Discussing the Coronavirus Pandemic

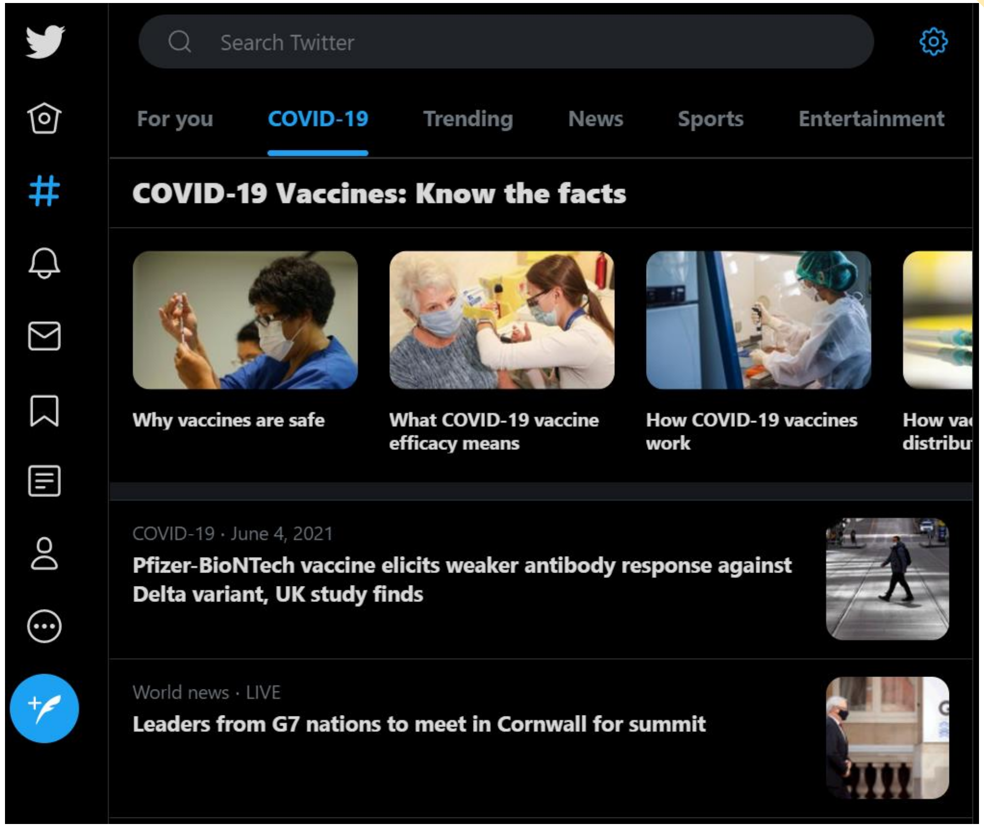
Capstone - Progress Standup

By:

Shreyas Chitransh

Introduction

- Covid-19 culminated in fear and grief but also some of the best solidarity amongst the global citizens;
- Analyze Twitter data from the early months of the pandemic using Unsupervised ML;
  - Topic Modeling - highlight the topics being discussed;
  - Entity analysis - for government bodies, entities and other influencers.



