**Minimum Viable Product (MVP)**

# Coronavirus tweets NLP - Text Topic Modeling

The goal of this project is to answer questions like: ***Throughout covid pandemic outbreak time (2020 Q1), what are the hot topics that people care about?***

***What is people’s attitude towards covid (or certain topic, such as vaccination) as the covid situation heating up?***

**Solution Path**

To do this, I am using Covid relative tweets from 2020Jan to 2020Mar—which generally include tweet date, original tweet text and uploader’s location. The data consist of over 41k tweets.

Then I performed topic modeling analysis on the data using python NLP libraries.

**Work Completed**

In pre-processing the data, I stemmed the words in tweet using PorterStemmer() and used stop\_words=’english’ argument within TfidfVectorizer() function.

To begin building a topic model, I created a few baseline models using the CountVectorizer and TfidfVectorizer, with LSA and NMF topic modeler and a small number of topic components (*n* = 5);

**Recent Findings**

I have opted to tune a model using the TfidfVectorizer. Though I am continuing to tune the number of topics to use in this model, I have determined some important anchors to separate topics, including:

* ['store', 'groceri', 'food', 'supermarket', 'price']
* ['consum', 'oil', 'demand']
* ['hand', 'sanit', 'toiletpap', 'mask']
* ['worker', 'employe']
* ['shop', 'online', 'deliveri']

With only minimal tuning of the anchor strength, the topic model [also] produces some interesting sample topics, such as:

* **Food price**: ['store', 'groceri', 'food', 'supermarket', 'price']
* **Oil demand**: ['consum', 'oil', 'demand']
* **Covid time must haves**: ['hand', 'sanit', 'toiletpap', 'mask']
* **Employment situation during covid**: ['worker', 'employe']
* **Shop online**: ['shop', 'online', 'deliveri']

**Moving Forward**

I will continue to tune (1) the anchor words and (2) the number of topics modeled to make the topics more coherent/interpretable. The final topic model will be used to label tweet and sensitivity analysis will be done based on topics.