#Virus Trends on Twitter Dataset using Topic Modelling

A Data Driven Design Project Report

Submitted by,

Balbir Singh

Radhika Khandelwal

Rashmi Nagpal

Work done under the guidance of Sébastien Foucaud

Table Of Content

Problem Statement

To analyze and find insights on emerging trends related to the virus.

The situation surrounding COVID19 is dynamic and rapidly changing, on a daily basis. It has created anxiety among people and businesses. This situation has caused governments to lockdown entire cities and even countries. Since most of the people can’t go out, a rise in social media use has been observed. People are expressing their views, thoughts and fears on social media. It is for this reason we have chosen to analyse the twitter feed to understand the current situation.

Introduction

As terrible as the currently unfolding Coronavirus epidemic has been, it’s been fascinating to observe how quickly communities at social platforms like Twitter, Facebook et al across the world have scrambled to start understanding this virus and its potential impact. Because of this, there are lots of interesting **tweets** coming out fast on Twitter henceforth, made us ponder upon to see if there were any discernible patterns in the topics and conclusions these tweets are discussing. So, we manually scraped the dataset from twitter and performed Topic Modelling, which is a process in which we try to uncover abstract themes or "topics" based on the underlying documents and words in a corpus of text.

Approach

From the input dataset, i.e. tweets scraped from the Twitter dataset we performed **data-preprocessing** techniques, as mentioned below

1. **Tokenization** - Segregation of the text into its individual constituent words.
2. **Stopwords** - Throw away any words that occur too frequently as its frequency of occurrence will not be useful in helping detecting relevant texts. (as an aside also consider throwing away words that occur very infrequently).
3. **Stemming** - combine variants of words into a single parent word that still conveys the same meaning
4. **Vectorization** - Converting text into vector format. One of the simplest is the famous bag-of-words approach, where you create a matrix (for each document or text in the corpus). In the simplest form, this matrix stores word frequencies (word counts) and is often referred to as vectorization of the raw text.

Once pre-processing was completed, we performed basic **Exploratory Data Analysis**, for example : finding top 50 word frequencies after preprocessing.

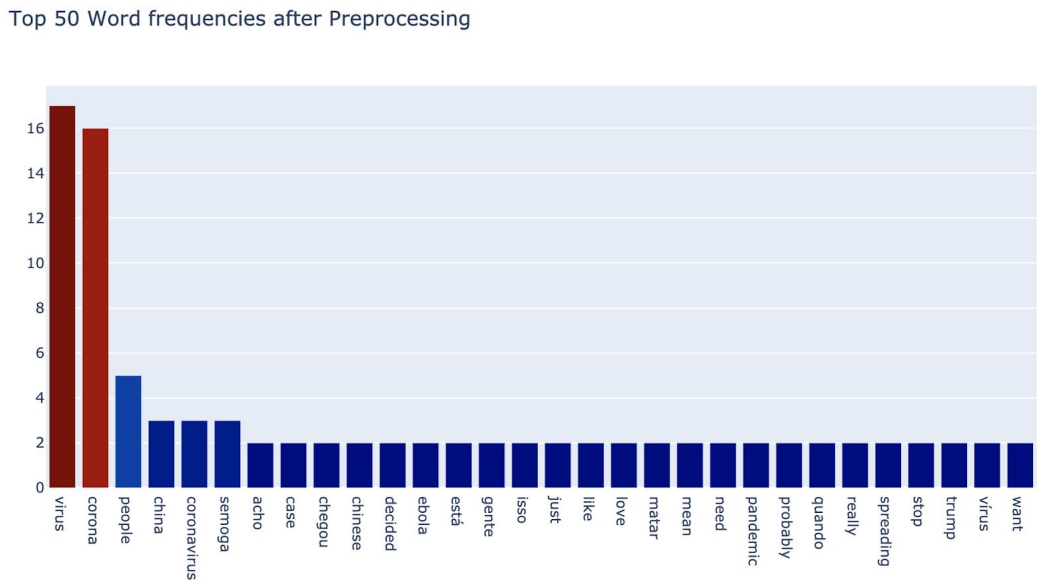

Figure 1. Top 50 words in the new dataset

Post that, we performed Topic Modelling approaches and created word clouds for the same.

