Data Science

Capstone Report - Spring 2020

Dynamic Topic Modeling: COVID-19

Alexander Bogdanowicz

supervised by
Professor ChengHe Guan

Abstract

COVID-19 has led to unprecedented changes in global politics, economics, and social interactions. Amongst its other impacts, this work aims to study, through the use of advanced Natural Language Processing and Dynamic Topic Modeling, social sentiment and reactions to COVID-19 through topics expressed on social media. We deploy an end-to-end data processing pipeline and implement a Sequential Latent Dirichlet Allocation model to track the daily growth and changes in topic composition of 8 million tweets over the period March 31st to April 13th. We successfully identify 12 diverse topics which exhibit strong macro-trends over time while experiencing micro topic variations, covering domains such as healthcare, politics, community, and economics.

1. Introduction

The last two decades have seen societies across the globe continue to evolve their means of socializing and self-expression, and consequently have led to simultaneous advancements in the primary statistical domain related to communication, Natural Language Processing (NLP). With the advent of a global pandemic impacting the lives of billions across the globe, the capability to assess how individuals, groups, and societies respond to, and cope with, the extraordinary consequences of COVID-19 is more pressing than ever.

1.1. Problem Motivation: COVID-19

In as early as late November of 2019, a SARS-like virus began spreading between neighborhoods in China's Hubei province with those infected reportedly experiencing pneumonia-like symptoms with an unknown cause. Since then, the world has experienced an unprecedented shakeup as the virus began to spread to parts of Europe and East-Asia, before rapidly infecting over 187 countries and territories with casualties numbering in the hundred-of-thousands. The viral outbreak has overwhelmed an increasingly global healthcare system, leading to shortages of personal protective equipment and crucial life-saving medical systems, while throttling local and global economies alike.

As a result of its increasingly digital nature, communication via micro-blogging social networks have imbued the study of human behavior, including individual sentiment, group topics, and even identity-politics with a big-data driven science. Twitter, with over 315 million active users generating over 500 million tweets per day has historically served as a reliable source of social expression, largely as tweets tend to contain the following useful properties:

- Textual Data (Topics and Sentiment)
- Temporal Data (Time-Series Component)
- Spatial Data (Geo-Tagging and Profiles)[1]

As is the case with any systematic shock of these proportions, individual and group reactions have been varied, with some resorting to denial, others to scapegoating, and still others to the spread and consumption of misinformation. In an effort to gauge public sentiment and the global pandemic's impact on social thoughts and behavior, we attempt to answer the following questions:

1. What is the general sentiment, over time, and grouped by location, in the United States specific to the pandemic?
2. What kinds of topics are individuals and groups vocalizing in relation to the pandemic?
3. Are there any noticeable topic trends and if so how do these topics change over time and in response to major events or news?

We will examine the contemporary literature relating to sentiment analysis and topic modeling as it relates to twitter data and epidemiology, appreciating recent developments in NLP, delving into the solution, understanding data acquisition, structure, and preprocessing before finally explaining the model and results.

2. Related Work

2.1. Sentiment & Topic Modeling: History

In the context of extracting topics from primarily text-based data, Topic Modeling has allowed for the generation of categorical relationships among a corpus of texts. The origins of contemporary

[1] As of June 2019, Twitter has made geo-tagging an opt-in feature (only 1-2% of tweets are now geo-tagged)

topic modeling techniques can be traced to the late 1980's (Deerwester et al., 1990)[1], with the emergence of Latent Semantic Analysis (LSI). LSI itself however is really only an application of Singular Value Decomposition, attempting to identify a subspace of a Document Term Frequency Space in order to capture the majority of variance in the corpus. Latent Dirichlet Allocation, discovered by a team of University of California, Berkley Researchers in 2003 [2], unlike it's discriminative counterparts, is a generative model. In LDA, documents are represented by random mixtures of words over latent (i.e. emerging during the modeling process) topics. Therefore, LDA is able to identify the probability of a given document being in a given topic through a "bag-of-words" interpretation of its contents. We delay further technical discussion of LDA to Section 3. Solutions.

Since it's emergence in 2003, LDA has played a benchmark role as a model for Topic Modeling, and has since seen various domain-specific improvements and adjustments. In 2011, Du et al. (2011)[3] introduced a temporal component to topic modeling, referred to as Sequential Latent Dirichlet Allocation, which focuses on modeling how topics in a corpus of documents evolve over time, by relating the antecedent and subsequent segments. In 2012, Zhang and Sun. (2012)[4] added a parallel probabilistic generative model to LDA, which included correlations between users to generate topics and saw modest improvements to accuracy. At the same time, Huang et al. (2012)[5] attempted to simplify the feature space by first implementing a single-pass clustering algorithm, before utilizing traditional LDA on the new vector space. The year after, Zhao et al. (2013)[6] furthered the micro-blog topic domain by utilizing a greater feature space relating to micro-blog posts, focusing however on a simplified LSA model. Table 1. depicts a summary table containing many of the contemporary variations of Latent Dirichlet Allocation.

LDA Advancements and Variants			
Authored	Year	Variant	Description
Blei et al.	2003	LDA	Original Generative Model
Blei and Lafferty	2006	SeqLDA	First Dynamic Topic Model
Du et al.	2011	SeqLDA	Time-Series Topic Model
Zhang and Sun.	2012	MB-LDA	Feature Space Expanded to User Network
Huang et al.	2012	-	Clustering Feature Space prior to LDA
Zhao et al.	2013	MB-LSA	Expanded Micro-Blog Feature-Set + LSA
Wang et al.	2015	SH-LDA	Updated Temporal & Hashtag-Graph-Based Topic Model
Xu et al.	2016	TUS-LDA	Joint Temporal and Emotional Probability Space LDA
Wang et al.	2018	-	Dynamic Spatial & Temporal LDA
Du et al.	2019	MF-LDA	Analyzed the life-cycle of "hot-topics" with Dynamic LDA
Yao and Wang	2019	-	3-Step Geo-Topic Generation and Tracking LDA

Table 1: Recent Literature on Topic Modeling

2.2. Domain Specific Topic Modeling: Epidemics Literature & Twitter

Studying the spread of disease has been the primary focus of epidemiologists for the last century, but only recently has the advent of social media enabled the study of how individuals and groups think about and react to viral outbreaks. However intuitive the combination of Natural Language Processing and epidemiology may seem however, textual data is naturally precarious to work with and filled with user-bias. In 2008 for example, Google famously claimed to have effectively detected areas of influenza epidemics two weeks earlier than the CDC by analyzing search frequencies and terms [7], only to have their model, Google Flu Trends (GFT), fail consistently by over 140% during the peak of that year's flu season. Therefore, it is important to understand the limitations of natural language processing when pursing potential predictive models.

Influenza is also a good example of a yearly re-occurring viral outbreak that has allowed for

research relating to tracking the spread of and sentiment around the virus. One of the earliest topic modeling techniques centered around virology and using Twitter data was Paul and Dredze's (2011)[8] Ailment Topic Aspect Model (ATAM), which added isolated ailments such as influenza, infections, and even obesity, filtered from the Twitter corpus and providing for a secondary latent ailment variable in an LDA-style topic model. The model was improved upon with the development of ATAM+ with an added predictive component, built on a more medically-specific corpus of hand-picked articles relating to specific diseases [9].

Much of the literature concerning virology and topic modeling is concerned with predicting outbreaks given the historically available geo-location data from micro-blogging applications. It is vital to mention here that works relating to predicting outbreaks on the basis of models trained specifically on Tweets from Twitter are no longer feasible as of June 2019. This comes as a result of Twitter, Inc.'s decision to make geo-tagging an opt-in user feature. The result is that only 1-2% of the total population of tweets is geo-tagged, and this sub-population tends to be very biased towards a younger demographic and commercial uses of Twitter. We therefore focus our analysis on the entire-population of Twitter users and continue with a review of works analyzing public discussions, reactions, and sentiment towards outbreaks. In Section 3. we discuss possible solutions to this data-quality issue and potential next-steps.

