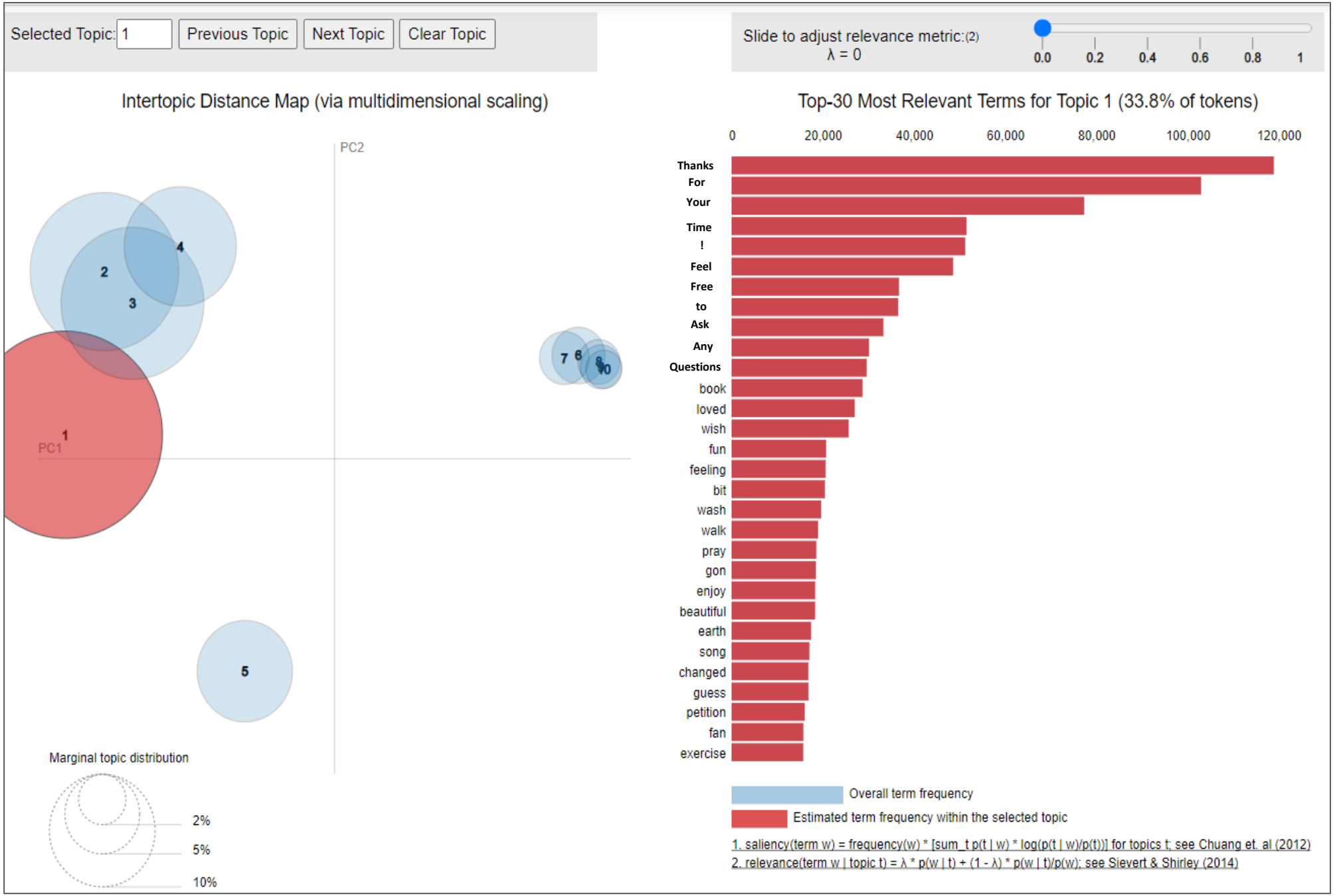
- Goal:
  - To automatically identify the underlying topics being discussed with context to Covid19;
- Generated 4 Latent Dirichlet Allocation Models
  - Preprocessing Changes
  - Data Comprehension
  - Consistent improvements
- Outcome:
  - 9 Coherent Topics and 1 Random Topic

Model Type (name)	Corpus Characteristics	Number of Topics	Coherence Score (%)
LDA Model 1	corpus_LDA_with_Covid19_synonyms	5	~35.0
LDA Model 2	corpus_LDA_with_Covid19_synonyms	20	~46.0
LDA Model 3	corpus_LDA_with_Only_Covid19	5	~47.5
LDA Model 4	corpus_LDA_with_Only_Covid19	10	~51.5

Further Work

- Use different vectorizers:
  - Tf-Idf;
- Bigger proportion of dataset leveraging cloud computing;
- Tweets from other months of Covid19;
- Entity Recognition;

Questions


Selected Topic: 1
Previous Topic
Next Topic
Clear Topic
Slide to adjust relevance metric:(2)
λ = 0
0.0
0.2
0.4
0.6
0.8
1
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
1
2
3
4
5
6
7
8
9
10
Marginal topic distribution
2%
5%
10%
Top-30 Most Relevant Terms for Topic 1 (33.8% of tokens)
0
20,000
40,000
60,000
80,000
100,000
120,000
Thanks
For
Your
Time
!
Feel
Free
to
Ask
Any
Questions
book
loved
wish
fun
feeling
bit
wash
walk
pray
gon
enjoy
beautiful
earth
song
changed
guess
petition
fan
exercise
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Backups

Issues with Stemming

Stemming and Lemmatization

```
words = ["connects", "connected", "strange", "is", "am"]

stemmed = ["connect", "connect", "strang", "is", "am"]

lemmatized = ["connect", "connect", "strange", "be", "be"]
```

Saliency and Relevance

Saliency is not only a measure of how frequently the term is used but also how important it is in the corpus for detecting a topic.

