Topic Modeling of Coronavirus Tweets

Capstone Project 2

Alex Chung

Project Goal:

2020 has been a unique year with the Coronavirus putting the world in lockdown, forcing us to adjust to “new normals” . There hasn’t been a time when people all over the world go into self-seclusion for an entire year such as we have done.

Our goal is to work with twitter feeds scraped from March 30th to April 30th, 2020 and to see if we could explore some questions in this worldwide laboratory:

- *What happens when we live in seclusion for an extended period of time?*
- *What behavioral challenges and changes occurred during this time?*
- *What new trends emerged, either positive or negative?*
- *Was there an increase in stress? Depression? Fear? Or Loneliness?*
- *Do people experience different stages in seclusion similar to the stages of grief? If so, what are these stages?*

The original data comes from from a kaggle dataset:

https://www.kaggle.com/smid80/coronavirus-covid19-tweets

Our Workflow:

1. **Prep** our data thru exploratory data analysis and data wrangling
2. **Clean**, lemmanize, tokenize our text
3. **Vectorize** our processed text to start the natural learning process using scikit-learn and spaCy.
4. Use KMeans Clustering to **discover** topical **clusters**
5. Finally, we will **extract** important topical **keywords** using Latent Dirichlet Allocation

1.

Prepping our Data

Initial exploratory data analysis

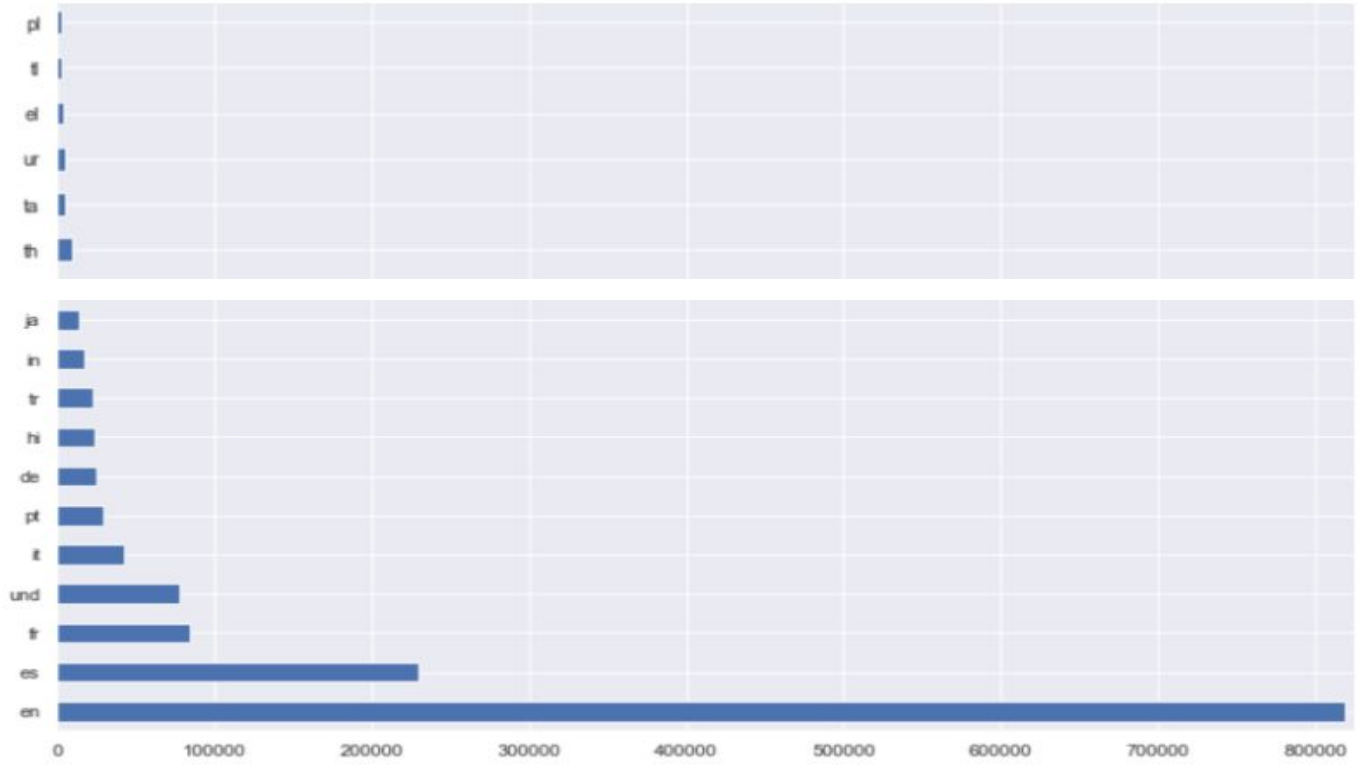
The original dataset:

- Contains 30 csv files
- Each file is a single day of tweets from 3/30/20 to 4/30/20
- Is almost 4 GB combined
- Contains 7.16 million total rows of data