Despite the lack of geo-tagging capabilities in modern applications of Twitter data, there remains literature related to topic extractions and sentiment analysis outside of a geo-specific context. As early as 2011, Alessio Signorini et al. (2011)[10], utilized Twitter data to track public concern of the spread of the 2009 Influenza A H1N1 Pandemic, identifying strong influxes of topics revolved around hand-washing and mask safety. Roy et al. (2019)[11] focused instead on understanding the overall proximate blame tendency of online users on Twitter, with relation to the Ebola Outbreak. Instead of predicting the spread, Roy et al. (2019)[11] utilized a sequential LDA model to follow the evolution of localized blame in a retroactive study of the outbreak. Ahmed et al. (2019)[12] implemented a similar technique in a study of topics surrounding the H1N1 pandemic, identifying a subset of misguided Twitter users that believed pork could host and/or transmit the virus. The 2015 Zika outbreak led to an examination of millions of tweets geolocated in North and South America by Pruss et a. (2019) [13], who discovered increases in public attention to Vaccinations, Viral Testing, Symptomatic Topics, and increases in polarizing political topics.

2.3. Contributions of this work

It is important to highlight the various advancements to Topic Modeling and Natural Language Processing that have been made over the last two decades and to understand the findings that these advancements helped generate in an epidemiological context. This work hopes to attain a two-fold contribution to the contemporary analysis of the most impactful virological outbreak since the Spanish Flue, COVID-19, and to the topic modeling domain at large by focusing on the following factors:

1. A Big Data Approach to Sequential Latent Dirichlet Allocation (100 million+ Tweets)
2. A reproducible work, with a focus on an end-to-end custom reusable data pipeline
3. An understanding of the evolution of topics surrounding the COVID-19 pandemic
4. A novel exploration of methods to combat recent geo-location limitations

Most works concerning either theoretical topic modeling or its domain-specific application are limited by resources, the topic scope and size, and the accessibility of historical sentiment data (e.g. tweets). As a result of the size, scope, and length of the COVID-19 pandemic, the breadth of topics concerning the outbreak and consequently the number of tweets generated on the topic is exponentially greater to that of any other outbreak in recent history. As a result, we have

collected millions of tweets per day specifically relating to commentary on the outbreak, a scale that is unattainable in any other of the previous works reviewed - no work concerning viral outbreaks has reviewed a twitter dataset of over 10 million tweets.

Additionally, most of the works reviewed do not offer a publicly available operational code-repository with an end-to-end, from streaming to modeling, custom developed pipeline (akbog/urban_data).

Finally, as a result of Twitter's recent policy changes regarding the geo-location of tweets, we recommend a novel approach to backing out geo-coordinates from public twitter user profiles, using a locally built instance of the OpenStreetMaps database for geo-encoding.

3. Solution

3.1. Data Acquisition

In recent years, accessibility to large datasets sourced from social-media platforms have become more-and-more scarce, given increases in privacy concerns for user-data and an increase in enterprise (i.e. paid access) interest in social media data. Twitter has famously remained a relatively open platform for data analytics, making it a popular source amongst academics. Yet there are still certain hurdles faced when acquiring twitter data samples, namely streaming limitations and more recently geo-location restrictions. See Table 2 for a description of Twitter's Developer API Services and Limitations.

Twitter API Access		
Service	Tweets/Month	Description
30-Days Sandbox	25k	Tweets only available from within the last 30-days
Full Archive	5k	Tweets from the full twitter archive (since 2008)
Standard Stream	Rate-Limited	Stream Live Tweets from the last 14-days (Excessive requests can generate rate-limits)

Table 2: Twitter API Services & Limitations

As a result of the data-size limitations posed on the 30-Days Sandbox and Full Archive datasets, as well as the on-going problem we are studying (as of May 2020, the virus is still being hotly discussed on social media), we decided to stream as many tweets as possible (24/7 streaming through Twitter's API Endpoint) starting from March 31st 2020.

Twitter's Streaming API allows for a number of request parameters that can filter for language, location, and key-words. Specified conditions act as "OR" statements, not "AND", meaning specifying a location and key-words produces geo-tagged tweets from a specific location ***or*** matching a given key-word. Therefore we are only concerned specifically with tweets pertaining to COVID-19 and as have streamed an average of *3.314* **million** tweets per day between March 31st and April 13th. We use the following key-words to filter COVID-19 related tweets during the streaming process:

```
["coronavirus", "corona virus", "covid19", "covid", "covid-19", "wuhan",
"sars-cov-2", "pandemic", "epidemic", "outbreak", "virus", "infect"]
```

Additionally, after a connection is broken or stalled, Twitter's API may return duplicate tweets in each request. We therefore build dictionaries to test whether a given tweet_id (which is a unique identifier) has already been streamed prior to outputting into files. We output files into compressed gzipped bundles of approximately 5000 tweets.

3.2. Data Structure, Reading, & Storage

An early issue encountered while importing tweets was that NLTK's default twitter corpus reader was unable to stream tweets from gzipped files. We therefore extended the reader with a custom `GzipStreamBackedCorpusReader` to load tweets dynamically from compressed files. By default, tweets are outputted into the familiar, dictionary-like, JSON format. The main component to a tweet object are the tweet itself and the user object, with certain tweets containing extra retweet, quote, or reply objects depending on the type of tweet.

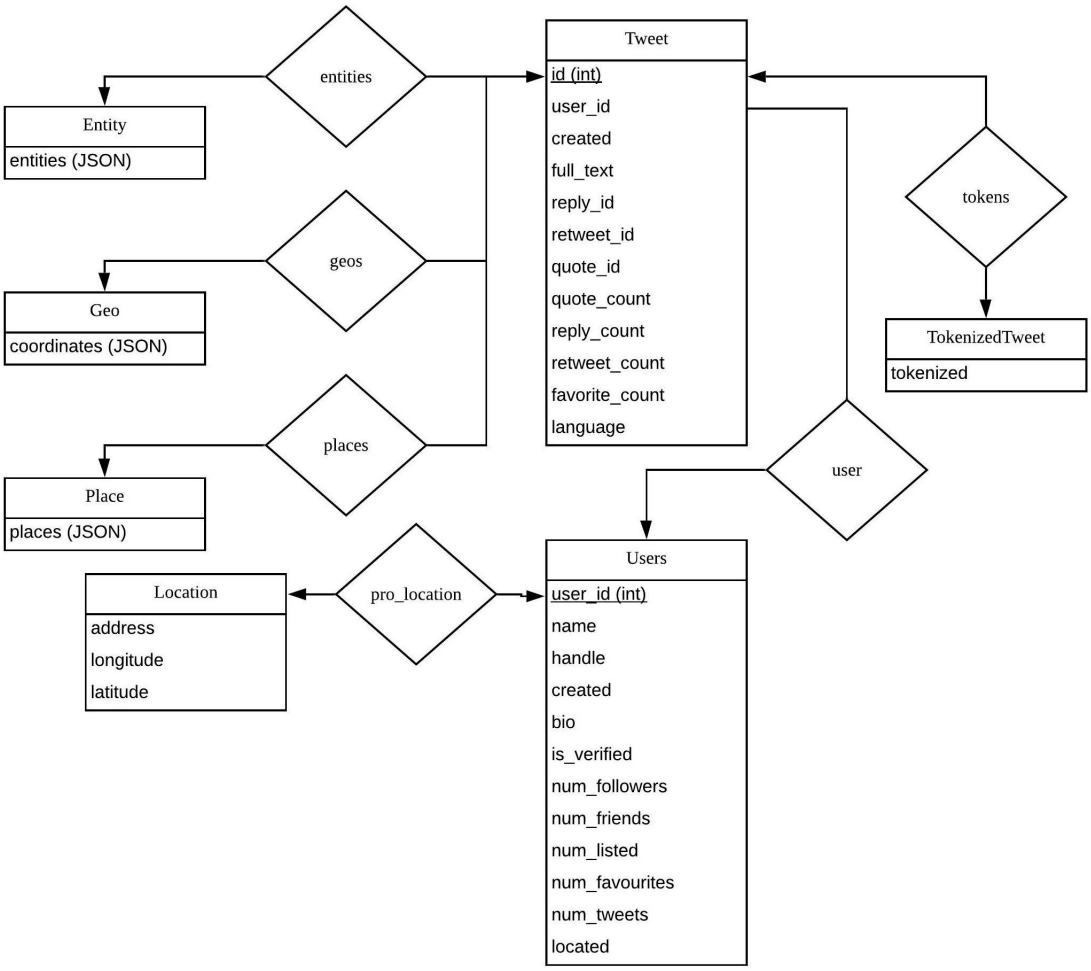

Figure 1: Entity-Relationship Diagram of the PostgreSQL Database

As Figure 1 suggests, the data is stored in a PostgreSQL Database on an NYU Shanghai based Virtual Machine, based on the custom schema defined in Figure 1. Tweets are ingested by instantiating a custom built `TwitterDatabase` object, which builds the tables via the Python Object-Relational-Mapper (ORM) SQLAlchemy, reading in the Gzipped JSON file corpus, parsing each tweet, and finally utilizing the `C++` Library LangDetect to identify the language of each tweet before database ingestion.

