Report 3

17PT17 - Muthualagesan Suryavarathan
17PT28 - Sanjay Kumar S

Roadmap

1. About the Dataset
2. Preprocessing
3. Analytical Methods used
4. Code
5. Results
6. Future Work

About the Dataset

- Contains Tweets from 29th March to 30th April 2020.
- The dataset contains variables associated with Twitter: the text of various tweets and the accounts that tweeted them, the hashtags used and the locations of the accounts.
- There are around 1,67,000 tweets.
- Dropped user privacy sensitive columns such as location, followers etc from the table.

Data Cleaning and Pre processing

- Removal of emoticons and hyperlinks from tweets.
- Removal of punctuations
- Added new stopwords relevant to the problem
- Removed all the stop words and tokenized the tweets

Top Hashtags and Top Mentions

- The top hashtags and top mentions are found from the data based on their count.
- This result tells what hashtags was trending through the period chosen.
- It also tells who was more embedded in the network and had more influence.

Extracting the sentiments of these tweets

- TextBlob is a python library for Natural Language Processing (NLP).
- TextBlob actively used Natural Language ToolKit (NLTK) to achieve its tasks.
- NLTK is a library which gives an easy access to a lot of lexical resources
- Allows users to work with categorization, classification and many other tasks.
- Performed Sentiment Analysis on the tokenized data to obtain a polarity score ranging from [-1,1]
- +1 for most positive, 0 for neutral and -1 for most negative tweets.

Hashtags, Mentions and Sentiments

- Based on the sentiment polarity score, we extract the top 100 positive, neutral and negative hashtags and mentions.
- It is observed that the top 50 hashtags and mentions are common for all sentiments.
- Hence we choose the next top 50 (51-100) hashtags for further analysis.

T-SNE

- t-Distributed Stochastic Neighbor Embedding (t-SNE).
- Technique primarily used for data exploration and visualizing high-dimensional data.
- t-SNE gives you a feel or intuition of how the data is arranged in a high-dimensional space.

Visualize similarity through t-SNE

- Train a Word2Vec model.
- It is a more recent model that embeds words in a lower-dimensional vector space using a shallow neural network.
- For example, *strong* and *powerful* would be close together and *strong* and *Paris* would be relatively far.

Future Work

- Improve the results on Sentiment Analysis
- Social Network Analysis on the dataset

References

https://arxiv.org/pdf/1607.00534.pdf