The original dataset

2020-03-30 Coronavirus Tweets.CSV	204,112 KB
2020-03-31 Coronavirus Tweets.CSV	234,792 KB
2020-04-01 Coronavirus Tweets.CSV	207,864 KB
2020-04-02 Coronavirus Tweets.CSV	204,713 KB
2020-04-03 Coronavirus Tweets.CSV	189,412 KB
2020-04-04 Coronavirus Tweets.CSV	162,010 KB
2020-04-05 Coronavirus Tweets.CSV	158,178 KB
2020-04-06 Coronavirus Tweets.CSV	205,846 KB
2020-04-08 Coronavirus Tweets.CSV	185,464 KB
2020-04-09 Coronavirus Tweets.CSV	169,416 KB
2020-04-10 Coronavirus Tweets.CSV	152,229 KB
2020-04-11 Coronavirus Tweets.CSV	115,286 KB
2020-04-12 Coronavirus Tweets.CSV	90,263 KB
2020-04-13 Coronavirus Tweets.CSV	87,025 KB
2020-04-14 Coronavirus Tweets.CSV	159,867 KB
2020-04-16 Coronavirus Tweets.CSV	194,240 KB
2020-04-17 Coronavirus Tweets.CSV	181,944 KB
2020-04-18 Coronavirus Tweets.CSV	134,571 KB
2020-04-19 Coronavirus Tweets.CSV	120,619 KB
2020-04-20 Coronavirus Tweets.CSV	147,660 KB
2020-04-21 Coronavirus Tweets.CSV	150,119 KB
2020-04-22 Coronavirus Tweets.CSV	149,129 KB
2020-04-23 Coronavirus Tweets.CSV	140,497 KB
2020-04-24 Coronavirus Tweets.CSV	154,604 KB
2020-04-25 Coronavirus Tweets.CSV	114,756 KB
2020-04-26 Coronavirus Tweets.CSV	110,532 KB
2020-04-27 Coronavirus Tweets.CSV	142,502 KB
2020-04-28 Coronavirus Tweets.CSV	127,750 KB
2020-04-29 Coronavirus Tweets.CSV	140,805 KB
2020-04-30 Coronavirus Tweets.CSV	128,251 KB

The original dataset



Language

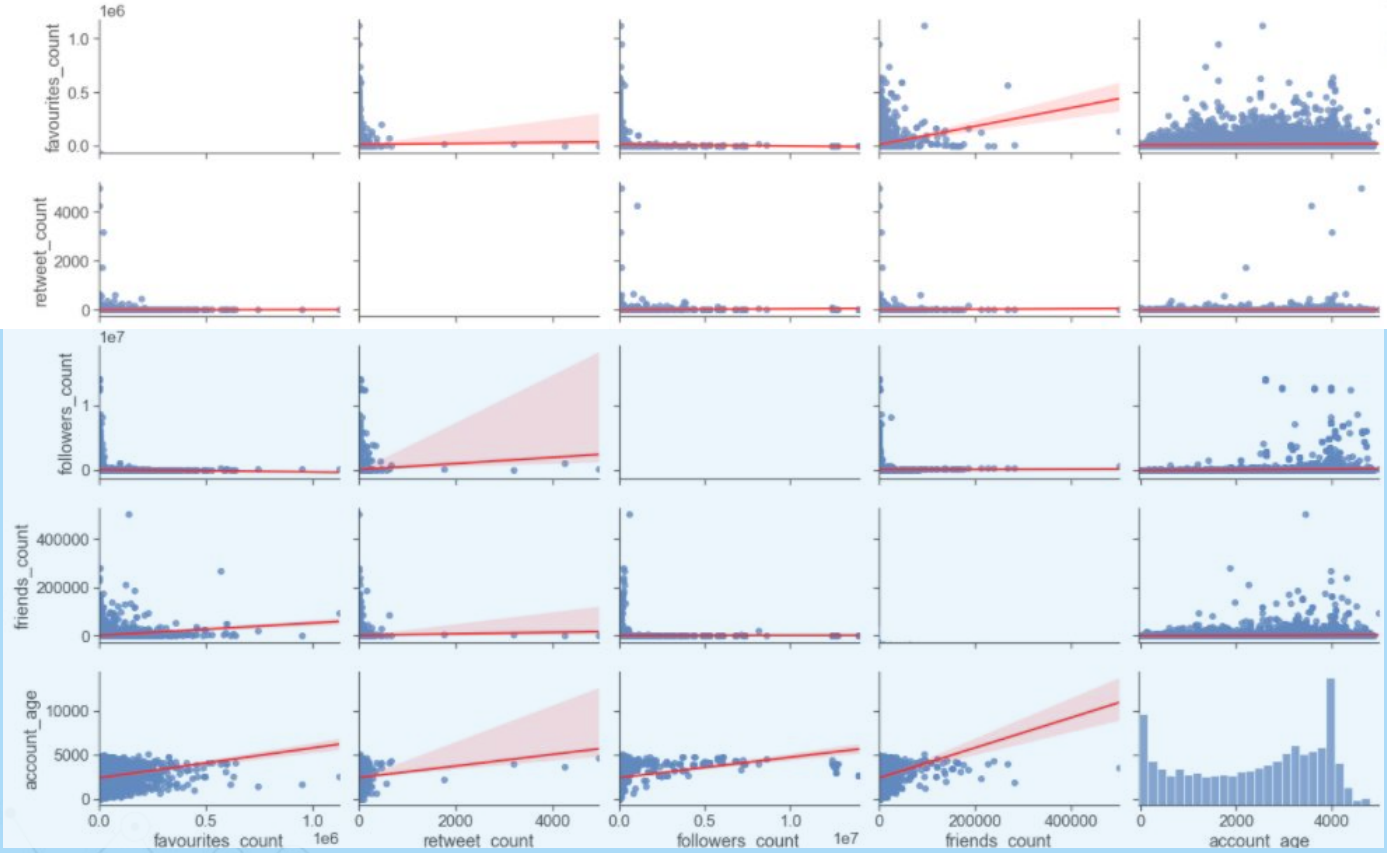
The majority of the tweets are in English. Spanish comes in second with the rest trailing behind.

