Classifying Tweets

Erik Paulson

The Data

- https://www.kaggle.com/datasets/datatattle/covid-19-nlp-text-classification?select=Corona_NLP_train.csv
- Approximately 45k tweets regarding COVID-19
- The only column of interest for this project was the “OriginalTweet” column

The Process

Original Tweet:

"When I couldn't find hand sanitizer at Fred Meyer, I turned to #Amazon. But $114.97 for a 2 pack of Purell??!!Check out how #coronavirus concerns are driving up prices. https://t.co/ygbipBflMY"

Processed Tweet:

couldn hand sanitizer fred meyer turned 114 97 2 pack purell check concerns driving prices

The Process

Converting the series of strings to tokens

Creating a token matrix of TF-IDF features to show how relevant each token is per tweet

[8]:

	000	10	19	20	2020	amid	amp	available	avoid	banks	...	water	way	week	weeks	went	work	workers	working	world	year
0	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	...	0.0	0.0000	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	...	0.0	0.0000	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	...	0.0	0.0000	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	0.0	0.0	0.000000	0.0	0.0	0.0	0.328878	0.0	0.0	0.0	...	0.0	0.0000	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	...	0.0	0.0000	0.423729	0.0	0.0	0.0	0.0	0.0	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
44950	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	...	0.0	0.0000	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0
44951	0.0	0.0	0.270908	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	...	0.0	0.5588	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0
44952	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	...	0.0	0.0000	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0
44953	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	...	0.0	0.0000	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0
44954	0.0	0.0	0.190835	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	...	0.0	0.0000	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0

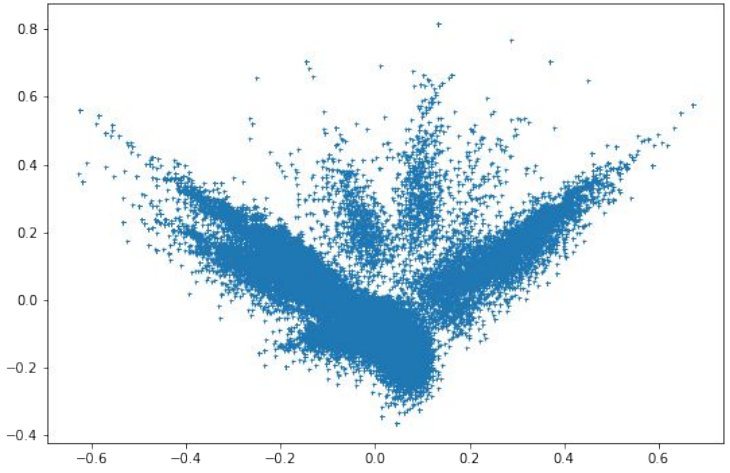
44955 rows × 191 columns

The Process

Plotted tweets after using PCA to reduce further to 2 dimensions:

- Lowering dimensions to 2 allows us to view plot in 2-D

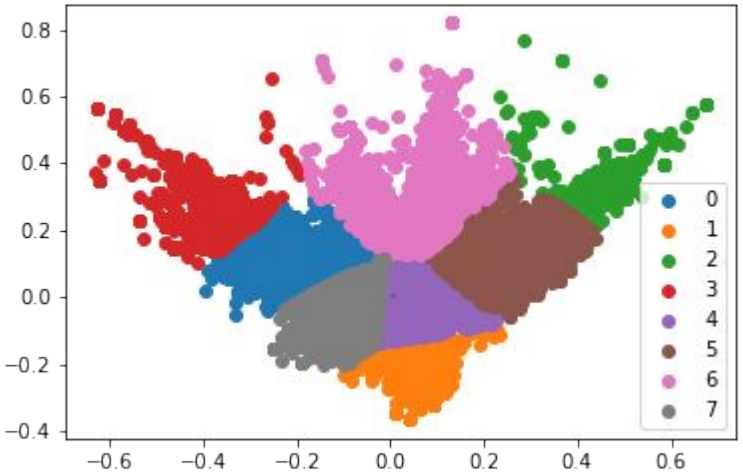
- This may give us an idea on how many clusters to use



The Process

Using K Means to group the tweets into clusters, each cluster will have a topic.