COVID-19 Vaccines: Know the facts

Why vaccines are safe

What COVID-19 vaccine efficacy means

How COVID-19 vaccines work

How va distribu

COVID-19 · June 4, 2021

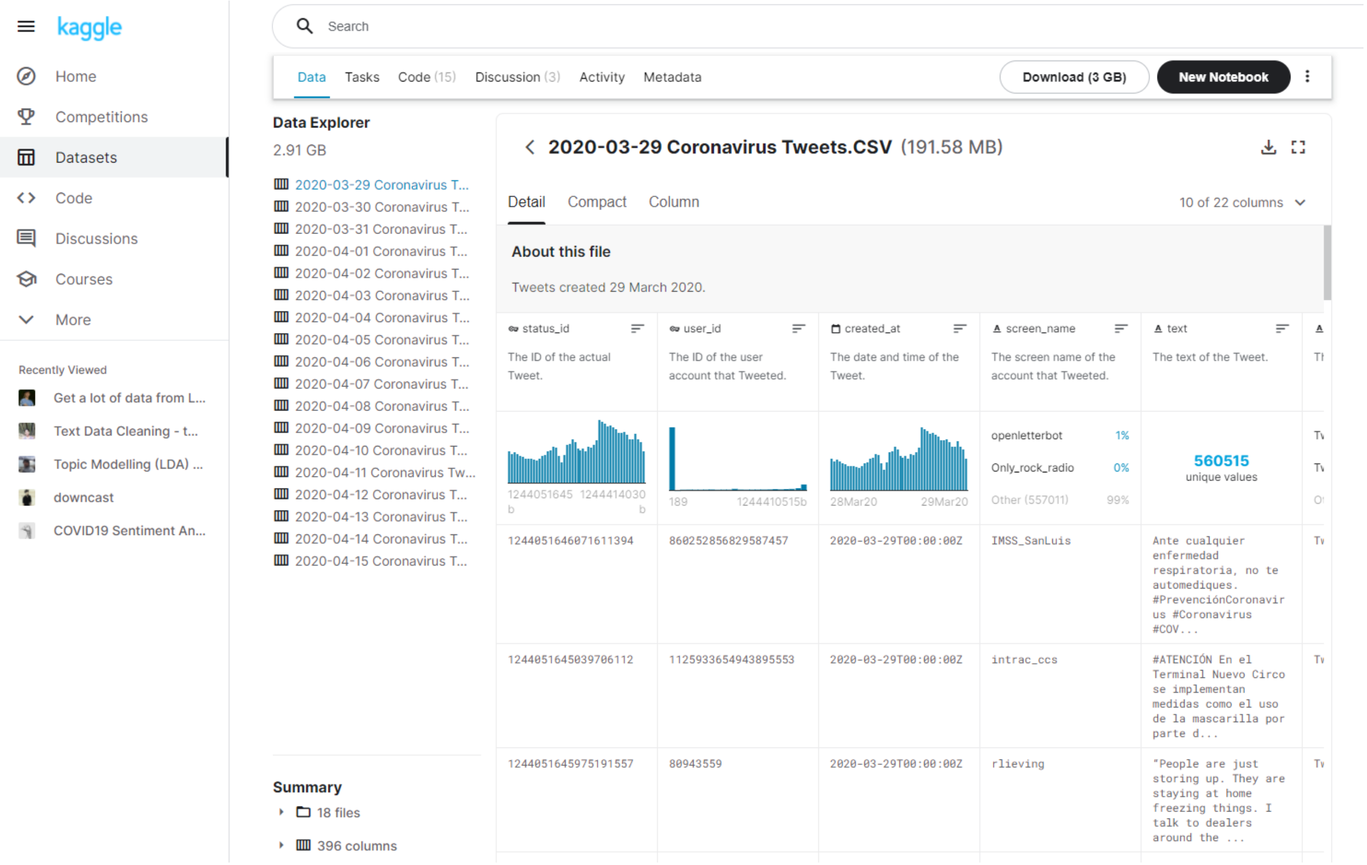
Pfizer-BioNTech vaccine elicits weaker antibody response against Delta variant, UK study finds

World news · LIVE

Leaders from G7 nations to meet in Cornwall for summit

Data Gathering

- Initial attempt to hydrate ongoing Tweets;
  - Get csv of Tweet ID and download them yourself;
  - Multiple Twitter application suspensions.
- Kaggle dataset >4GB of Tweets spanning 1 month;
  - CSV Files



Data Explorer

2.91 GB

- 2020-03-29 Coronavirus T...
- 2020-03-30 Coronavirus T...
- 2020-03-31 Coronavirus T...
- 2020-04-01 Coronavirus T...
- 2020-04-02 Coronavirus T...
- 2020-04-03 Coronavirus T...
- 2020-04-04 Coronavirus T...
- 2020-04-05 Coronavirus T...
- 2020-04-06 Coronavirus T...
- 2020-04-07 Coronavirus T...
- 2020-04-08 Coronavirus T...
- 2020-04-09 Coronavirus T...
- 2020-04-10 Coronavirus T...
- 2020-04-11 Coronavirus Tw...
- 2020-04-12 Coronavirus T...
- 2020-04-13 Coronavirus T...
- 2020-04-14 Coronavirus T...
- 2020-04-15 Coronavirus T...