The original dataset

Language

Once we isolate english language tweets for analysis and drop unused columns, we drop down to a little over 800,000 rows of data.

	user_id	favourites_count	retweet_count	followers_count	friends_count
count	8.187250e+05	8.187250e+05	818725.000000	8.187250e+05	8.187250e+05
mean	3.333511e+17	1.323988e+04	3.588336	5.730710e+04	2.273526e+03
std	4.912573e+17	3.982695e+04	82.064074	5.894429e+05	1.156747e+04
min	2.650000e+02	0.000000e+00	0.000000	0.000000e+00	0.000000e+00
25%	1.377155e+08	2.550000e+02	0.000000	1.410000e+02	1.580000e+02
50%	1.151877e+09	1.877000e+03	0.000000	7.870000e+02	5.480000e+02
75%	8.779354e+17	9.266000e+03	1.000000	4.072000e+03	1.658000e+03
max	1.250570e+18	1.989070e+06	26508.000000	8.121826e+07	1.496242e+06



Correlations

We do a scatterplot with a simple regression line

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1$$

to see how well correlated each column is to the other.

We see a **correlation** between friends_count and favourites_count.

What's interesting is there is **no correlation** between friends_count and retweets_count.

Also interesting is there's **no correlation** between followers_count and favourites_count or retweets_count.

Some Interesting Insights

What we see is friends will most likely favorite a tweet, but not necessarily retweet. Followers, on the other hand, are less likely to favorite a tweet but more likely to retweet than friends.

These are insightful, however we need to delve into the text column in order to answer some of the questions we started with.

Most Retweeted in Descending Order

screen_name	text
AdamMilstein	#coronavirus News Alert: Dr. Vladimir Zelenko, a board-certified family practitioner in NY, has now treated 699 Covid-19 patients with 100% success using Hydroxychloroquine Sulfate, Zinc and Z-Pak. All symptoms of shortness of breath resolved within 4-6 hr https://t.co/siCvNg845Q https://t.co/lOrUvBHp0A
BenjAlvarez1	This is how Angela Merkel explained the effect of a higher #covid19 infection rate on the country's health system.\n\nThis part of today's press conf was great, so I just added English subtitels for all non-German speakers. #flattenthecurve https://t.co/VzBLdh16kR
DineshDSouza	Something doesn't add up. The global panic and extreme action seem massively disproportionate to what we know about #CoronaVirus (for instance, its mortality rate). So either we are overreacting to a ridiculous degree, or they are not telling us something
eileenguo	#Wuhan residents estimate, based on calculations of cremations and urns now being returned to families, that between 42k-46k (!!) died in city + surrounding areas in the 2.5 months of lockdown. Far more than official figure of 2535 deaths. #COVID19 \n\nhttps://t.co/L1OsFv0VEf
gautam_adani	ADANI FOUNDATION is humbled to contribute Rs. 100 Cr to the #PMcaresfund in this hour of India's battle against #COVID19. ADANI GROUP will further contribute additional resources to support the GOVERNMENTS and FELLOW CITIZENS in these testing times.
marcorubio	Some in our media can't contain their glee & delight in reporting that the U.S. has more #CoronaVirus cases than #China\n\nBeyond being grotesque,its bad journalism\n\nWe have NO IDEA how many cases China really has but without any doubt its significantly more than why they admit to
RealCandaceO	The number one killer in America is Heart disease. 1,002 people a day. \n\nDid you know that if you die from heart disease right now, and they determine you to be an asymptomatic carrier of Covid-19 in your post-Mortem, they legally add your death to the #Coronavirus death toll?
LindseyGrahamSC	If it were up to me the whole world would send China a bill for the #CoronavirusPandemic. \n\nThis is the third pandemic to come out of China.