3.3. Dataset & Preprocessing

Natural Language Processing requires transforming a semi-structured dataset of random words into a machine-learning readable format. In order to accomplish this, we first take some initial considerations as to the current structure of our dataset, noting that, as it stands, we have settled on a collection of 46,000,420 tweets over a 14 day period starting March 31st. Of these, over 50%

are retweets, meaning they represent duplicate tweets that are already existing in our database. We defer the merits of a metric for weighting retweets to future works and proceed to remove retweets from the dataset at hand.

Additionally, as proposed in Yao and Wang (2019)[14], we consider the possibility that certain users may exhibit features outside the realm of normal twitter use (i.e. excessive tweets, retweets, followers etc.). We therefore deploy a similar method to remove users - and therefore tweets - whose Z-score is outside 3 standard deviations from the norm, calculated as follows:

$$Z_{(tweets,followers,friends)} = \frac{log_{10}V_{(tweets,followers,friends)} - \mu_{(tweets,followers,friends)}}{\sigma_{(tweets,followers,friends)}} \quad (1)$$

We take the log of each value (i.e. tweets, followers, friends) to account for their extremely right skewed distribution. Figure 2 depicts the resulting tweet distribution after removing retweets and users with abnormal behaviors (e.g. bots).

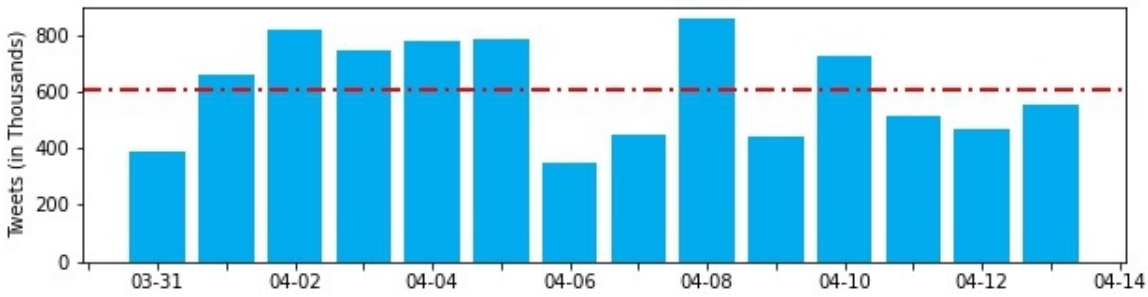

Figure 2: Distribution of Preprocessed Tweets

The variation in the tweets is not a result of our preprocessing stages but rather a result of the behavior of Twitter's streaming API, which will rate-limit when requests are especially high.

As a result of computing limitations on our database machine, which is optimized for querying but not parallel processing, we setup a copy of our dataset on NYU Shanghai's HPC, utilizing a column grouped distributed file structure, Apache Parquet, which optimizes for column operations and distributed tasks. Figure 3 depicts the next steps, which feature a collection of Natural Language Processing tasks on the data. These include standard operations, such as normalizing sequences of repeated characters and lowercasing. We extend an MIT developed reddit & twitter text tokenizer, `RedditScore`, to perform stop-word removal using `SpaCy`'s and `NLTK`'s stopword dictionaries, punctuation and numerical operations, twitter-handle manipulation, and finally lemmatization with `NLTK`'s `WordNetLemmatizatizer`. This operation is necessary to remove non-semantic words such as "the", "of", or "and". A unique aspect of this process is that we are able to split hashtags, which are always continuous strings, into their respective words (e.g. #DoctorsAgainstCOVID19 to Doctors Against COVID19), thereby preserving this context.

It is also important to remove words that are perceived to be common across all topics. Given that a pre-requisite of membership in the dataset is a key-word match with our key-word dictionary and given that COVID-19 goes by many different names (e.g. Coronavirus, Sars-Cov-2 etc.), we first substitute all references to the virus with the string "coronavirus" prior to adding the term to our stop-word dictionary for removal. We identify other terms to remove during the model training process.

As a final note, it is often the case that models contain units of bigrams or trigrams which are sets of words that co-occur together in two's or three's respectively. While it is common-practice to include these in the preprocessing and tokenization stages, we opt to utilize `Gensim's` Phraser modeler, which is trained separately, and can help to limit the size of the dictionary of words for modeling by only adding terms that co-occur above a certain frequency. We utilize this process

during the modeling phase because it requires hyper-parameter tuning to identify the correct threshold.

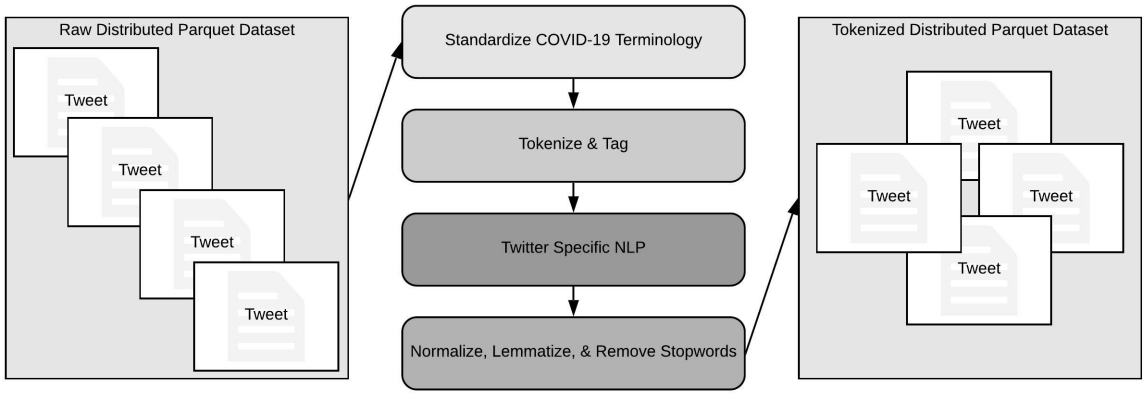

Figure 3: Dataset Preprocessing Stages

We run our tokenization process in parallel using Dask, an open source python library for parallel computing, which effectively takes advantage of high performance computing. Finally, we store our preprocessed tokens in a second distributed parquet dataset on NYU Shanghai's HPC, ready for modeling.

3.4. Geo-Tagging: OpenStreetMaps

Initially, we set out to Dynamically Model Topics over space, inspired by Yao and Wang (2019)[14], but were handicapped by Twitter's new opt-in geo-tagging policy which severely limited the size of our potential dataset. It is important to note that this is an issue we expect many will face when working with Twitter Data in the future. In light of this, we developed a novel approach to combat the lack of geo-specific information by instead utilizing User's self-specified profile locations and reverse geo-encoding User's locations from their profiles. While certainly less accurate, the text-form version of User's profile location is unusable outside of a focused context (i.e. user's in New York). Whereas certain database APIs exist that perform this operation, almost all of them have strict rate and query limits, and none of them can realistically geo-encode a dataset of millions of tweets. Many of these APIs however, such as the OpenStreetMaps (OSM) project, are publicly available and offer solutions for hosting PostgreSQL instances of their databases locally. Considerations relating to storage, computational power, and memory are all necessary.

In order to gauge the accuracy of this solution, we suggest the following accuracy function, which can be run utilizing a sample of already geo-encoded tweets, tested against the same sample of profile reverse geo-encoded tweets:

$$Accuracy = \frac{\sum_{i=1}^{n} hav(\varphi_{Pred} - \varphi_{Actual})}{n} \tag{2}$$

Here we utilize haversine distance to calculate the great-circle distance between the prediction and actual result. This function can be used to both test the average deviation of our estimate and as a measure to bias OSM Database results for areas that are over-represented in Twitter but underrepresented in OSM. To our best knowledge, such a technique has yet to be deployed to combat the lack of geo-encoded tweets.