Summary

- 18 files
- 396 columns

2020-03-29 Coronavirus Tweets.CSV (191.58 MB)

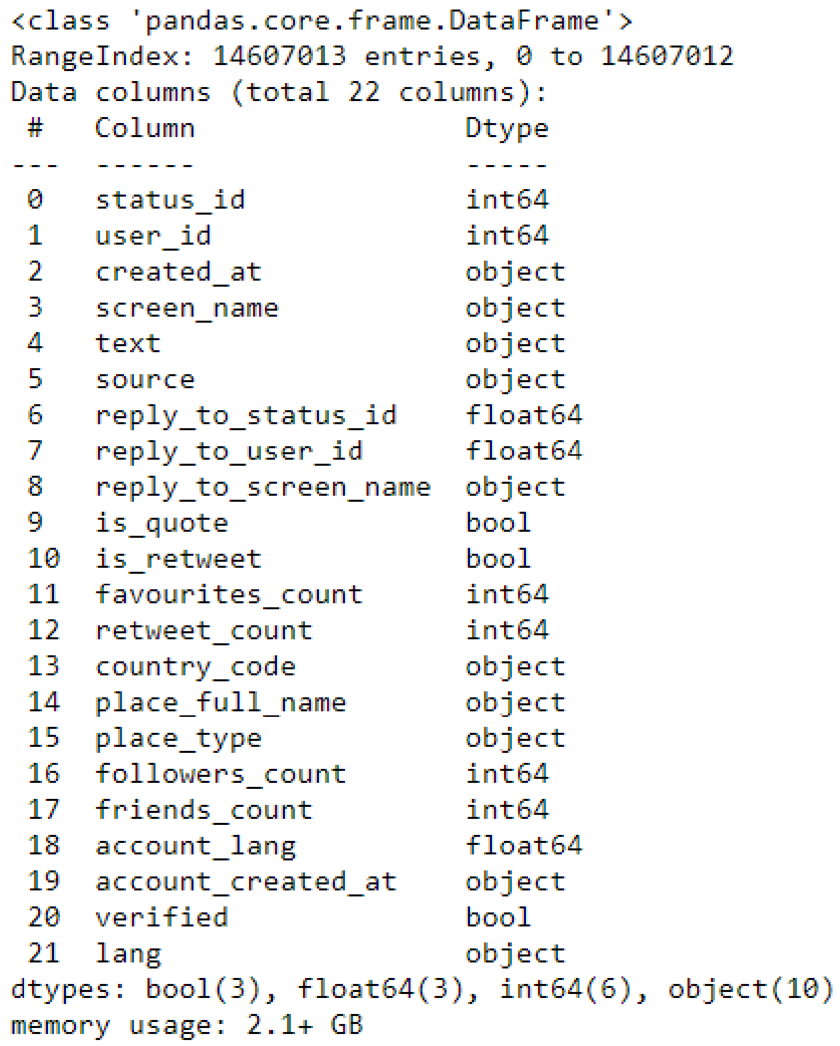
About this file

Tweets created 29 March 2020.

status_id	user_id	created_at	screen_name	text
The ID of the actual Tweet.	The ID of the user account that Tweeted.	The date and time of the Tweet.	The screen name of the account that Tweeted.	The text of the Tweet.
1244051646071611394	860252856829587457	2020-03-29T00:00:00Z	IMSS_SanLuis	Ante cualquier enfermedad respiratoria, no te automediques. #PrevenciónCoronavirus #Coronavirus #COV...
1244051645039706112	1125933654943895553	2020-03-29T00:00:00Z	intrac_ccs	#ATENCIÓN En el Terminal Nuevo Circo se implementan medidas como el uso de la mascarilla por parte d...
1244051645975191557	80943559	2020-03-29T00:00:00Z	rlieving	"People are just storing up. They are staying at home freezing things. I talk to dealers around the ...