Most Favorited in Descending Order

favourites_count	screen_name	text
1995152	ChelseaAMusic	It's time @SmithfieldFoods took responsibility for what they did & how they didn't protect their workers when the first case of #COVID19 came out in their factory like so not right at all!!! 😡😷 https://t.co/8YWBA2HPX2
1562269	David_Leavitt	The @WhiteHouse Gift Shop is selling "World vs. Coronavirus" coin for $100...and the description says there's only 1,000 but somehow I feel like this is another @realDonaldTrump scam to cash in on #COVID19 \n\nSurprised it's not a "Trump vs. Coronavirus" coin tbh https://t.co/FTEKXaEKBB
1422809	MiguelCalabria3	People under lockdown are showing their gratitude to front-line healthcare workers worldwide by applauding them.\n\n#coronavirus #COVID19\n#QuedateEnCasa #StayAtHome #RestezChezVous #sanitarios @famartinez2001 https://t.co/FvYzuvv1V1
1311806	littlebytesnews	Terrible, may she RIP. Hopefully this doesn't become a trend and medical professionals get the mental healthcare they need to overcome so much suffering and death.\n#Coronavirus: Top NYC doctor takes her own life https://t.co/YLgP04CxxM https://t.co/DnTxwJWksQ
1266919	SueRMichael	This is a wonderful story of surviving #COVID19 and of hope https://t.co/2fhFVPV8BV
1258128	hazelglasgow	"Coronavirus: WHO warns 'the worst is yet ahead of us' in outbreak" #Coronavirus https://t.co/mdtDhIJlvq
1251292	madanabhat	Listen to the most recent episode of my podcast: Digital democracy as a consequence of the #Coronavirus crisis https://t.co/116Y5G3UWK
1186068	amor_vuelveTX	@MissClioMurray @come_for_t @setzacat @SamusAran2020 @LauraEastlick1 We're waiting for one tonight. We're far down south, not even #Coronavirus #COVID19 won't come around mol #BabyYvette lubz frenz https://t.co/xE0Djuh7un
1131771	fahma311	@OntHospitalAssn Stay safe heroes #StayHomeStaySafe #FlattenTheCuve
1126530	paoloigna1	Continueremo a parlarne per esigenze Geopolitiche ed elettorali di #Trump #covid19:BBC News - #Coronavirus: US intelligence debunks theory it was 'manmade'\nhttps://t.co/dDZVtm0SPg
1065477	ben10dinosaur	That concludes the One World: #TogetherAtHome benefit concert! Thanks to @jimmykimmel, @jimmyfallon and @StephenAtHome for hosting! And thanks to the artists for their great performances! And most of all, thanks to the healthcare workers in the front lines against #COVID19!
1059413	geoffrey_payne	gee! did you compliment the President on the US #COVID19 death toll? #Auspol #Covid19usa https://t.co/Ug7HpOPdiA
1048085	Solutioneer72	@RonaldKlain ARBs* - esp valsartan/sacubitril\n\nvs ARDS** caused by #COVID19 #SARSCoV2\n\n(Q: what abt losartan (reduces clotting)?)\n\n*Angiotensin Receptor Blockers\n\n**Acute Respiratory Distress Syndrome \n.\n.\nhttps://t.co/4SZC2AbOHD

Most Retweeted

The most retweeted news is about the positive results of Hydroxychloroquine Sulfate, Zinc and Z-Pak on Covid19 patients. This was early on in the year and everyone was hoping for a quick solution to the virus. There was some blame for China's initial cover-up. Some doubted the seriousness of the virus while others reported deaths from the frontline. The 2nd most retweeted tweet was Angela Merkel's explaining the importance of flattening the curve. Then there were concerns about the economic effects of going into a shutdown.

Most Favorited

The most favorited tweets range from gratitude for front line workers, survival stories, political criticism, etc … There were some requests for and discussions about medical information about Covid19 treatment, etc …

This is what we see so far. Let's prep our data for some Natural Language Processing to delve deeper into the dataset.

Selecting a random subset of our data

```
# Selecting a random fraction sample of the full dataframe
trunc_df = truncated_df_eng.sample(frac=0.015)
trunc_df = trunc_df.drop_duplicates(subset = ["text"], keep='last')
```

```
trunc_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 109257 entries, 2852980 to 7274277
Data columns (total 12 columns):
 #   Column              Non-Null Count   Dtype
---  ------              --------------   -----
 0   user_id             109257 non-null  int64
 1   created_at          109257 non-null  datetime64[ns, UTC]
 2   screen_name         109257 non-null  object
 3   text                109257 non-null  object
 4   is_quote            109257 non-null  bool
 5   is_retweet          109257 non-null  bool
 6   favourites_count    109257 non-null  int64
 7   retweet_count       109257 non-null  int64
 8   followers_count     109257 non-null  int64
 9   friends_count       109257 non-null  int64
 10  account_created_at  109257 non-null  datetime64[ns, UTC]
 11  account_age         109257 non-null  int64
dtypes: bool(2), datetime64[ns, UTC](2), int64(6), object(2)
memory usage: 9.4+ MB
```

Random subset

We reduce our original data of 7.3 million rows to just over 100,000 rows by picking a random fractional sample in order to work with our available resources.

2. & 3.

Processing our Data

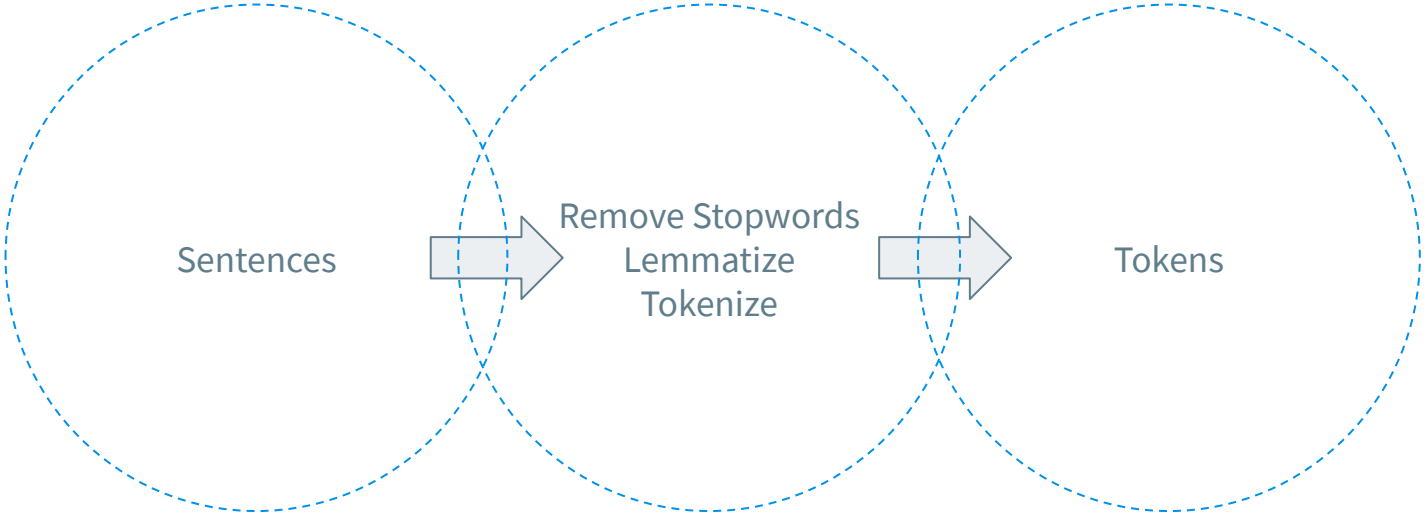
Clean, lemmatize, tokenize and vectorize our text