While we have built a local instance of the OSM North America Database and are still pursuing this option, it is unfortunately outside the scope of current resources as a practical application will require a full planet import; an operation on the scale of weeks to months given current resources. We therefore defer this implementation to future works.

3.5. Modeling & Latent Dirichlet Allocation

In order to work with LDA models, we must first build out a data pipeline to vector encode our tokenized tweets into a term-frequency matrix with m rows (# of tweets in our corpus) and n columns (# of unique terms). For multiple reasons, including Gensim LDA integration and grid search, we settled on utilizing a Scikit-Learn pipeline for model testing, which allowed us to build models with different vector encoding (i.e. TF-IDF, One-hot etc.) and utilize grid-search for hyper-parameter optimization in parallel. One-hot vector encodings simply encode the presence of a word in a string as 1 if present and 0 if not, without keeping track of quantity. Whereas we tested TF-IDF and normal (absolute) vector encodings of the dataset, the decision was made to utilize One-hot vector encoding as a result of 2 factors; our dataset is made up of short (typically <140 character) tweets which do not typically repeat terms and as LDA is a word generating model, TF-IDF score representations are not readily interpretable.

Sequential LDA were first discovered by the original co-creaters of LDA, Blei and Lafferty in 2006[15]. Whereas many incremental changes have been made to LDA since then - as described in Section 2.1 - the dynamic component of the original 2006 SeqLDA work is sufficient for our purposes. In normal static LDA, documents are represented as random mixtures of latent topics, where each topic is characterized by a distribution of word probabilities.

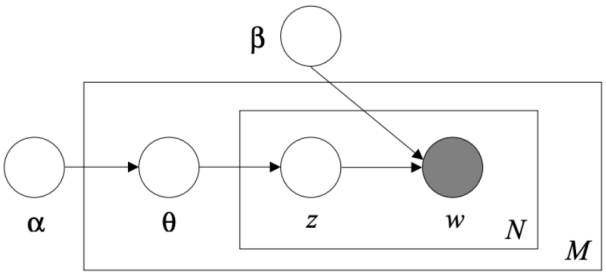

Figure 4: Original LDA Representation [2]

The plate diagram shown in Figure 4 shows the process for generating M documents (i.e. tweets), each with N words. β depicts a probability matrix representing the probability of a word being in a given topic, whereas α is a vector representing the probabilities of a given topic being present in a given tweet. Both are hyperparameters that effectively establish the Dirichlet distributions that govern the model. In essence, classic LDA is a method of attempting to understand what words make up which topics and how these topics make up a document through words.

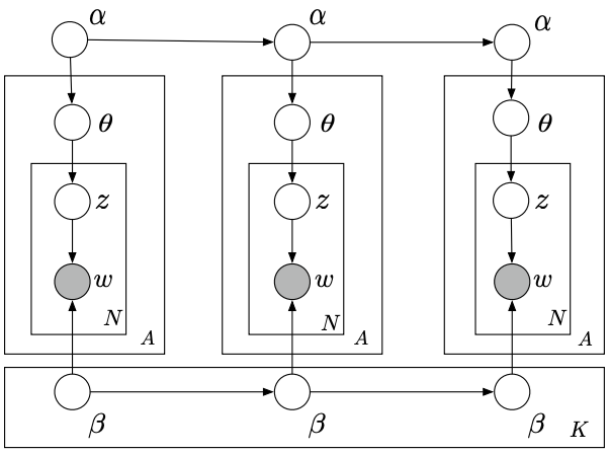

Figure 5: Original DTM Representation [15]

Sequential LDA provides static LDA with a dynamic component by utilizing a state space model, as depicted in Figure 5, which replaces the Dirichlet distributions with log-normal distributions with mean α, chaining the Gaussian distributions over K slices and effectively tying together a sequence of topic-models.

In order to implement the logic outlined in Blei and Lafferty (2006)[15], we first train a collection of normal LDA models on a subset of our data (March 31st - April 2nd) in order to establish the hyperparameters of our model. As `Gensim` already utilizes KL-Divergence to estimate α and β Dirichlet priors, we only test for the optimal number of topics.

Whereas `Gensim's` default scoring function is perplexity, we choose instead to use a measure of topic coherence which operates by maximizing the following function:

$$UMass_{(w_i,w_j)} = log\frac{D(w_i,w_j)}{D(w_i)} \tag{3}$$

UMass scores higher when words appear together more frequently than they do by themselves, operating under the assumption that topics that are "coherent" will feature words that appear together more often than not.

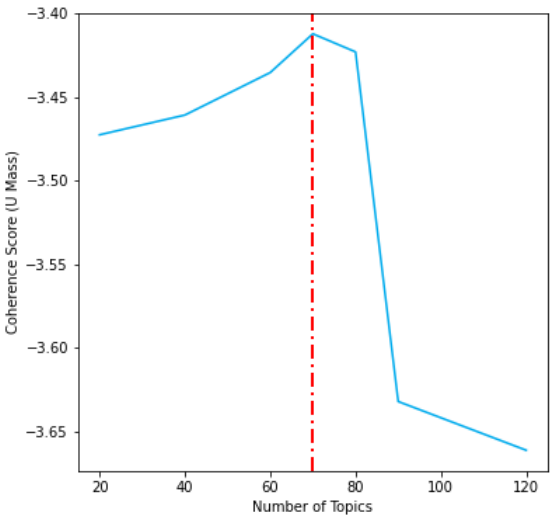

Figure 6: Static LDA Coherence Scores for Varied Numbers of Topics

The resultant models and their corresponding coherence scores can be seen in Figure 6. There was a clear benefit from increasing topic size until the number of topics reached 70, at which point there was a decided drop in coherence scores. As a result, we choose 70 Topics for our Sequential Model.

As `Gensim` features an existing implementation of the Sequential LDA algorithm presented in Blei and Lafferty (2006)[15], we initialize our model with the pre-calculated hyperparameters, and proceed to build the model. Prior to using Gensim's implementation, we attempted two other LDA implementations; LDA Mallet, a java based implementation with a python wrapper and guaranteed faster convergence and sklearn. However, on the provided sample, LDA Mallet refused to converge and Sklearn's default LDA implementation proved both time-consuming and produced lower Coherence Scores than Gensim.

After multiple rounds of testing, our final model took $\sim$ 34 hours to complete, passing over the dataset 5 times each training iteration (i.e. 5 passes / time slice), updating assumptions every 1000 tweets. A summary of our model configuration and cluster resources used can be found in Figure 3. We theorize that model training times could be improved through either compilations in cpython or utilizing an Apache Spark (i.e. optimized for Distributed Computing) LDA model.

4. Results and Discussion

4.1. Topic Distributions

As a result of the qualitative nature of working with textual data, evaluating the results of an LDA model are partly quantitative and partly qualitative in nature. This is similar to the logic presented when choosing a proper optimization function. At the end of the day, our model is designed to extract qualitative measures of the topics that individuals are discussing, and follow how these topics change over time.

Cluster Configuration	
Nodes	2
Cores/Node	16
Memory/Node	32GB
Partition	Parallel
Model Configuration	
Dataset Passes	5
Update Model	Every 1k tweets
Scoring	UMass
Train Time	$\sim$ 34 hrs

Table 3: Cluster & Model Configurations

First, we breakdown the most popular topics present in our dataset. Whereas we optimized our model to account for 70 topics, which was arrived at utilizing a grid-search strategy and optimizing for Umass, our topics do not feature equal representation in the dataset. In fact, as Figure 7 depicts, for any given day between April 3rd and April 13th, the top 12 topics over each day make up between 70 and 80% of the topics present in our dataset.

