Infected Tweets:

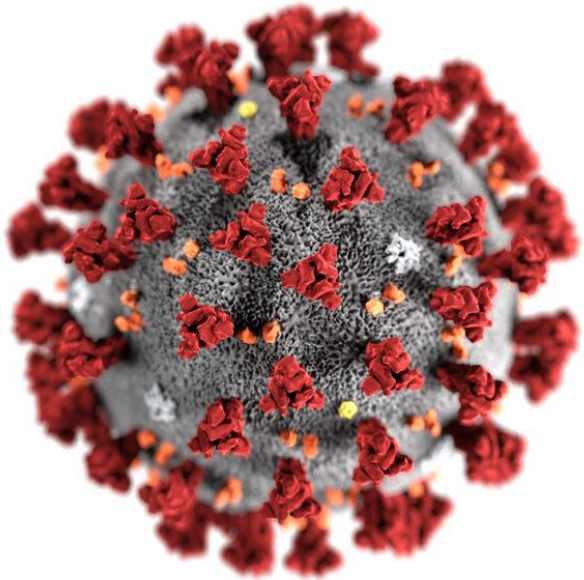
Unsupervised Modelling of Randomly-Sampled Tweets to Determine COVID-19 Contents

By Max Mazel, Najiha Boosra, and Adam Cohen

Problem Overview

COVID-19 Quick Facts:

- Economic and social disruptor
- Mass quarantines and isolation
- Global in scope
- Clustered mainly in major population centers



COVID-19 virus. Source: Statsnews.com

Problem Statement

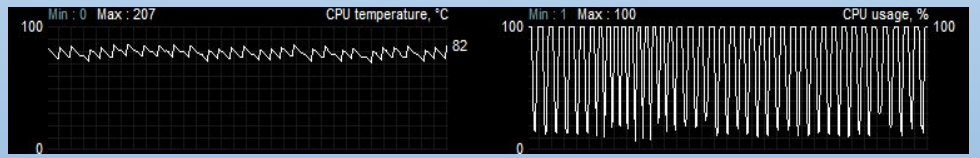
When the COVID-19 pandemic is running rampant, current information is paramount for staying safe. Graphs and models are created daily, producing figures relied on by hospitals and government agencies. Although this data is official, it does not tell the entire story of outbreak impact. On social media, specifically Twitter, people are talking about their experience with COVID-19, potentially indicating COVID impact sites before they become apparent medically.

As a team of data scientists consulting with New Light Technologies, we are tasked with creating a model that can dynamically classify Tweets as COVID-related. To solve this unsupervised learning problem we will train a w2v vectorizer on COVID-19 related tweets and use the weights of those words in a DBscan cluster. We will then check the Tweet clusters to assess relationships, and determine which clusters are our targets and where the Tweets originated from.

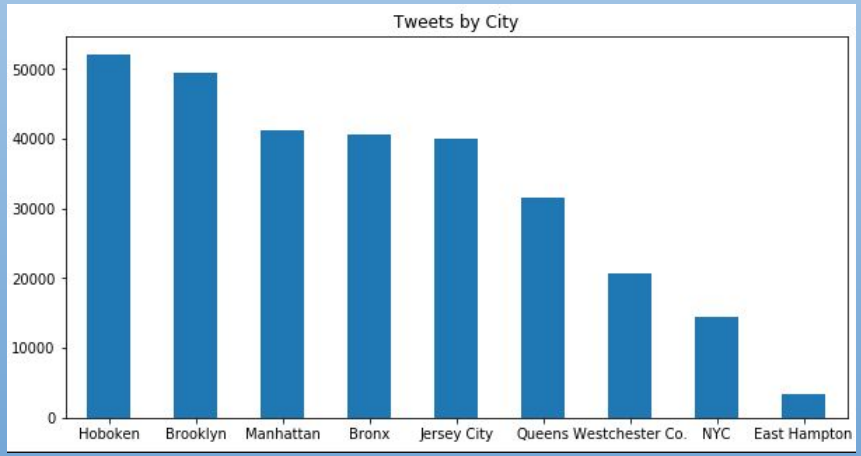
EDA/Cleaning

Data Collection and Preparation

- Used a custom wrapper of twitterscraper Python library.
- 100k biased tweets queried on targeted keywords
- 200k variance tweets queried on random keywords from RandomWordGenerator
- Wrapper was very inefficient
  - Bias set used 1,170 query calls
  - Variance set used 5,850 query calls



This is what peak server sadness looks like.