Processing our Data

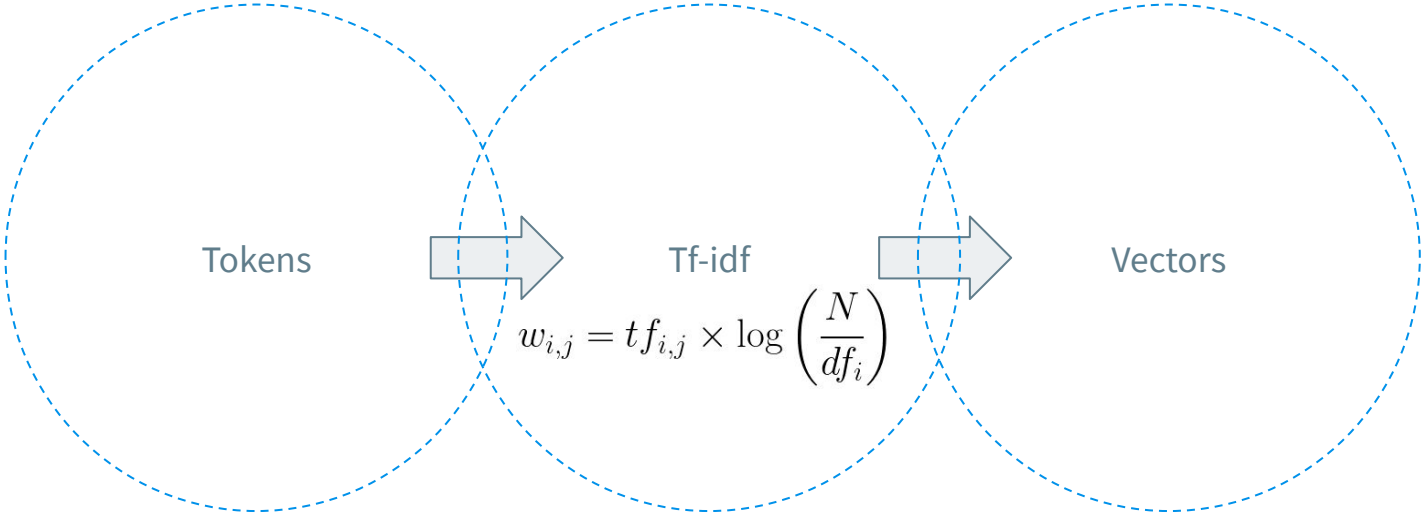
To begin machine learning on our data we will need to process our texts by turning our sentences into tokens and our tokens into vectorized data.

We will see a breakdown of the process and the outcome in the following slides.

Clean, Lemmatize, Tokenize our Data



Vectorize our Data



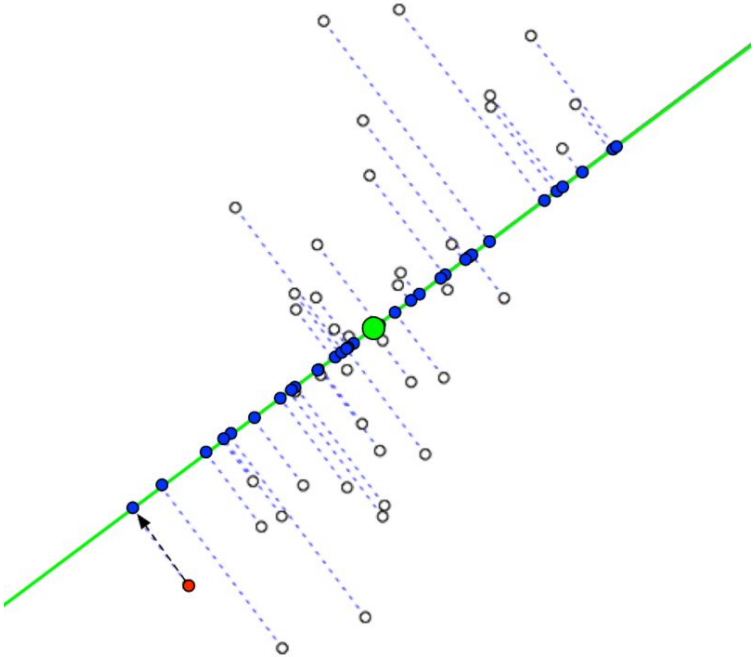
Side by Side Comparison of the Processed Text