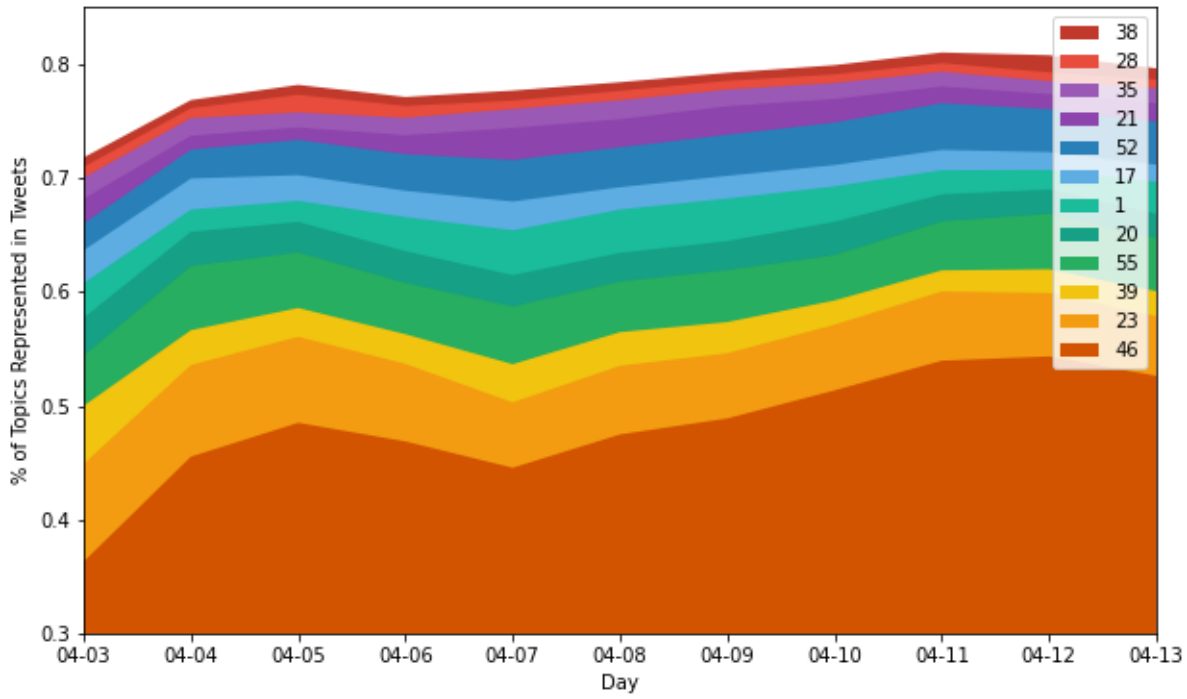

Figure 7: Time-Series Representation of the Changes in Topic Dominance Over Time

We also observe an increasing trend in the representation of the top 12 topics over the specified period. We note a few potential reasons for this:

1. As the Sequential Model trained over each successive time-slice, those topics making up the bulk of our dataset were well represented in our dataset, and therefore strengthened by the sample sizes (note the steep increase on the first day of training)

2. There exists a strong correlation between the change in size of our day-to-day training population and the representation of the top 12 topics

3. The model represents the true phenomenon within the dataset

To our third point, later iterations of our training algorithm trained on smaller than average representations of the data (due to Twitter's Rate Limits), but continued to exhibit this trend. It may be that these topics are in fact "trending" on Twitter and therefore accumulating in representation over the specified period.

Additionally, each of the top 12 topics exhibited coherence scores in the range -1.12 and -3.37 suggesting that terms found in these topics co-occur in topics to a greater extent than they appear separately. We will therefore investigate whether this is a result of terms that appear in many topics together, or whether these terms are relatively unique to these topics.

4.2. Interpreting Topic Representations

As Figure 9 in Appendix A. demonstrates, we broke down each day into its respective topic representations, sampling tweets that scored high in their respective topics and providing the top 10 words that best indicate a given topic. Table 4 represents the results of our labeling process, as of April 13th. Almost every topic in our top 10 list is unique, and all are readily interpretable, a strong sign of a successful topic model.

Topic Interpretations		
Topic #	Words	Label
46	time, like, need, know, world, day, life, think, going, good	Status Updates
55	trump, president, american, america, democrat, vote, response, china, republican, obama	US Politics
17	mask, worker, nurse, ppe, hospital, patient, medical, frontline, face_mask, healthcare_worker	Healthcare
52	case, death, new, state, new_york, total, update, county, city, reported	Reports
23	death, test, number, testing, case, vaccine, infection, rate, data, patient	Infection & Testing
21	support, community, thank, help, crisis, health, response, team, excellent, time	Positive Response
1	online, business, help, student, support, resource, free, program, new, school	Personal Finances
35	supply, staff, ppe, company, equipment, worker, player, medical, employee, testing	Medical Resources
20	stay_home, stay_safe, social_distancing, safe, stay, home, lockdown, save_lives, healthy, easter	Social Distancing
39	trump, state, january, response, election, economic, warned, american, government, warning	American Response

Table 4: Topic Word Representations for April 13th and Custom Labels

For example, it is interesting to note that topic 20, which we labeled Social Distancing, features the word "easter" as one of its key-words. This term was not present in topic 20 on April 3rd which indicates a greater concern about the upcoming Christian holiday and its associated social gatherings. We observe a similar phenomenon in topic 1, Personal Finances, which experiences a slight shift in importance from terms associated with small businesses, such as loans to programs related to schools and students, which is in-line with both student graduations and the government stimulus timeline. Topic 35, Medical Resources, which is similar to topic 17, Healthcare, but with a greater focus on supplies and equipment, experiences topic drift in line with ventilator deliveries, which were finally distributed in the United States in the first week of April, at the peak of the outbreak. A sample summary of some of the changes visible in the dataset is available in Figure 8.

We have also come to understand that one of the largest topic pools relates to American politics and policy. Topics 55 and 39 both lead with the term Trump, with topic 55 focusing more on the upcoming election, with terminology focusing on votes, democrats, and republicans, while topic 39 focusing more on terms relating to the governments response, as evidenced by the prominence of the term "january", which is featured in tweets primarily discussing the delayed reaction of the United States to the virus. In fact, the Dynamic Topic Model also effectively captured the president's son-in-law Jared Kushner's increased role in the U.S. COVID-19 taskforce, which was announced on April 3rd, before commentary on his role slowly diminished over time.

4.3. Unpopular Topics & Over-Generalization

While we have discussed the state and strengths of our model, specifically as it is able to effectively and intuitively capture term and topic trends over time, it is important to discuss certain weaknesses of this Sequential LDA implementation and discuss some less-popular (as a % of Dataset) categories.

To begin with, the most-popular category, at times with presence in ∼50% of tweets, which we

have labeled as Status Updates due to the general nature of the corpus of words that represents it, is a bit too general. We note that terms such as "time", "like", and "need" tend to appear together in tweets and therefore in our topics. Our reliance on topic coherence, a standard practice in LDA modeling, may skew the proportion of tweets that belong to this category by scoring it higher as a result of the co-occurence of these terms. This is confirmed by the degree to which these terms occur in other topics, relative to the other top topic categories.

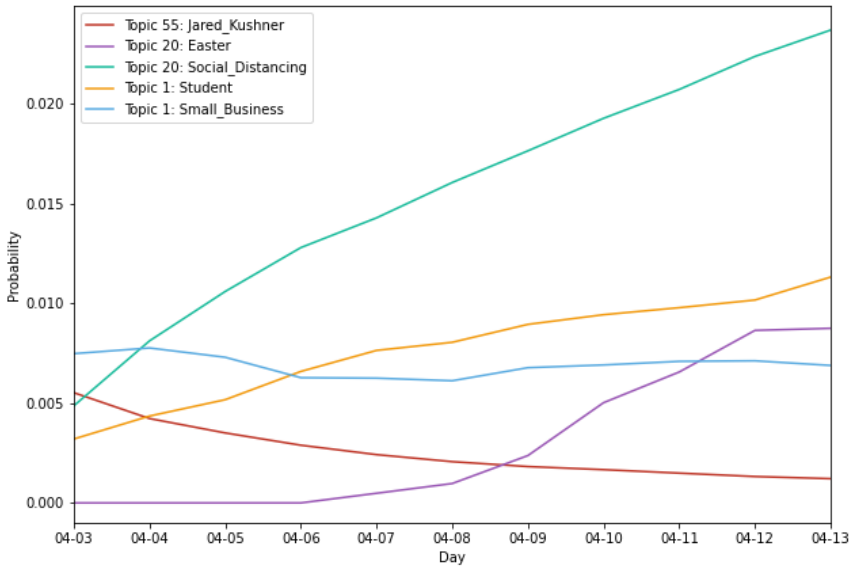

Figure 8: Changes in Topic-Word Probabilities over Time

A potential solution for this issue is to add a stop-word filter for terms such as "like", "going", or "think" but we must do so with a high degree of confidence that these terms do not in-fact relate to any latent topics. For instance, the term "think" might express a larger amount of self-expression as compared to other categories (hence our Status Update Label). Topic models are designed to splice documents into their root topics through their representational word probabilities. In this case, we should ask whether the relative size of the topic is proportional to its contribution to the document. If not, as may be the case here, removing certain key-terms of these topics may benefit the model.