Data Collection and Preparation

- Initial bias set of 30k tweets generated using outside knowledge for COVID keywords.
- 30k biased tweets processed with Word2Vect to produce additional query terms.
- Intent was for EDA set to have high bias through careful targeting, while testing set would have high variance through maximum volume.

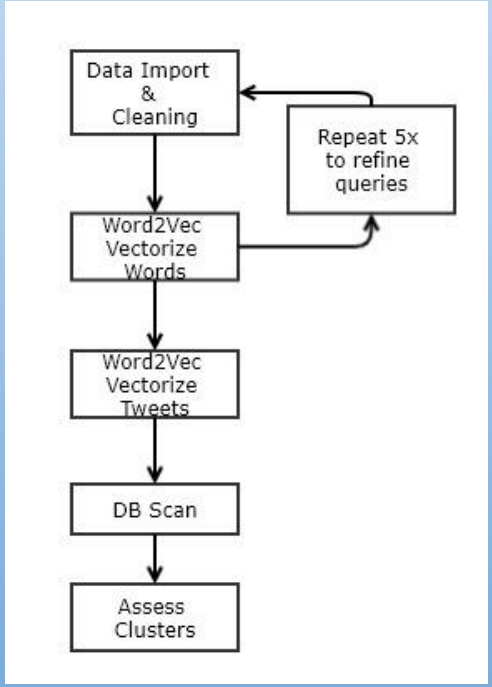
```
[('right', 0.09585417807102203),
 ('twitter', 0.08774752914905548),
 ('president', 0.07988424599170685),
 ('waiting', 0.07965778559446335),
 ('chinese', 0.07624496519565582),
 ('safety', 0.07341757416725159),
 ('able', 0.07280883193016052),
 ('helping', 0.06696482747793198),
 ('united', 0.06436911970376968),
 ('community', 0.06277808547019958)]
```

Words most similar to ‘covid’

Modeling

Model Workflow

Our model calculating
Twitter vectors



w2v: Word Vectorizer

Keyword: ‘Testing’

```
('dying', 0.1552795320749282 8),
('families', 0.13991212844848633)
('virus', 0.13581764698028564),
('word', 0.1346871256828308),
('cuomo', 0.126844942569732 67),
('covid19', 0.125726655125617 98),
('coronavirus', 0.120611436665058
('wait', 0.118201725184917 45),
('hit', 0.11507762223482132),
('hospital', 0.11272245645523071)
```

- Vectorized words in corpus
  - 300 dimensions
- Create “vocab”
  - Cosine similarity
  - Build query list
- Used to create Tweet Vectors

w2v: Tweet Vectorizer

- Assigns weight of vectorized words to tweets
  - Loops through each word in Tweets
- COVID-related Tweets are assessed
- Feeds into clustering algorithm
  - DBscan
  - Kmeans

DBscan

0	73453
-1	33798
2	80
1	39
3	33
4	29

- Autonomously creates clusters
  - Epsilon score
  - Minimum size
- Excellent for large data sets
  - Analyze clusters after creation
  - Observe which have data that we need
- Parameter tuning
  - Smaller epsilon
    - More accurate clusters
    - More noise
  - Larger minimum size
    - Less Clusters

had to go to three diff convenient stores to find real milk not soy and they were all out of the lower priced non organic stuff i buy organic so i was fine with it but wow covid 19 panic buying begins in manhattan	[go, three, diff, convenient, stores, find, real, milk, soy, lower, priced, non, organic, stuff, buy, organic, fine, wow, covid, 19, panic, buying, begins, manhattan]	0
covid coronovirius in nursing home setting kirkland more than 50 with symptoms 3120	[covid, coronovirius, nursing, home, setting, kirkland, 50, symptoms, 3120]	0
from kofi anan epidemics like covid19 are problems without passports tall walls and immigration officials will not keep them out funds4researchletscienceeducatelawmakerscc drriveramindt 6dm4cnyconnections emorycfar	[kofi, anan, epidemics, like, covid19, problems, without, passports, tall, walls, immigration, officials, keep, funds4researchletscienceeducatelawmakerscc, drriveramindt, 6dm4cnyconnections, emorycfar]	0

DBscan

- Autonomously creates clusters
  - Epsilon score
  - Minimum size
- Excellent for large data sets
  - Analyze clusters after creation
  - Observe which have data that we need
- Parameter tuning
  - Smaller epsilon
    - More accurate clusters
    - More noise
  - Larger minimum size
    - Less Clusters