Data Details

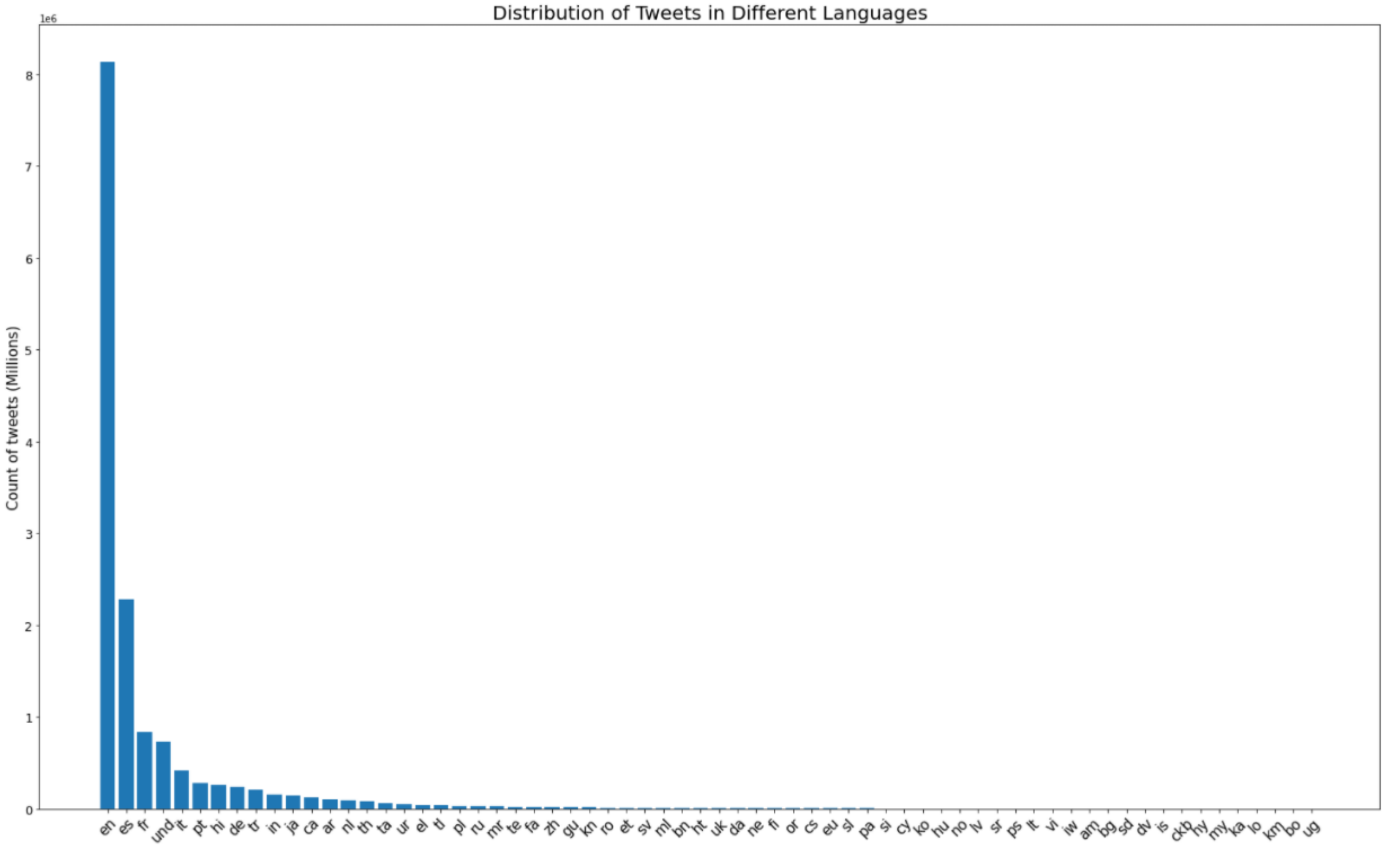
- **>14M Rows and 22 Columns**
  - **Rows : Separate Tweets**
  - **Columns : Tweet and User Metadata**
- **Metadata:**
  - **Tweet:**
    - **Status ID, Date, Time, Favorites, Retweets , Autodetected language etc.**
  - **User:**
    - **User ID, Screen Name, Account Creation Date etc.**



```
<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
```