We must also point out that, although certain topics within the model evolve well over time, other smaller topics, which cover more niche but still widely discussed subjects have more drastic evolutions over time. Because we are working in the time-span of weeks, topics are more likely to rise and fall, with their word-topic probabilities evolving accordingly. This is also a detriment of considering a fixed number of topics from the beginning, which amounts to a pseudo-zero-sum effort among topics. In doing so, we limit emerging topics to a pre-existing and fixed topic-word space, forcing existing topics to potentially change as a result. Major topics presented in this study however, do exhibit relatively consistent trends over time and tend to encapsulate domains rather than events rather well.

5. Conclusion

At the beginning of this study, we set out to build a working dynamic topic model to be applied to a large (and growing) dataset of tweets specifically concerning the 2019-2020 COVID-19 pandemic. We demonstrated a reproducible, robust technical solution that spanned the entire data processing pipeline, from Data Acquisition to Data Modeling, covering an online storage solution and thorough preprocessing, tokenization, and vectorization effort in between.

Our approach differentiated itself in both scale and scope, utilizing advanced Sequential Latent Dirichlet Allocation to study the emerging topics concerning COVID-19 at a scale that few works have been able to achieve. By grasping the topic early, we were able to stream a sufficiently large corpus of tweets live (measuring in the 100's of millions), building a domain-specific corpus to be

used in both current and future works. In this way, we contributed to the cross-sectional field of Urban Research and Big Data.

Through our SeqLDA model, we contributed to an understanding of both the topics surrounding the COVID-19 pandemic and their evolution over time. Specifically, we identified 12 of the most popular topics present in our dataset over the period spanning April 3rd to April 13th and discussed their growth and changes over time. These topics were both robust, in that they covered specific domains, not simply events, and dynamic, in that they were able to change over time in response to rising trends in our dataset. They spanned politics, healthcare, community, and the economy, and experienced macro-level growth over time, while also exhibiting micro-level changes in topic composition.

As a result of our research into Dynamic Topic Modeling with a spatial (geographical) component, we proposed a novel solution to accommodate Twitter's recent changes in geo-location policies. We fully intend to extend this dynamic research effort into understanding not only the composition of topics across the Twitter platform and its user base but also understanding their respective geographic topic compositions and their changes over time in response to the growth (and subsequent retreat) of the pandemic.

Whereas we are optimistic towards the future, we also understand that this is an unprecedented time that will have lasting impacts on individuals and society at large, impacting not only the economy or geo-politics, but human behavior and psychology. Therefore, in more ways than one, this research is just beginning to scratch the surface of what will be a concerted research effort into studying the history and repercussions of COVID-19.

References

[1] Deerwester, Dumais, Landauer, Furnas, and Harshman, "Indexing by latent semantic analysis," *Association of Information Science Technology*, vol. 41, no. 6, pp. 391–407, 1990. [Online]. Available: http://lsa.colorado.edu/papers/JASIS.lsi.90.pdf

[2] D. Blei, A. Ng, and M. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, no. 3, 2003. [Online]. Available: http://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf

[3] L. Du, W. Buntine, H. Jin, and C. Chen, "Sequential latent dirichlet allocation," *Knowledge and Information Systems*, no. 31, pp. 475–503, 2011. [Online]. Available: https://link.springer.com/article/10.1007/s10115-011-0425-1

[4] C. Zhang and J. Sun, "Large scale microblog mining using distributed mb-lda," *International World Wide Web Conference*, 2012. [Online]. Available: https://www2012.universite-lyon.fr/proceedings/companion/p1035.pdf

[5] B. Huang, Y. Yang, A. Majmood, and W. Hongjun, "Microblog topic detection based on lda model and single-pass clustering," *International Conference on Rough Sets and Current Trends in Computing*, vol. 7413, pp. 166–171, 2012. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-642-32115-3_19

[6] H. Zhao and X. Yan, "Chinese microblog topic detection based on the latent semantic analysis and structural property," *Journal of Networks*, vol. 8, no. 4, pp. 917–923, 2013. [Online]. Available: https://www.semanticscholar.org/paper/Chinese-Microblog-Topic-Detection-Based-on-the-and-Yan-Zhao/2a00a3ec453e784430d9e61526a9c48ecbb73d03

[7] J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant, "Detecting influenza epidemics using search engine query data," *Nature*, vol. 457, pp. 1012–1014, 2009. [Online]. Available: http://dx.doi.org/10.1038/nature07634

[8] M. Paul and M. Dredze, "A model for mining public health topics from twitter," *Human Language Technology Center of Excellence*, 2010. [Online]. Available: http://www.cs.jhu.edu/~mpaul/files/2011.tech.twitter_health.pdf

[9] ——, "You are what you tweet: Analyzing twitter for public health," *Human Language Technology Center of Excellence*, 2011. [Online]. Available: http://www.cs.jhu.edu/~mpaul/files/2011.icwsm.twitter_health.pdf

[10] A. Signorini, A. M. Segre, and P. M. Polgreen, "The use of twitter to track levels of disease activity and public concern in the u.s. during the influenza a h1n1 pandemic," *PLoS ONE*, pp. 579–596, 2011. [Online]. Available: https://doi.org/10.1371/journal.pone.0019467

[11] M. Roy, N. Moreau, C. Rousseau, A. Mercier, A. Wilson, and L. Atlani-Duault, "Ebola and localized blame on social media : Analysis of twitter and facebook conversations during the 2014-2015 ebola epidemic," *Culture, Medicine, and Psychiatry*, no. 44, pp. 56–79, 2019. [Online]. Available: https://doi.org/10.1007/s11013-019-09635-8

[12] W. Ahmed, P. Bath, L. Sbaffi, and G. Demartini, "Novel insights into views towards h1n1 during the 2009 pandemic: a thematic analysis of twitter data," *Health Information and Libraries Journal*, vol. 50, no. 36, pp. 50–72, 2019.

[13] D. Pruss, Y. Fujinuma, A. Daughton, M. Paul, B. Arnot, D. Szafir, and J. Boyd-Graber, "Zika discourse in the americas: A multilingual topic analysis of twitter." *PLoS ONE*, no. 5, pp. 579–596, 2019. [Online]. Available: https://doi.org/10.1371/journal.pone.0216922

[14] F. Yao and Y. Wang, "Tracking urban geo-topics based on dynamic topic model," *Computers, Environment and Urban Systems*, vol. 79, 2019. [Online]. Available: https://doi.org/10.1016/j.compenvurbsys.2019.101419

[15] D. Blei and J. Lafferty, "Dynamic topic models," *Proceedings of the 23rd International Conference on Machine Learning*, no. 20, 2006. [Online]. Available: https://mimno.infosci.cornell.edu/info6150/readings/dynamic_topic_models.pdf