	text	processed_text	vectors
72549	First day at work. Working remotely until further notice. Tweets and email only. Payday #covid19 dependent. The 2020's is reflected as things were 100 years ago. Someone called me a 'wartime' creative director as well. I'm doing it. The socially distant 'walking alive' as well.	day work working remotely far notice tweets email payday covid19 dependent 2020 reflect thing 100 year ago wartime creative director good socially distant walk alive good	(0, 982)\t0.2992287366392971\n (0, 396)\t0.39476660817347137\n (0, 268)\t0.3083281060127434\n (0, 57)\t0.2893424282475256\n (0, 1017)\t0.23126559602693025\n (0, 4)\t0.2720243872709879\n (0, 914)\t0.23172305388597783\n (0, 18)\t0.22991965839362785\n (0, 219)\t0.07019823428041078\n (0, 306)\t0.30880926676531184\n (0, 606)\t0.3291098020759458\n (0, 349)\t0.2461658539950927\n (0, 100...
87824	@mattkatz00 will you #AskGovernorMurphy if he will #FreeThemAllForPublicHealth during #COVID19 crisis? Being undocumented is not punishable by death. @GovMurphy #FreeThemAll before NJ jails holding ICE detainees become death camps	@mattkatz00 askgovernormurphy freethemallforpublichealth covid19 crisis undocumented punishable death @govmurphy freethemall nj jail hold ice detainee death camp	(0, 982)\t0.2992287366392971\n (0, 396)\t0.39476660817347137\n (0, 268)\t0.3083281060127434\n (0, 57)\t0.2893424282475256\n (0, 1017)\t0.23126559602693025\n (0, 4)\t0.2720243872709879\n (0, 914)\t0.23172305388597783\n (0, 18)\t0.22991965839362785\n (0, 219)\t0.07019823428041078\n (0, 306)\t0.30880926676531184\n (0, 606)\t0.3291098020759458\n (0, 349)\t0.2461658539950927\n (0, 100...
56832	You knew. #TrumpLiesAmericansDie #BorisLiesBritsDie #coronavirus \nhttps://t.co/FuVbiWw0QP	know trumpliesamericansdie borisliesbritsdie coronavirus https://t.co/fuvbiww0qp	(0, 982)\t0.2992287366392971\n (0, 396)\t0.39476660817347137\n (0, 268)\t0.3083281060127434\n (0, 57)\t0.2893424282475256\n (0, 1017)\t0.23126559602693025\n (0, 4)\t0.2720243872709879\n (0, 914)\t0.23172305388597783\n (0, 18)\t0.22991965839362785\n (0, 219)\t0.07019823428041078\n (0, 306)\t0.30880926676531184\n (0, 606)\t0.3291098020759458\n (0, 349)\t0.2461658539950927\n (0, 100...
59387	Preparedness\n#Comic #Comics #Ink #Selfie #Selfies #Copic #Copicart #GilbertComics #Dog #Dogs #Pets #CoronaVirus #Covid_19 #Masks #N95 #DarthVader #StarWars https://t.co/bpu4gUDIXX	preparedness comic comics ink selfie selfies copic copicart gilbertcomics dog dogs pets coronavirus covid_19 masks n95 darthvader starwars https://t.co/bpu4gudixx	(0, 982)\t0.2992287366392971\n (0, 396)\t0.39476660817347137\n (0, 268)\t0.3083281060127434\n (0, 57)\t0.2893424282475256\n (0, 1017)\t0.23126559602693025\n (0, 4)\t0.2720243872709879\n (0, 914)\t0.23172305388597783\n (0, 18)\t0.22991965839362785\n (0, 219)\t0.07019823428041078\n (0, 306)\t0.30880926676531184\n (0, 606)\t0.3291098020759458\n (0, 349)\t0.2461658539950927\n (0, 100...

Reducing Dimensionality

While vectorizing our data is necessary, it also increases the dimensions of our data.

We went from 12 variables to 2,048 variables for each row of data. We reduce our data using Principal Component Analysis.