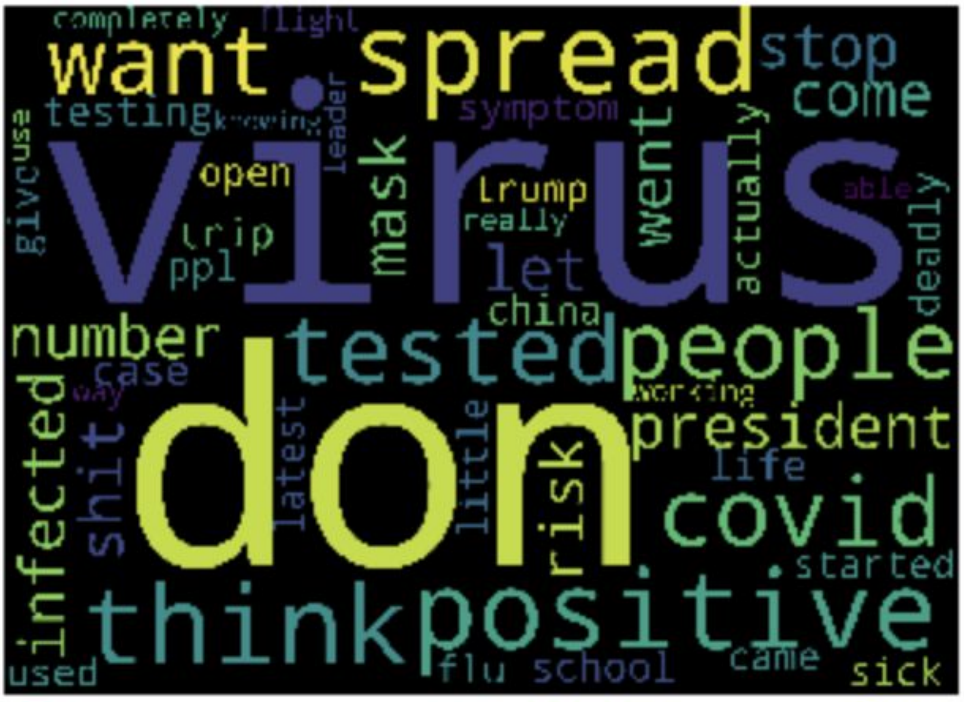

Figure 2. Word Cloud based on topic modelling

Design and Implementation

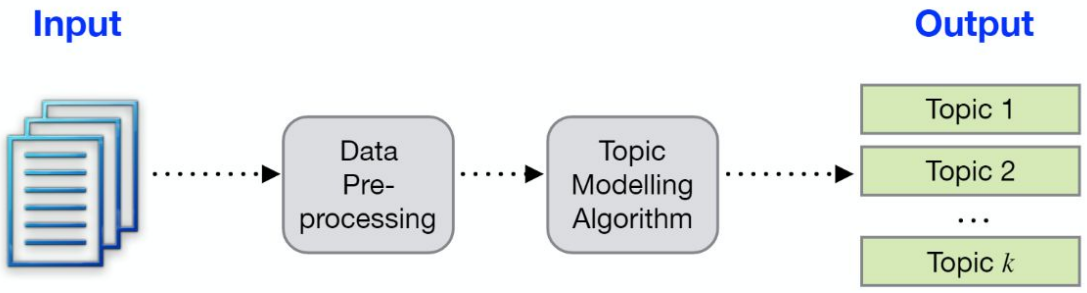

Figure 3. Workflow

As shown in the above infographics, we classified twitter dataset into various topics and have implemented below mentioned techniques :-

1. **Latent Dirichlet Allocation** - Probabilistic, generative model which uncovers the topics latent to a dataset by assigning weights to words in a corpus, where each topic will assign different probability weights to each word.
2. **Non-negative Matrix Factorization** - Approximation method that takes an input matrix and approximates the factorization of this matrix into two other matrices, with the caveat that the values in the matrix be non-negative.

Results Obtained

One can view the results of the various NLP approaches in the github repository over here.

We started out by making a graph of the word frequency with the unclean data to explore the tweets dataset. We then did topic modelling and created a word cloud from four of the topics to check the importance of the words in these tweets.

Figure 4. Word Cloud #4

As we can see from the above word cloud, words like ***covid19, PM*** and ***pandemic*** hold most importance in the tweets quered for “virus”. This helped us understand what people all over the world are thinking out loud on twitter when it comes to viruses.

Conclusion

Some of the major conclusions we found are :

- China: Among the countries, China is most talked about. Which may not seem obvious at first because the highest number of cases and deaths are happening in the US right now and not in China. China seems to be back on its feet. But Trump mentioning China again and again while also accusing it for pandemic might be one of the reasons why China is most talked about on social media.

- Donald Trump: Among people, Donald Trump is the most talked about. It may be because of multiple reasons like the US having the highest number of cases and deaths or the blame game between Trump and China.

- Ebola: People are comparing previous pandemic Ebola with Covid19.

- PM: Two possible people responsible for this word showing up in the most talked about could be UK PM Boris Johnson and Indian PM Mr. Modi. The UK PM is in talks because he had the Coronavirus. Similarly, the Indian PM is being talked about because of his bold move to put the whole nation under lockdown and doing better than most other countries in terms of dealing with the pandemic ; WHO also praised the Indian PM for the same.

These are some generic conclusions but the same approach can be used for answering very specific questions. For instance, the government might want to know people's reaction to

certain restrictions that are being put on cities and states. Businesses might also want to know how consumer habits are changing because of the pandemic. The word cloud shown above has words like *online* and *school* which can help draw insights about schools that are operating online. Similarly, a lot of other insights can be inferred from the tweets even in the post COVID19 world.