0	242166
-1	50734
10	84
20	49
9	39
16	35
14	30
5	28
8	27

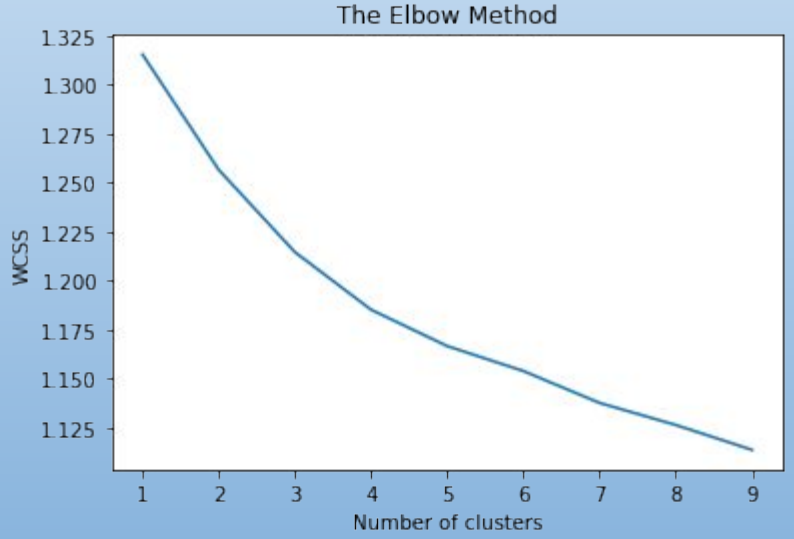
other	20925
test	17807
#COVID-19	12877
#covid19	11895
hot	11787
student	11162
medical	9893
sale	9578
round	8258
chocolate	8197

[looks, like, adios, muchachos, us, covid, covid19us, coronavirus]	0
[humoursnlon, covid, coronavirususa]	0
[covid, coronovirius, nursing, home, setting, kirkland, 50, symptoms, 3120]	0
[kofi, anan, epidemics, like, covid19, problems, without, passports, tall, walls, immigration, officials, keep, funds4researchletscienceeducatelawmakerscc, drriveramindt, 6dm4cnyconnections, emorycfar]	0
[un, health, agency, warns, coronavirus, covid19, criminal, scams]	0

K-Means & Elbow Method

- Calculate SSE for Values of k
- Use ‘Elbow method’
- K-means algorithm to predict the labels