Chose 8 clusters based on Inertia Plot, and Silhouette Score (see appendix)



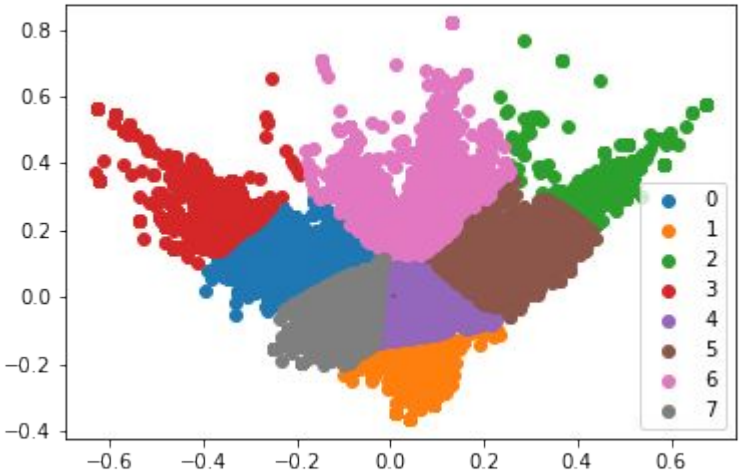
The Process

Analyzing each cluster's topic by checking most relevant words:

0: Consumer Demand + Shopping/ Grocery Prices

1: Supermarket Items, Panic Buying, Toilet Paper

2: Grocery Stores, Workers, People Going Out



The Process

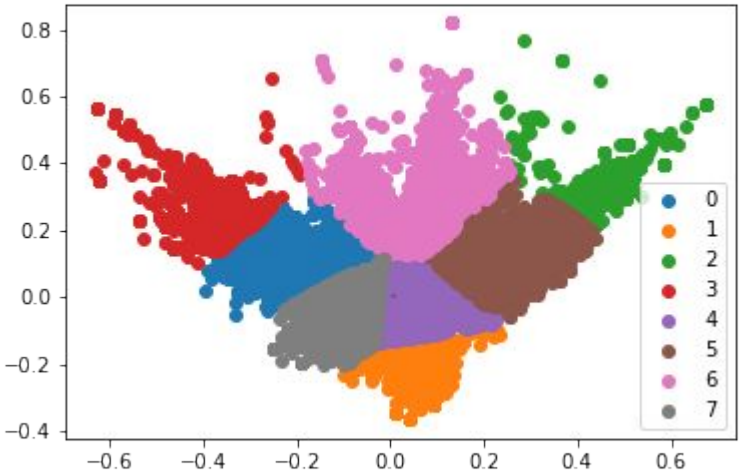
3: COVID-19, Consumers, Prices, Oil Crisis, Online Shopping

4: Online Shopping, Food, Need Help, People

5: Grocery Store, People, Going

6: COVID-19, Online Shopping, Retail, Grocery

7: Prices, Consumers, Oil, Demand, Pandemic



The takeaway

For me it seems like many tweets regarding covid have a similar “vibe”. Many people are concerned about topics such as shortages, prices, online vs in person grocery shopping, people, essential workers and use twitter to get those thoughts out into the open.

Conclusions

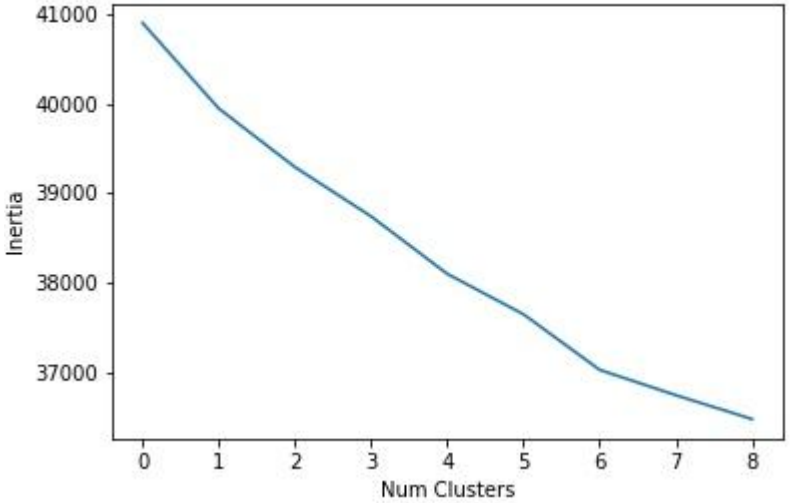
In this dataset there seemed to be some overlap in the clusters, many topics had some shared themes even though some were also very distinct.

K Means clustering worked the best for me, compared to DBSCAN or other clustering/partitioning method. K=8 provided a good amount of clusters so there may be some extra overlap, the silhouette score was above average but it is still close to 0 (meaning there are overlapping clusters)

Future Work

The next steps of this project will be utilizing n-grams (groups of words) to see if this will make analysis more easy and reduce overlapping clusters.

I focused on using the top percent of words from the tweets based on TFIDF values, I will move forward and increase that number slowly to see if it helps diversify clusters and reduce overlap.



Silhouette Scores:

0.027765541646708688, k=2
0.03107208854793987, k=3
0.035550483073850565, k=4
0.042147724994378806, k=5
0.04807748049678138, k=6
0.047102798537917046, k=7
0.05671003192559292, k=8
0.046863467125585714, k=9
0.047767514384750444, k=10

Inertia plot (checking densities of clusters)(sum of distances of points to centers)(lower=better)