**NLP Project Write-up**

**Coronavirus Tweet (2020) Analysis**

**Abstract**

The goal of this project was to use unsupervised NLP analysis to answer questions like: what are the major topics people talk about on tweet? What is the hottest topic during a month in 2020? What is people attitude (positive or negative) towards certain topic?

**Design**

This project originates from the challenge: **Coronavirus tweets NLP - Text Classification.** The data is provided by [Kaggle](http://taarifa.org/). Tweets that are relative to coronavirus in 2020 are the analysis object. In this project, I used PorterStemmer and TFIDF to preprocess the corpus and used NMF to do topic modeling. Also, sentiment analysis is done by using VADER to show people’s attitude towards certain topic.

**Data**

Since this is an unsupervised learning project. I combined the test and train data to be my full data. The dataset contains 45k tweets with tweet time and tweet location. After using stemmer to stem the words and used TFIDF to transform corpus into doc-term matrix. The final terms left are 8300.

**Algorithms**

I first used TFIDF to derive a doc-term matrix for further analysis. Then I tried all LSA, NMF, LDA and select NMF as the final PCA algorithm as it returns me topics that makes sense most. The final 5 topics I find are:

* + Panic Food Buying
  + Social Distance
  + Online Food Delivery
  + Oil Price
  + Sanitary Product

Then I assigned each tweet with the most relative topic by topic weight from NMF analysis. And plot the topic number along month to get topic heat trend in 2020:

Chart, line chart

Description automatically generated

Finally, I used VADER to do sensitive analysis on every tweet. If the compound score is larger than 0.05, then it’s classified as ‘positive’. If less than -0.05, then ‘negative’. Otherwise it’s ‘neutral’. Final sentiment trend line is as below (topic sentiment breakdown could be found in my tableau page: https://public.tableau.com/app/profile/jasmine7670/viz/CoronaNLP-Tweetbytopic/Dashboard1?publish=yes)

Chart, line chart

Description automatically generated

**Tools**

* Numpy and Pandas for data manipulation
* Scikit-learn for modeling
* Nltk for NLP analysis
* vaderSentiment for sentiment analysis
* Tableau for interactive visualizations

**Communication**

A PowerPoint is made and presented to audience to show my key findings and visuals.