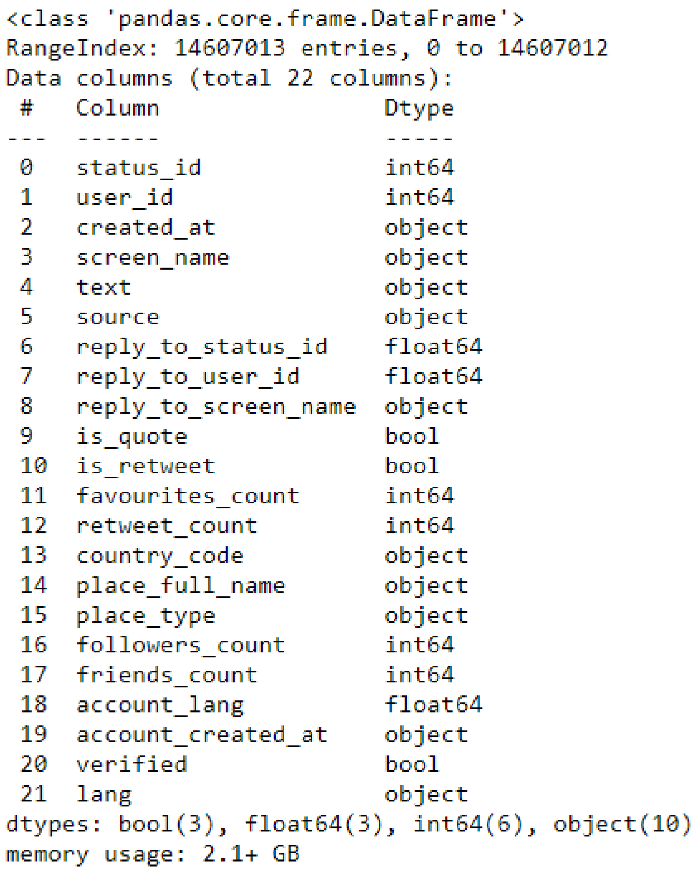
Data Clean-Up

- 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



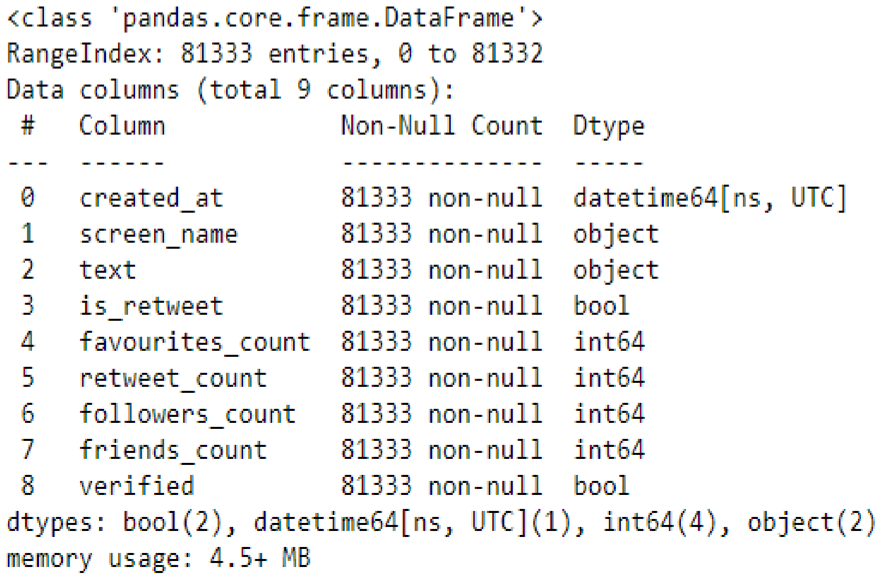
Subset Dataset – Improve Manageability

14.6M Rows with 22 Columns   -->   81.3K Rows with 9 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
```

-->

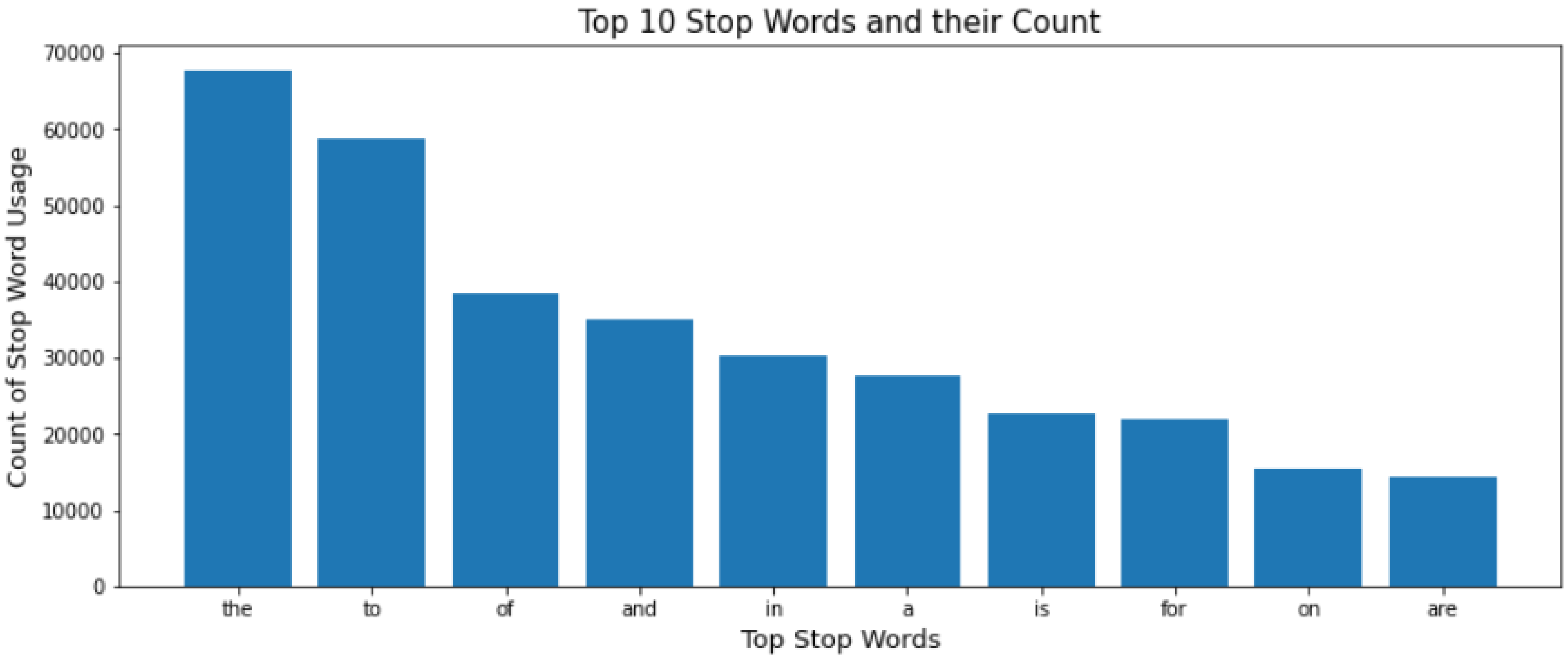