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Increased Characters

Out[10]:
```
179      889
1475     872
2448     482
9465     650
11917    464
         ...
79575    749
79750    921
80767    485
80851    858
81178    917
Name: text, Length: 76, dtype: int64
```

We will check one of the Tweets below. We have selected index 79750 as it shows a character count of 921.

In [11]:
```
analysis_df.iloc[79750]
```

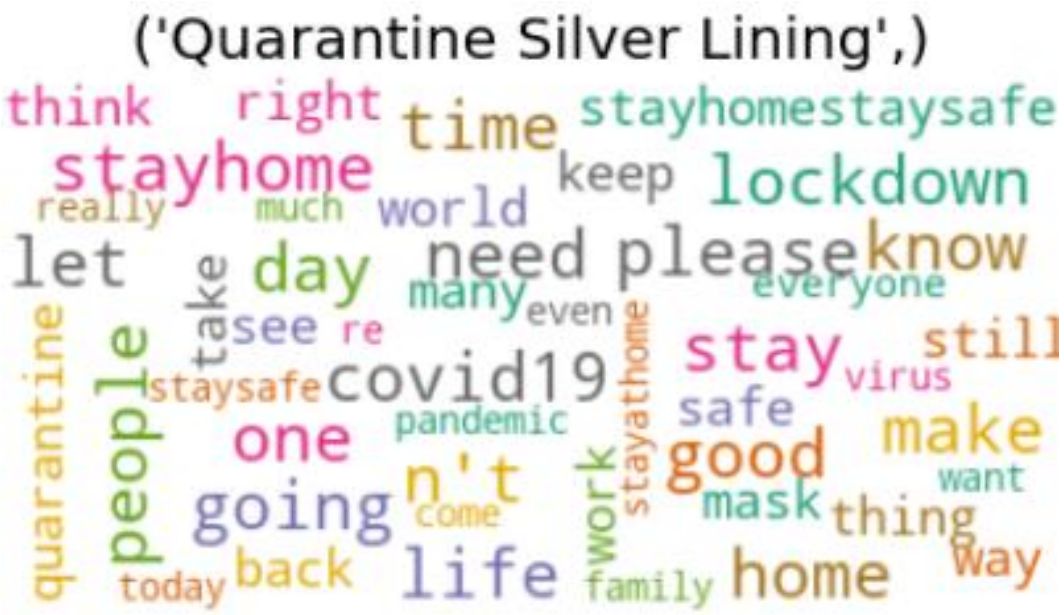
Out[11]:
```
Unnamed: 0                                                    79750
created_at                                  2020-04-30 07:15:17+00:00
screen_name                                               PhillyComptonMW
text               @laurasessions10 @SusanBordo @Sammysgranny @su...
is_retweet                                                    False
favourites_count                                              31740
retweet_count                                                     0
followers_count                                                1966
friends_count                                                  1034
verified                                                      False
Name: 79750, dtype: object
```

It's apparaent that the Tweet contains a high number of mentions, however we are unable to see the full text of the Tweet. Let's isolate that and check the text below.

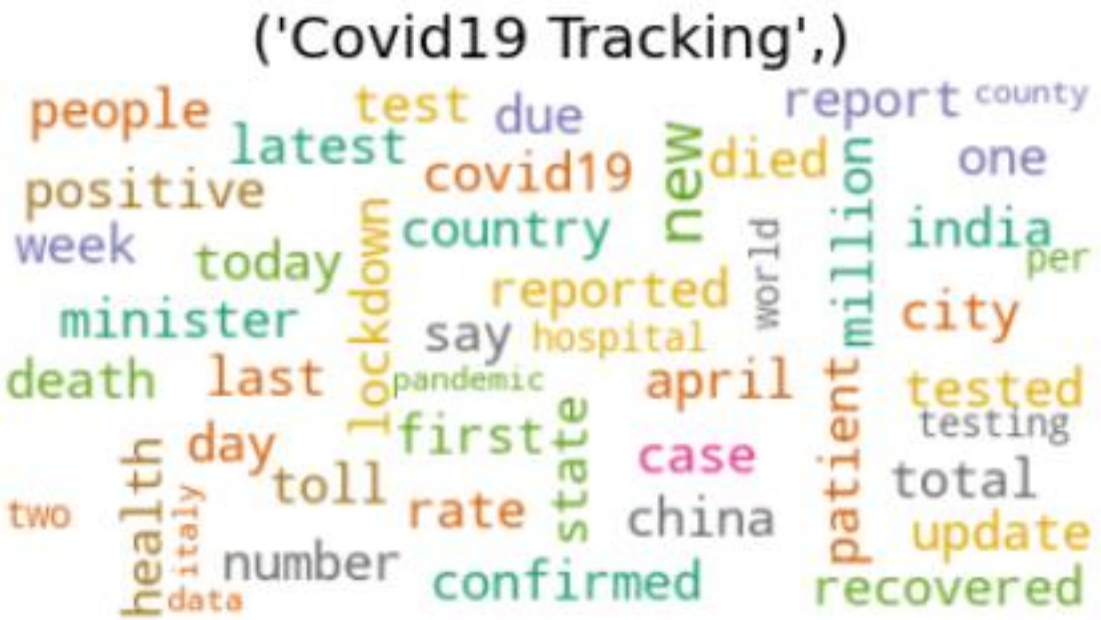
In [12]:
```
analysis_df['text'].iloc[79750]
```