109,257 x 12

Preprocessed Data

109,257 x 2,048

Processed Data

109,257 x 1,735

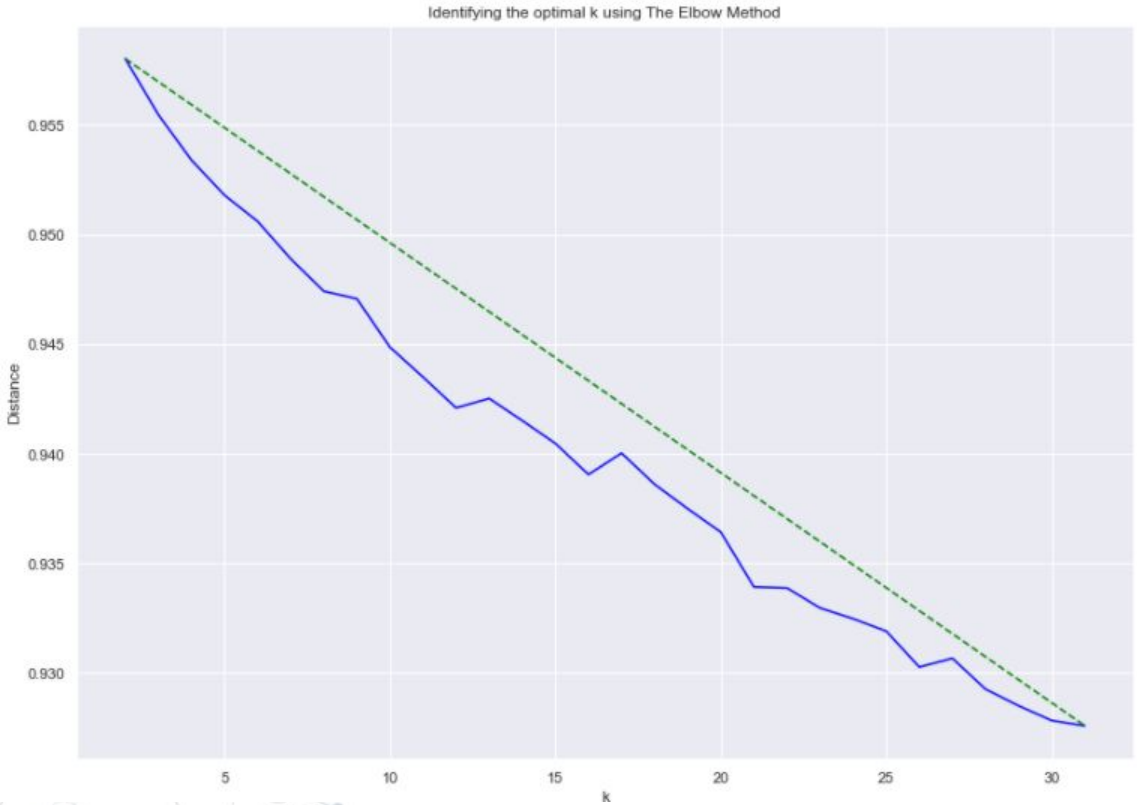
Applying Dimension Reduction

4.

Mapping Topic Groupings

Using KMeans to Discover Topical Clusters in our Texts

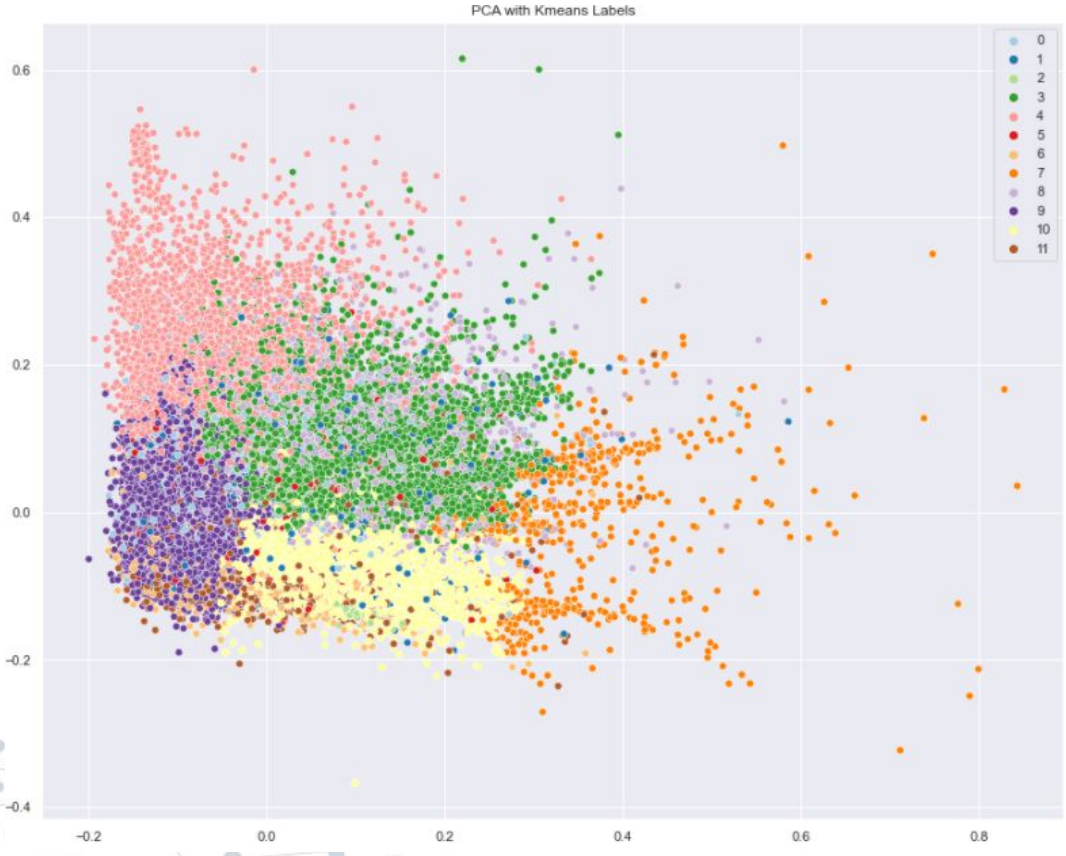
Finding Optimal Topic Groupings



Optimal Groupings

To find the optimal number of topic groupings we use the Elbow Method. The optimal k or grouping is at the turn elbow begins to turn or diminish. For our data it happens around 12 topic groups.

Visualizing the Topic Groups



Topic Groups

Although there are some overlaps, we can see some good topic clusters.

The Precision, Recall and Accuracy of our model is pretty high. Which gives us confidence in our model's performance.

```
Precision: 0.95
Recall: 0.95
Accuracy: 0.987
```

5.

Extract Important Topic Keywords

Using Latent Dirichlet Allocation

5.