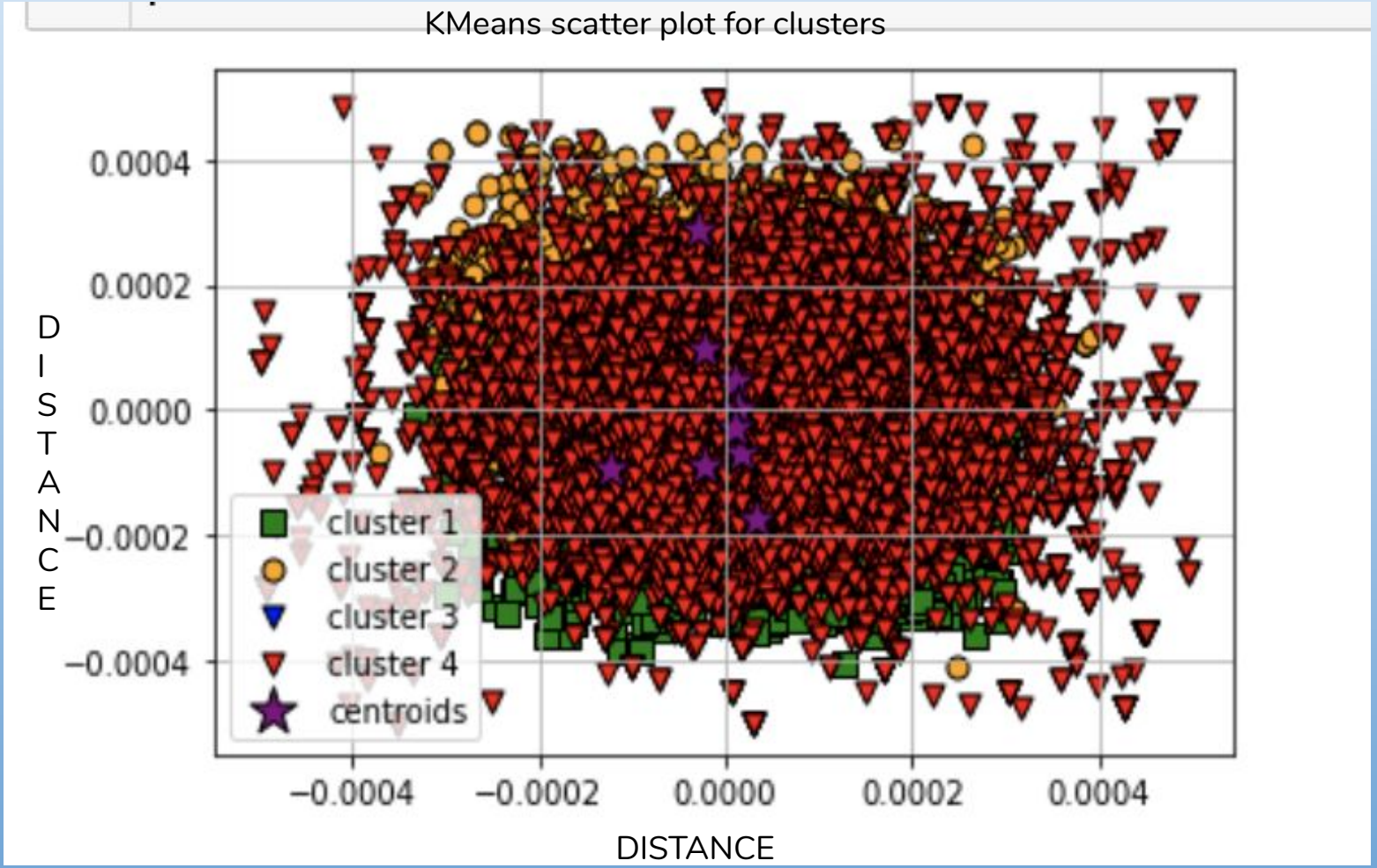
```
panic', 'buying', 'begins', 'manhattan']
3:['covid', 'coronovirius', 'nursing', 'home', 'setting', 'kirkland', '50', 'symptoms',
'3120']
1:['kofi', 'anan', 'epidemics', 'like', 'covid19', 'problems', 'without', 'passports', '
tall', 'walls', 'immigration', 'officials', 'keep', 'funds4researchletscienceeducatelawm
akerscc', 'drriveramindt', '6dm4cnyconnections', 'emorycfar']
3:['nurse', 'encouraging', 'people', 'protect', 'especially', 'immune', 'issues', 'inclu
des', 'wearing', 'mask', 'thorough', 'hand', 'face', 'washing', 'avoiding', 'crowds', 's
praying', 'wiping', 'surfaces', 'alcohol', 'chlorox', 'products', 'getting', 'flu', 'sho
t', 'covidcoronavirus']
3:['trump', 'saying', 'covid', 'covid19us', 'hoax', 'instead', 'activity', 'public', 'he
alth', 'protocols', 'something', 'preventable', 'may', 'veering', 'maga', 'pandemic']
1:['un', 'health', 'agency', 'warns', 'coronavirus', 'covid19', 'criminal', 'scams']
3:['alcohol', 'kills', 'covid']
```



Elbow in the graph is optimal k value, where

k = 4

Scatter Plot for Clusters



DBscan

Model Selected

Model Evaluation

- DBscan accounts for high dimensionality
  - Word vector has 300 dimensions
- Epsilon selection: .4
  - Bias Data vs Complete Data
  - 23 clusters created
- Multiple COVID clusters
  - 242,166 of COVID concern tweets
  - 50,734 “noise” tweets
  - 577 “spam” tweets across 21 clusters
- Identified Spam Tweets in Dataset

Infected Tweets Identified!

Conclusion and Recommendations

- Social media posts can be used to determine whether or not COVID is in an area
- This unsupervised data set can be clustered and assessed
  - Word2Vec Cosine Similar words
  - Weighted Tweet Vectorizer
  - DBscan autonomous clustering
- Using this process we can:
  - Separate COVID tweets from any area
  - Determine impact of outbreak on community
    - Concentration of COVID tweets
  - Heatmap the locations of virus hotspots
- Make a world a better place!

Future Improvements

- Inherent issues with scrape system:
  - Library doesn't use Twitter API, so less control over data pulled.
  - Intense CPU resource requirements
  - Queries must be run in sequence.
- Inherent issues with model system
  - Twitter subject to many repeat spam posts.
  - Model parameters lack scalability. Large datasets performed poorly on small-dataset parameters.
  - DBScan very resource-hungry, especially at large scales, and doesn't natively support CUDA.

Questions?

References

https://www.statnews.com/2020/02/11/disease-caused-by-the-novel-coronavirus-has-name-covid-19/

Word2Vec Image: https://hackernoon.com/word2vec-part-1-fe2ec6514d70

Black Twitter bird Image: https://webstockreview.net/explore/twitter-bird-png-transparent/