```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 81333 entries, 0 to 81332
Data columns (total 9 columns):
 #   Column            Non-Null Count  Dtype              
---  ------            --------------  -----              
 0   created_at        81333 non-null  datetime64[ns, UTC]
 1   screen_name       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

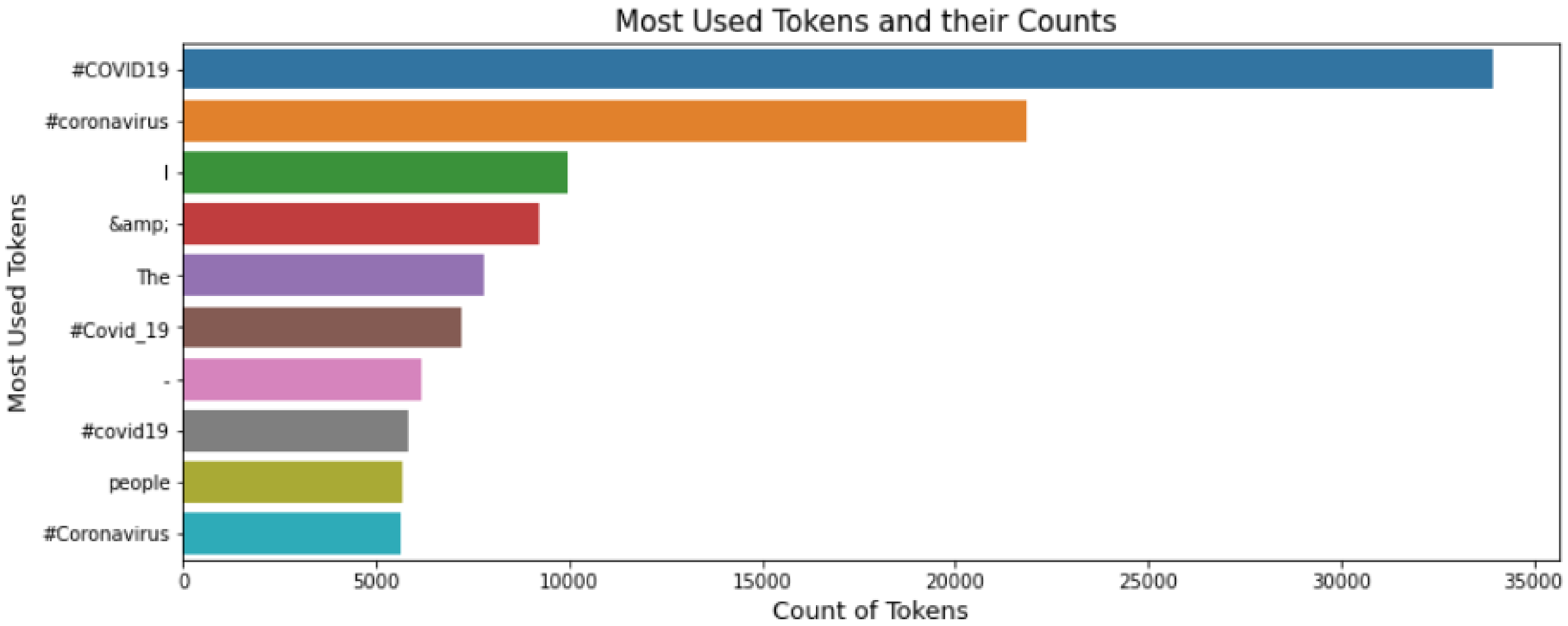
- Removing URLs (Completed)