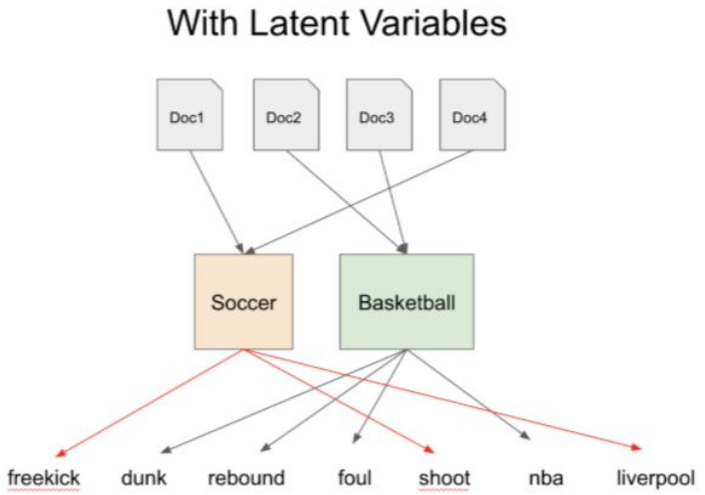
Extract Important Topic Keywords

Using Latent Dirichlet Allocation

Extracting Important Topic Keywords

We use Latent Dirichlet Allocation (LDA) to find important topic keywords for each cluster. LDA works by creating a latent layer (unsupervised topics) bridging between documents and tokens.

Let’s see how well it does.



Topic 0: *america, usa, gop, country, administration, election, donaldtrump, obama, biden, china, maga, cnn*

We see the political wheels already turning even in April. The early seeds of the red & blue war that will continue until the beginning of 2021.

Topic 1: *lockdown, extend, stayhome, quarantinelife, stay safe, face, time, online, end, lift, market, business, government, world*

Mid-March was the beginning of the covid lockdown so this was a natural topic of debate around that time. How long will it extend to? How will it affect businesses and markets? How will it end?

Topic 2: *usps deliver copy,usps deliver copy official, sign support, act deliver, sign act deliver, enforce mask usage*

USPS was a big topic back then. It was in deep financial trouble and Trump refused to bailout the Postal Service. Delivery was also a big topic of discussion since stay home policies were mandated, everything relied on delivery - food, goods, medicine, masks and even ballots - many became a huge arena for debate.

Topic 3: *outbreak, spread, crisis, pandemic, quarantine, coronavirus outbreak, china, government, india*

This was an obvious topic - the word pandemic was avoided as long as possible in describing coronavirus - until early March when we saw the spread of virus reach crisis levels within a few short months.

Topic 4: *positive, rate, case, record, number, death, surpass, rise, test positive, total, test, infect, population, jump, million*

One of the hot topics is the escalation of covid cases and the rising numbers of infection and death. This was obviously an area of great concern.

Topic 5: *economy, business, market, debt, student, adapt, long term, shrink, gdp, risk, long term*

Another topic is the economy, the market, the shrinking gdp and concern over the long term economic effects

Topic 6: *stay safe, order, stay home, socialdistancing, mask, away, inside, stayhomesavlives, kind, care*

Social distancing and self isolation became the new normal. Along the same line - stay home, save lives, stay safe.

Topic 7: *listen, great artist, discover, read, music, coronalockdown, watch, listen rotation, click link, time*

With a lot of time spent indoors people turn to online entertainment and discovering new virtual hobbies

Topic 8: *patient, positive, hospital, test, health, death, care, case, report, confirm, emergency, staff, relief*

There's definitely a focus on front line and health care workers with hospital resources being taxed beyond the limit with the increase in covid patients and deaths.

Topic 9: *test, health, china, world, realdonaldtrump, spread, die, million, food, vaccine, mask, action, government, quarantine, borisjohnson, national*

This list of keywords is a more generalized list about the spread and rise of covid and covid deaths. It seems to center on government steps or missteps in meeting the crisis, respectively receiving praise or blame for it.

Topic 10: *stayhomestaysafe, crisis, thank, response, learn, free, join, read, share, look, great, team, care, good*

This topic seems to center on encouraging one another to continue our resolve to fight, to encourage, to express gratitude as each one is doing their part in beating this global crisis.

Topic 11: *crisis, fight, time, need help, fund, resource, business, help people, local, covid help, want help, group, save, protect, healthcare, worker, relief*

This topic centers around helping and providing resources or funds to people and businesses in need. There was definitely a focus on essential workers and first responders who risked their lives in order to provide services to the rest of the world who stayed indoors. There was a rise in financial needs for the unemployed and for struggling businesses who needed the relief funds to survive.

Conclusion

We've explored a month's worth of data from twitter from the end of March to the end of April 2020. We used unsupervised learning techniques to perform natural language processing on a smaller portion of the entire set (there were originally 7.3 million rows of data, of which we randomly chose about 110,000 rows from). After we cleaned, tokenized and vectorized our text data, we decided on a few topic clustering methods to find out what important topics emerged.

Conclusion

Our original intent was to answer if we can discern what trends / sentiments / behaviors / challenges arose during our period of self-seclusion? Which would in turn could provide insight and lead to future policy decisions. The above 12 topics give a good snapshot of important topics of discussions that impacted the worldwide community during the lockdown.

Future Studies

We set out to find out what happens when the world goes into self isolation for a long period of time through natural language processing tweets. I think it's still a good area of study. The study was able to uncover some topic groupings. One weakness in this study is that the data only scraped tweets with the covid / coronavirus hashtag. This limits us to data around the subject of covid.

Future Studies

A better set of data is all tweets during the lockdown whether or not it was directly meant to relate to covid. We want to examine what trends emerged out of the extended time in lockdown and I think we may be able to get topics such as “exercised more”, “gained weight”, “painted my walls”, “where are my friends?”, etc because they aren’t necessarily covid related but lockdown related.

However, even with the limits of the data, in the end, the nlp still produced really nice results with topic modelling the data we have.