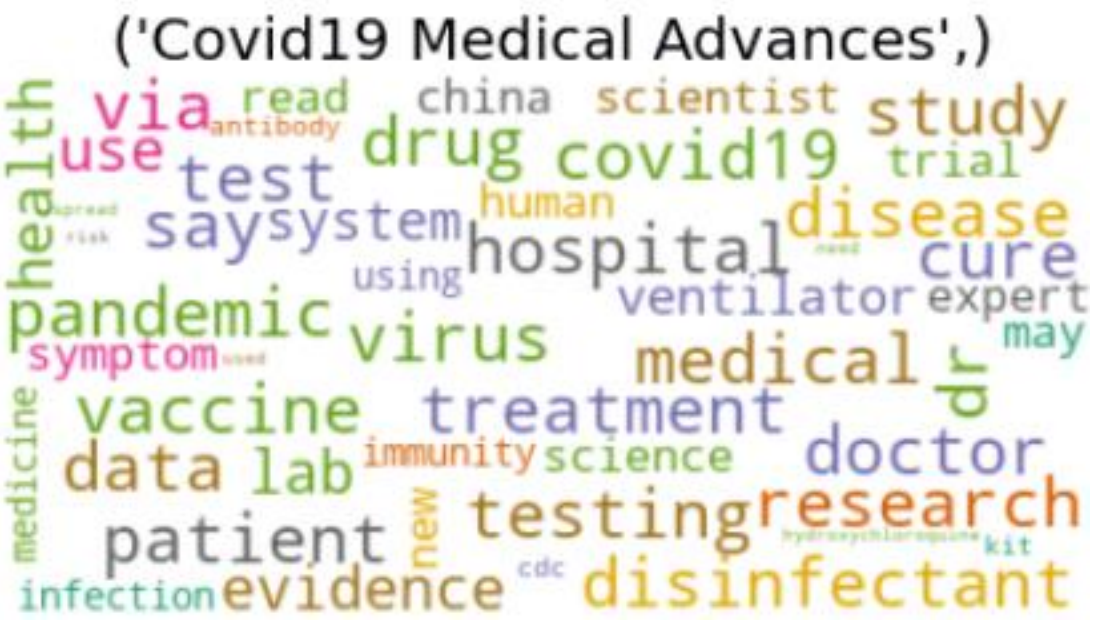
Out[12]:
```
'@laurasessions10 @SusanBordo @Sammysgranny @suewashko @bannerite @grammy4lphhl @kjoerwin @SophieInCT @WPalmerCurl @messerjs @P
uestoLoco @MooPersists @Retinalia @doctordill3 @letat_lechat @NWMouzer @KayTweetTweet @last_person_on_ @morgfair @veedubyoo @m3
3gs @workingtrucker @goprapebuster @PodcastObsessed @BrindaStar @ginadem @Bellarealness @AHamiltonSpirit @co_rapunzel4 @NaphiSo
c @EricWolfson @TheBaxterBean @Pandeism @Bvweir @TheWomensWatch @RonSupportsYou @admiralmpj @linksteroh @tedlieu @SenWarren @ju
dapeters @ColtSTaylor @kalpenn @JohnWesleyShipp @jonfavs @ProudResister @JuliaLikesFrogs @TheDailyShow @crzyfkinworld @Starz_Wa
yne Trump &amp;amp; Republicans: What virus? https://t.co/W1QmC9Lp5F\n#Trump #coronavirus #Georgia \n#TrumpPandemic #GOPGenocide #G
OP #TrumpIsTheWORSTPresidentEVER #TrumpGenocide #Republicans #TrumpLiedPeopleDied #TrumpLiesAmericansDie #RepublicansAreKilling
Us #MoscowMitch #COVID19 #Wisconsin'
```

Final Topic WordClouds Topics 1 & 2



Final Topic WordClouds Topics 3 & 4





Final Topic WordClouds Topics 5 & 6

Final Topic WordClouds Topics 7 & 8

('USA Democratic and Canadian Politics',)

('Random',)

Final Topic WordClouds Topics 9 & 10