A. Appendix

Day	Topic	Words	Tweet
3	46	['like', 'time', 'need', 'know', 'help', 'day', 'world', 'work', 'going', 'right']	A virtual influencer is helping to drive donations in WHO's new COVID-19 campaign https://t.co/tWdBRxSjDv https://t.co/J65xNzareA
3	17	['mask', 'face_mask', 'help', 'ppe', 'wear', 'worker', 'hospital', 'nurse', 'need', 'wear_mask']	Thank you from the Manila Protective Gear Sewing Club!! https://t.co/rihggYi4Jp
3	21	['time', 'support', 'help', 'news', 'need', 'new', 'crisis', 'case', 'community', 'thank']	Apple highlights apps to help families manage autism amid the coronavirus - CNET https://t.co/UzeWZnQB1D
3	52	['case', 'county', 'death', 'new', 'state', 'breaking', 'hospital', 'reported', 'trump', 'news']	A 13th employee at the Cook County Circuit Court clerk's office has tested positive for COVID-19. https://t.co/yNP7T6HKHa
3	35	['staff', 'supply', 'ppe', 'fight', 'medical', 'hospital', 'help', 'testing', 'need', 'country']	When an IAS officer wears a full PPE kit and doctors wear normal clothes and mask.....I don't think Modi's promise is working here #uglyindianbureaucracy @PMOIndia @narendramodi @drharshvardhan @DrHarjitBhatti @drpankajsolanki @UnitedRda @FordaIndia @RajCMO @CMODelhi https://t.co/Uqojffl5oF
4	46	['like', 'time', 'need', 'know', 'world', 'think', 'going', 'day', 'help', 'right']	@narendramodi @dm_ghaziabad @CMOfficeUP dear sir kindly spare some time to discuss the serious matter regarding pre paid electricity meter and maintenance charges in River Heights Ph 2 Raj Nagar Ext.,Ghaziabad as the builder is enjoying all the liberty against law.kindly act fast
4	52	['case', 'county', 'death', 'state', 'new', 'new_york', 'breaking', 'reported', 'update', 'total']	Medical personnel transfer bodies into and out of a refrigerated truck placed outside #Brooklyn Hospital Center in New York City #NYC Wednesday in order to deal with a spike in #COVID19 deaths #NewYork has confirmed 102,870 cases and at least 2,935 deaths https://t.co/jCCZla3Okq
4	35	['supply', 'staff', 'ppe', 'medical', 'fight', 'equipment', 'hospital', 'help', 'government', 'passenger']	@jensstoltenberg Mr.Stoltenberg, #Turkey is sending everyday #illegal #migrants with #COVID2019 to the Greek Islands. What kind of solidarity is this? Why #NATO didn't stop it? #europeanborderguard #greece_under_attack
4	21	['support', 'time', 'help', 'news', 'need', 'new', 'community', 'crisis', 'thank', 'excellent']	DOH says medical team from China will stay in the country until April 19
4	17	['mask', 'face_mask', 'wear_mask', 'wear', 'ppe', 'help', 'nurse', 'hospital', 'worker', 'patient']	@ME31017974 @Romeites @KKMPutrajaya wear mask 😷. Mask is important so that asymptomatic cases don't continue to spread C19. See video 👇.I will be spreading this news internationally but for the time being, since Msia is not receiving anyone frm outside, this is the time we can contain C19 frm inside #FightMrCovid https://t.co/g1j4TToGk5
5	46	['like', 'time', 'need', 'know', 'world', 'think', 'going', 'day', 'life', 'right']	What if the entire world shuts down completely until every last infected person has recovered and the virus is extinct before quarantine ends? And what if you're that last guy, and the world is tapping it's foot...
5	35	['supply', 'staff', 'ppe', 'equipment', 'medical', 'fight', 'hospital', 'player', 'passenger', 'ventilator']	#COVID19 Govt is giving liberty in paying EMIs, where as #landcraft builder imposing additional financial burden by starting deduction of maintenance charge from pre paid electric meter @dm_ghaziabad @gdagzb @myogioffice @JansunwaiAbhiyn
5	17	['mask', 'face_mask', 'ppe', 'nurse', 'worker', 'hospital', 'patient', 'wear_mask', 'wear', 'help']	In our #COVID19 blog post, we shared this 45 second "How to" video from the #SurgeonGeneral on creating your #facemasks at home. We'll be doing this to help ensure we aren't inadvertantly spreading the virus even though we are asymptomatic: https://t.co/GVOA15MQ5x #cdc https://t.co/3TwWWMr8C0
5	52	['case', 'death', 'county', 'new', 'state', 'new_york', 'breaking', 'total', 'update', 'reported']	'Now is not the time to discuss pay rise for nurses', says health secretary Matt Hancock. You still clapping for the NHS but voting Tory? Mad.
6	46	['like', 'time', 'need', 'know', 'world', 'think', 'going', 'day', 'life', 'right']	"No new fracking bans without scientific research" going well, @GavinNewsom. Way to be a climate leader! https://t.co/4dybYJ7FTo
6	21	['support', 'help', 'community', 'time', 'crisis', 'thank', 'excellent', 'need', 'news', 'new']	Korea is managing the COVID-19 crisis by emphasizing transparency and open communication, public-private partnerships, evidence-based deployment of public health measures, and innovative use of technology and data. https://t.co/4epOdXkh2k via @WB_AsiaPacific
6	35	['supply', 'staff', 'equipment', 'ppe', 'medical', 'fight', 'hospital', 'player', 'passenger', 'worker']	Working from home- the desperation 😬🤣 #covid19 #coronavirus #humour https://t.co/q2hSnyh2Vp
6	17	['mask', 'nurse', 'ppe', 'worker', 'hospital', 'face_mask', 'patient', 'medical', 'wear', 'help']	Q&A on coronaviruses (COVID-19) Most common symptoms of #COVID19 are fever, tiredness, & dry cough. Some: aches & pains, nasal congestion, runny nose, sore throat or diarrhea. People with fever, cough & difficulty breathing should seek medical attention. https://t.co/hOdLwM6qbV
6	52	['case', 'death', 'new_york', 'new', 'state', 'county', 'breaking', 'update', 'total', 'reported']	VDH Daily Update: 2878 Coronavirus Cases in Va. with 54 Deaths https://t.co/RiVf8Y857I
7	46	['like', 'time', 'need', 'know', 'world', 'day', 'think', 'going', 'life', 'right']	This ain't it... IDK what the city is doing but they need to reapply that money into the youth another way if they are not going to provide jobs.
7	17	['mask', 'nurse', 'worker', 'ppe', 'hospital', 'patient', 'face_mask', 'medical', 'health', 'help']	Nurses are not getting the protections they desperately need to fight #COVID19. @NationalNurses is demanding Congress act now. Add your name: https://t.co/YTc3RGD6Xw
7	52	['case', 'death', 'new', 'new_york', 'state', 'county', 'breaking', 'update', 'total', 'city']	Nearly 1 in 6 people who have died from the new coronavirus in New York state lived in a nursing home. #NY1Health https://t.co/ISWKiGRIxf
7	21	['support', 'community', 'thank', 'help', 'health', 'crisis', 'time', 'world', 'excellent', 'need']	2/ for @TheContentMine openVirus we need volunteers who know how to index abstracts and theses using SOLR. We have a great team who can retrieve and a great infrastructure. Somewhere in the existing scholarly literature are solutions to tackling the pandemic.
7	35	['supply', 'staff', 'equipment', 'medical', 'ppe', 'fight', 'player', 'worker', 'hospital', 'company']	Coronavirus Impact: Jackson Memorial Hospital CEO Carlos Migoya Questioned Over 'Furloughs, Pay Cuts, Not Providing Protective Equipment' https://t.co/9ks4JVkAoh
8	21	['support', 'community', 'thank', 'help', 'health', 'crisis', 'time', 'response', 'world', 'need']	Proud to have served, and holding vigil for all my brothers and sisters on the front lines in healthcare. #HealthcareHeroes #COVID19
8	46	['like', 'time', 'need', 'know', 'world', 'day', 'think', 'going', 'life', 'right']	My granny is 92 today. Here she is, celebrating away to herself while she stays isolated. Can you fathom the joy? 😇🥰🤗 https://t.co/Clzw1FrJIP
8	35	['supply', 'staff', 'equipment', 'medical', 'worker', 'ppe', 'player', 'fight', 'hospital', 'company']	Hotline service to begin for foreigners in Japan for virus inquiries https://t.co/iXtUdvv4Yi
8	52	['case', 'death', 'new', 'state', 'new_york', 'county', 'update', 'total', 'breaking', 'city']	My daily NY COVID-19 tracker: 149,316 confirmed cases 779 deaths
8	17	['mask', 'worker', 'nurse', 'ppe', 'hospital', 'patient', 'face_mask', 'medical', 'health', 'help']	#Researchers at @NotreDame have developed micro/nanofluidic devices for isolating cellular material such as #vesicles and #exosomes during #liquidbiopsies. They turn out to be the same size as the #coronavirus. - @NDCBE https://t.co/v3H4U2uu1b

