Coronavirus Tweets Analysis

Xufei Li

Abstract

The goal of this project was to use unsupervised machine learning to analyze text data in order to gain insights. I started with analyzing March 2020 tweets by non-negative matrix factorization(NMF) topic modeling, then comparing topic scores among different countries to see each country's top discussions. I visualized the comparison with a side-by-side bar chart as well as Scattertext. Then, I used a clustering model(k-means) to group my tweets to 6 distinct groups based on similarity and explore the top topics for each group. Finally, I ingested the most recent data via Twitter API and analyzed the top discussions by NMF topic modeling.

Design

By getting top discussions, the government can know what people care more about and what can be a problem to take action on. By comparing different countries, we can know the difference/priority topic for each country. By clustering, we can know for each distinct group of the tweets, what topics are being involved, then targeting the characteristics of the group to see if there's anything that the government can do for this distinct group of people .

Data & Data Cleaning

I got the data from Kaggle-Coronavirus Tweets which includes 41,000+ tweets. I filtered my data to March 2020 English tweets which includes 21,000+ data points, then preprocessed data & tokenized by TfidfVectorizer. I also used the Twitter API to get 5000 most recent tweets for comparison.

Algorithms

Topic modeling - NMF(n_components=10) (chosen n by trying different number)
Clustering - KMeans(k=6) (chosen k by elbow method & silhouette analysis)

Tools

- Tweepy for connecting twitter API
- Numpy and Pandas for data manipulation
- Geopy(Nominatim) for geographic mapping & retrieve country name from messy 'Location' column
- Scikit-learn for tokenization & topic modeling
- Matplotlib and Seaborn for plotting

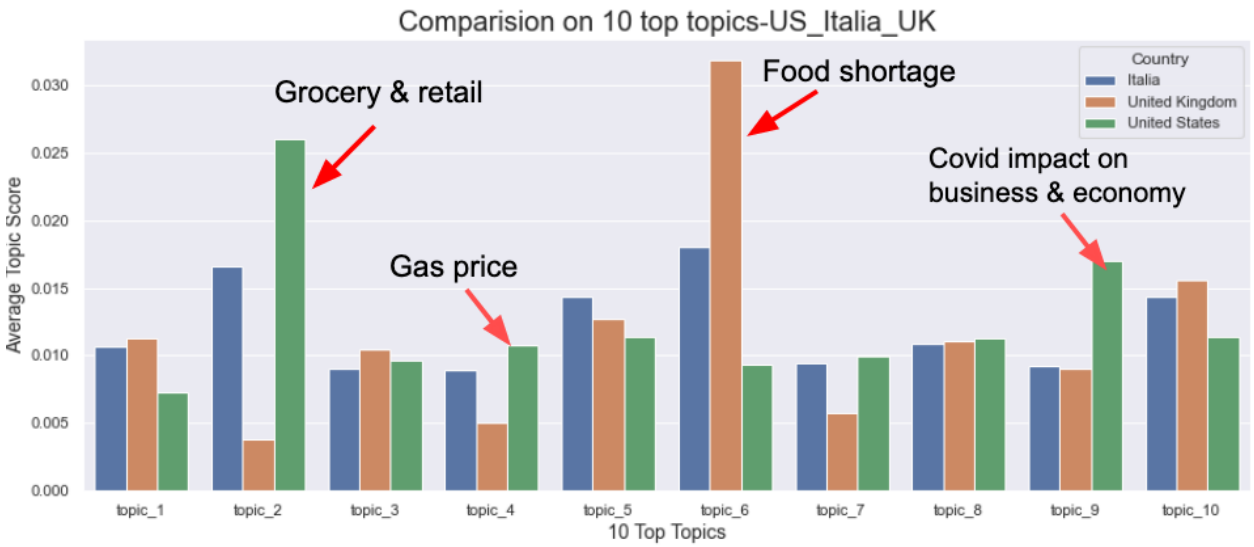
Communication

Conclusion for top 10 topic in March_2020:

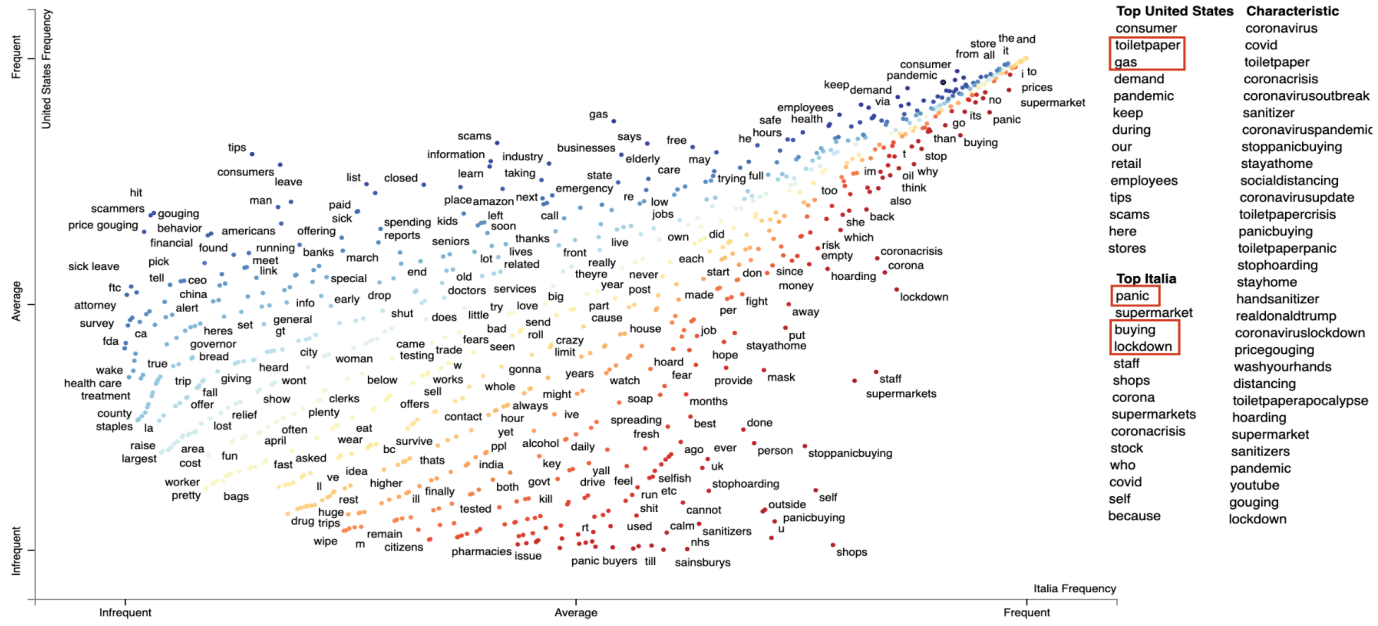
- Topic1: Panic buying & stocking food
- Topic2: Grocery & retail
- Topic3: Online shopping & delivery

- Topic4: Gas price decline
- Topic5: Toilet paper hoarding
- Topic6: Supermarket shelves are empty - Food shortage
- Topic7: Mask & Hand sanitizer shortage
- Topic8: Thanks ALL front-line workers
- Topic9: Covid impact on business & economy
- Topic10: StayAtHome

Visual - Comparison on 10 top topics-US_Italia_UK



Visual - Scattertext(USA vs. Italy)



USA vs. Italy

- 'panic'(39 vs. 99)
- 'tolietpaper'(71 vs. 36)
- Sharing words: 'jobs', 'testing'

Topics for May_26_2021:

- Vaccine - ‘appointment’, ‘second dose’, ‘side effects’
- News - 'LEAKED INTERNAL DOCS(fb)'
- News - ‘Tennessee anti-vaxxer arrested’
- News - 'COVID OUTBREAK IN MELBOURNE, Australia'

Future work:

1 day’s data can not represent the whole month's discussions. So, for the next step, I can get the whole month data for May 2020 then apply topic modeling to see what the top topics for the whole month are.