- Removing Stopwords (Completed)



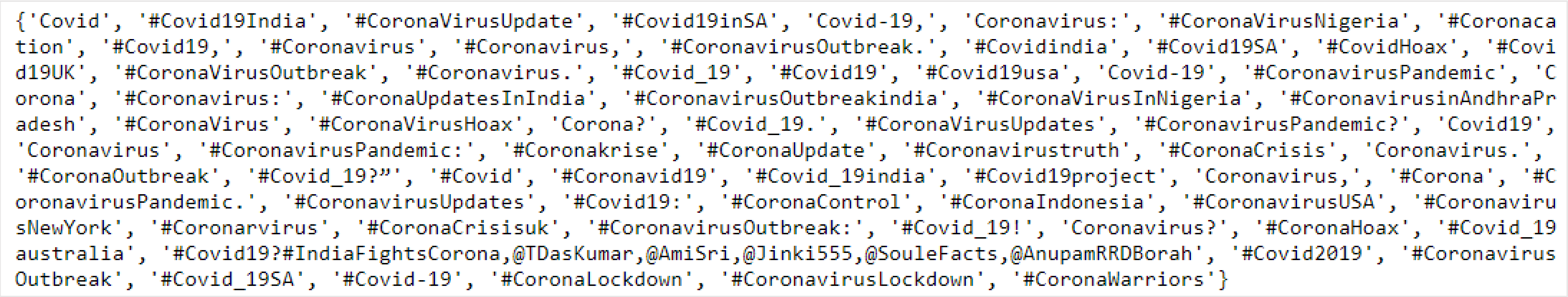
Preprocessing – Crucial Step

- Removing URLs (Completed)
- Removing Stopwords (Completed)
- Replacing Covid synonyms with “Covid19” (In Progress)