9	46	['like', 'time', 'need', 'know', 'world', 'day', 'going', 'think', 'life', 'good']	Best estimates of #R0 for #coronavirus seem to be 2-4. and yet it feels like it spreads very easily. So I can't even imagine what it would feel like if we were in an outbreak of something with a much higher R0 and we didn't have a vaccine, like #measles.
9	21	['support', 'community', 'thank', 'help', 'health', 'crisis', 'response', 'time', 'need', 'team']	No matter what challenges we face, we're always #VegasStrong. @UNLVathletics is going to make and donate 3,000 cloth masks to @UMCSN to respond to the #COVID pandemic—proud to see them stepping up and helping our community in need. https://t.co/yC3uA66eR8
9	52	['case', 'death', 'new', 'state', 'new_york', 'county', 'update', 'total', 'breaking', 'city']	Tasmania's north-west coast residents preparing for 'a whole new world' as coronavirus lockdown looms - ABC News https://t.co/MOqjprpB6U
9	17	['mask', 'worker', 'nurse', 'ppe', 'hospital', 'patient', 'medical', 'face_mask', 'health', 'protect']	The @CDCgov recently recommended anyone over 2-years-old should wear a mask outside. Dr. Julia Sammons, Medical Director of our Department of #InfectionPrevention & Control, spoke with @WhatToExpect about how to protect children under 2. #COVID19 https://t.co/lGBjmQRPn9
9	35	['supply', 'staff', 'equipment', 'worker', 'ppe', 'medical', 'company', 'player', 'hospital', 'fight']	As demand spikes, meds used alongside ventilators are in short supply. Manufacturers must ramp up drug production even more, says 💊drug shortage expert @foxerinr. All #COVID19 patients deserve access to #MedsWeCanTrust: @edsilverman https://t.co/gSJ7jl6z2Q
10	46	['like', 'time', 'need', 'know', 'world', 'day', 'good', 'going', 'think', 'life']	A Communal Virus and Our Collective Irrationality - India Gone Viral https://t.co/wgcyb8cOJU
10	52	['case', 'death', 'new', 'state', 'new_york', 'update', 'county', 'total', 'city', 'breaking']	New York sees record 1-day rise in COVID-19 deaths https://t.co/RUq1Vs8oJH
10	21	['support', 'community', 'thank', 'help', 'health', 'crisis', 'response', 'time', 'team', 'need']	@ASlavitt @ScottGottliebMD @NIHDirector @NYGovCuomo @GovNedLamont @GovMurphy are you watching #maddow covering #nursinghome #longtermcare #covid19 outbreaks nationwide? What can we implement as a national plan for this? #elderly #vulnerable #highriskcovid #coronavirus @maddow
10	35	['supply', 'staff', 'equipment', 'ppe', 'worker', 'company', 'medical', 'player', 'hospital', 'fight']	A #Canadian cargo company says international air carriers are #pricegouging the cost to ship products. It says there is minimal cargo space available due to the #COVID−19 pandemic, and prices are changing daily. @JGaidolaCHCH has the details WATCH: https://t.co/9jprmkdutt https://t.co/ZmECk2sRma
10	17	['mask', 'worker', 'nurse', 'ppe', 'hospital', 'patient', 'medical', 'face_mask', 'health', 'protect']	How Did the U.S. End Up with Nurses Wearing Garbage Bags? https://t.co/ibg2XtAUYE
11	46	['like', 'time', 'need', 'know', 'world', 'day', 'think', 'good', 'going', 'life']	Italian Prime Minister Giuseppe Conte floated the idea of all of Europe doing "things on their own" due to the European Union's poor response to Italy's tragic experience with the Chinese coronavirus. https://t.co/e1fv4z2OyR
11	17	['mask', 'worker', 'nurse', 'ppe', 'hospital', 'patient', 'medical', 'face_mask', 'frontline', 'health']	To the health care and essential workers putting themselves on the line during this pandemic -- we owe you a massive debt of gratitude. That's why this week, I proposed the COVID-19 Heroes Fund to give our frontline workers the compensation they deserve. https://t.co/2JlnH6fiyp
11	52	['case', 'death', 'new', 'state', 'new_york', 'update', 'total', 'county', 'city', 'breaking']	#COVID19 update as of 6:00 pm: #Austin now has 744 confirmed cases, up 54 from yesterday. We remain at 9 deaths & have had 133 people recover. We are starting to see an increase in cases in the 40-59 age range, & while cases are rising, they've been at a steady rate (1/7) https://t.co/DmtN2ORxyx
11	21	['support', 'community', 'thank', 'help', 'health', 'crisis', 'response', 'team', 'time', 'need']	leading Muslim charities from across the UK have united to pool their resources, skills and expertise ensuring that they can efficiently provide support to where it's most needed support and share if you can pls Campaign for National Solidarity COVID-19 https://t.co/R3JTXkIprF
11	35	['supply', 'staff', 'equipment', 'ppe', 'worker', 'company', 'medical', 'player', 'hospital', 'fight']	🛵Sign our petition calling on food delivery companies to cut their commissions! #CommunityOverCommission 👉 https://t.co/GMszvxuBPe
12	46	['like', 'time', 'need', 'know', 'world', 'day', 'life', 'think', 'going', 'good']	@HusainHacker Happy birthday! May God bless u with some jari butti so that tum safe raho Corona virus se 😂
12	52	['case', 'death', 'new', 'state', 'new_york', 'total', 'update', 'county', 'city', 'breaking']	Green Lake County Public Health said the person was hospitalized and is in stable condition. https://t.co/jOSl89vyMx
12	38	['easter', 'god', 'jesus', 'church', 'lord', 'pray', 'christian', 'wash', 'prayer', 'holy']	HAVE A BLESSED EASTER!🐣🙏 Rejoice in the risen Christ! Rejoice by faith! Rejoice when we are still in our night, (#covid19) for Jesus is the Light Who is sure to dawn. (2 Pt 1:19; cf Rv 22:16). Alleluia forever! https://t.co/Yr9XlQUhoN
12	21	['support', 'community', 'thank', 'help', 'health', 'crisis', 'response', 'team', 'excellent', 'time']	BP Oman has chartered a flight to bring back students and also dependents of Omanis from the UK in cooperation with @MofaOman, @OmanEmbassyUK and @omanair as part of our response to Covid-19. @BP_Oman https://t.co/o6JrZMXvzz
12	17	['mask', 'worker', 'nurse', 'ppe', 'hospital', 'patient', 'medical', 'face_mask', 'frontline', 'protect']	"It was the police who sloughed off social-distancing: physically handling him, cuffing him without wearing protective gear as unworn masks dangled from their belts" https://t.co/firTL6IJME
13	46	['time', 'like', 'need', 'know', 'world', 'day', 'life', 'think', 'going', 'good']	Coronavirus brings out the worst haircuts in all of us.
13	17	['mask', 'worker', 'nurse', 'ppe', 'hospital', 'patient', 'medical', 'frontline', 'face_mask', 'healthcare_worker']	A major California labor union that claimed to have discovered a stockpile of 39 million masks for health care workers fighting the coronavirus was duped in an elaborate scam uncovered by FBI investigators. https://t.co/CP1zeglOQl
13	52	['case', 'death', 'new', 'state', 'new_york', 'total', 'update', 'county', 'city', 'reported']	The new cases brings the total to 17 in the county https://t.co/7OYoHs34Kq
13	21	['support', 'community', 'thank', 'help', 'crisis', 'health', 'response', 'team', 'excellent', 'time']	Here's our roundup of the Top Ten resources about coronavirus. Atrium Health experts weighed in to separate fact from fiction. Keep this link handy as you prepare for the week ahead. https://t.co/m5JTofSY7S #AtriumHealthProud
13	35	['supply', 'staff', 'ppe', 'company', 'equipment', 'worker', 'player', 'medical', 'employee', 'testing']	Farm workers cannot carry on with harvesting and planting, with truckers and air freight capacity having ground to a halt due to #COVID19 crisis. #FoodAndRightsNow https://t.co/GBvQq7WRXP

Figure 9: Top 5 Topics in Each Day with Topic Word Representations and Tweet Sample