Preprocessing – Crucial Step

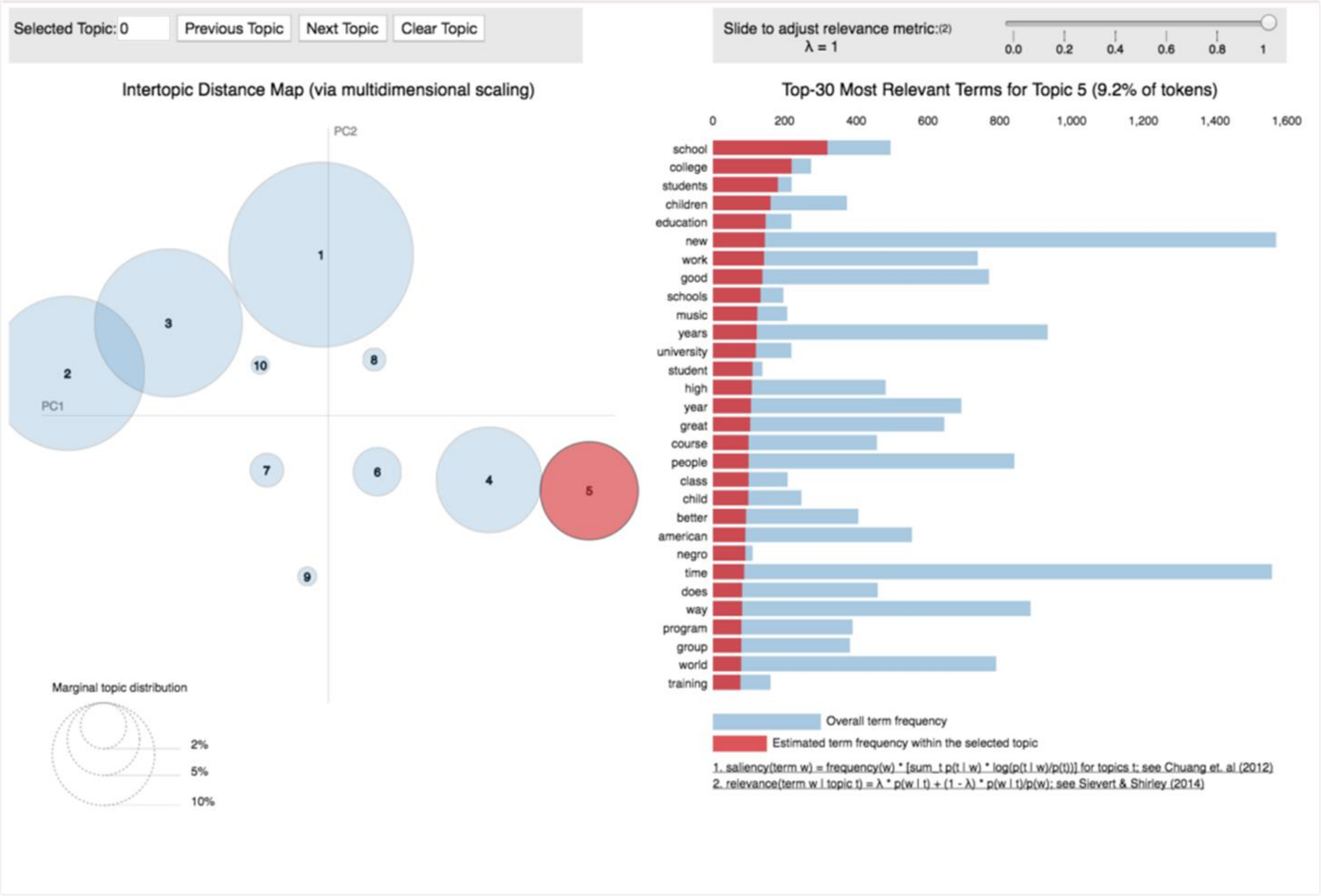
- Removing URLs (Completed)
- Removing Stopwords (Completed)
- Replacing Covid synonyms with “Covid19” (In Progress)
- Lemmatization(To Be Completed)



```
{'Covid', '#Covid19India', '#CoronaVirusUpdate', '#Covid19inSA', 'Covid-19,', 'Coronavirus:', '#CoronaVirusNigeria', '#Coronaca
tion', '#Covid19,', '#Coronavirus', '#Coronavirus,', '#CoronavirusOutbreak.', '#Covidindia', '#Covid19SA', '#CovidHoax', '#Covi
d19UK', '#CoronaVirusOutbreak', '#Coronavirus.', '#Covid_19', '#Covid19', '#Covid19usa', 'Covid-19', '#CoronavirusPandemic', 'C
orona', '#Coronavirus:', '#CoronaUpdatesInIndia', '#CoronavirusOutbreakindia', '#CoronaVirusInNigeria', '#CoronavirusinAndhraPr
adesh', '#CoronaVirus', '#CoronaVirusHoax', 'Corona?', '#Covid_19.', '#CoronaVirusUpdates', '#CoronavirusPandemic?', 'Covid19',
'Coronavirus', '#CoronavirusPandemic:', '#Coronakrise', '#CoronaUpdate', '#Coronavirustruth', '#CoronaCrisis', 'Coronavirus.',
'#CoronaOutbreak', '#Covid_19?”', '#Covid', '#Coronavid19', '#Covid_19india', '#Covid19project', 'Coronavirus,', '#Corona', '#C
oronavirusPandemic.', '#CoronavirusUpdates', '#Covid19:', '#CoronaControl', '#CoronaIndonesia', '#CoronavirusUSA', '#Coronaviru
sNewYork', '#Coronarvirus', '#CoronaCrisisuk', '#CoronavirusOutbreak:', '#Covid_19!', 'Coronavirus?', '#CoronaHoax', '#Covid_19
australia', '#Covid19?#IndiaFightsCorona,@TDasKumar,@AmiSri,@Jinki555,@SouleFacts,@AnupamRRDBorah', '#Covid2019', '#Coronavirus
Outbreak', '#Covid_19SA', '#Covid-19', '#CoronaLockdown', '#CoronavirusLockdown', '#CoronaWarriors'}
```

Plan of Action (Next 2 Weeks)

- **Use Latent Dirichlet Allocation (LDA) Topic Modeling**
  - Each document (Tweet) is represented by the distribution of topics and each topic is represented by the distribution of words.
- **Named Entity Recognition**
  - Entities are classified into predefined entity types like “Person”, ” Place” and ”Organization”
- (Stretch) Future plan to hydrate recent Tweets to run through completed model.



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)