# Background

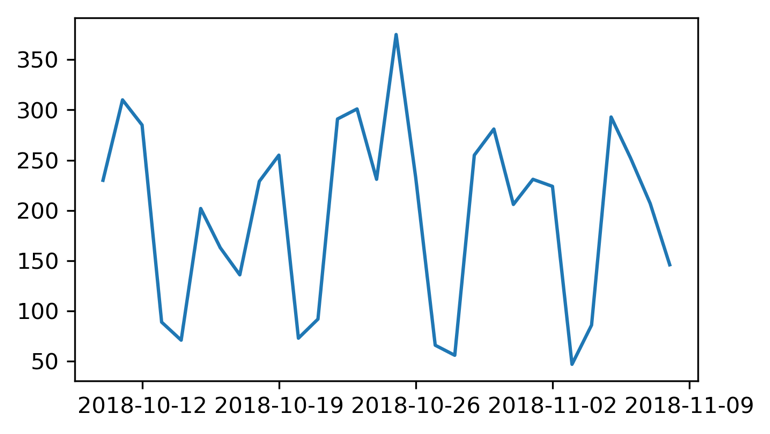
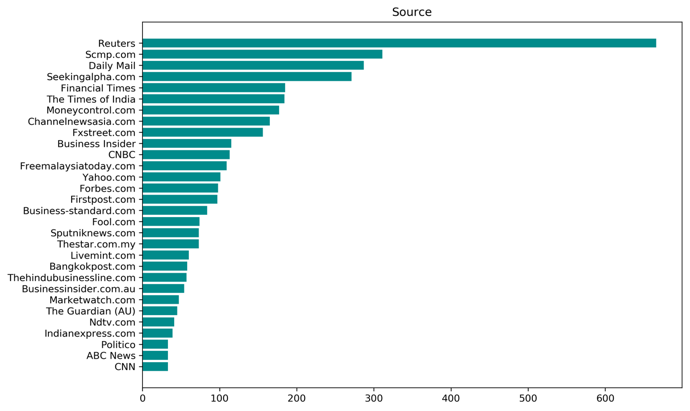
US and China trade war has caught the eyes of general public since the beginning of 2018. As it heats up with tariffs increasingly imposed against each other, a variety of industry has been influenced. Meanwhile, some countries have switched their attitudes towards US and China.

Once people want to make a decision based on ongoing status of trade war, they will first look at recent news since it reports recent happenings and opinions from domain experts. It is easy to gain some insights upon reading through one or several articles. However, figuring out trends, patterns, and correlations of different aspects of the trade war is both complicated and time-consuming.

Machine learning and nature language processing techniques can be very useful in mining insights from tons of news articles. Then, I set up goals of building up a model to help investors make wise decisions and assist the government to generate strategic diplomatic policies.

# Data

The metadata was collected by a free news API offered by newsapi.org, such as an article’s title, description, snippet of content, media source, published date and time, as well as URL link. The free version of API only allows end-users to pull metadata of 1,000 articles per day published within the most recent month. In total, I collected 6,000 data points and scrapped all the full article by iterating over the URL links. Popular sources and trend of number of posting are depicted in the figures below.

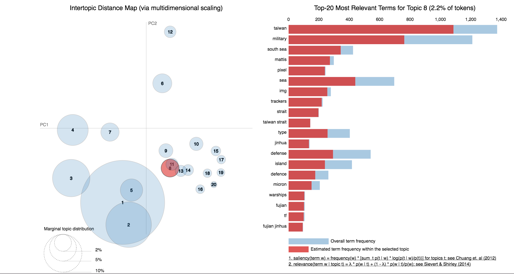
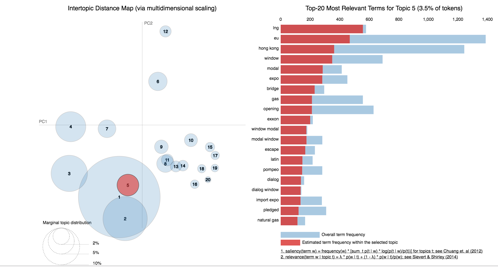
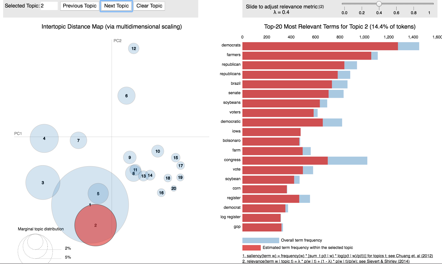


# Design

The general task flow contains six steps: data collection by querying API and scrapping, preprocessing (cleaning, tokenization, vectorization), Latent Dirichlet Allocation (LDA) Topic Modeling, Sentiment and Similarity Analysis, Data Visualization with Tableau and Clustering in time plot.

## LDA Topic Modeling

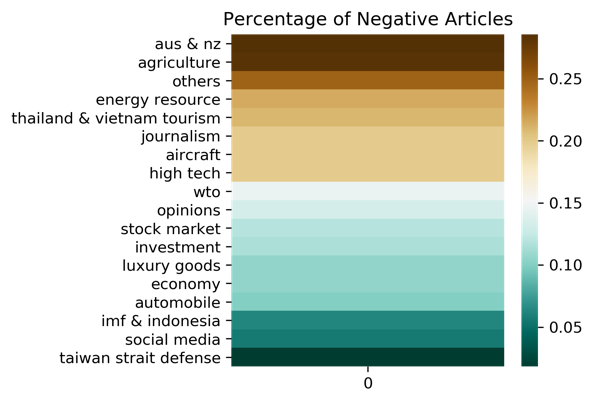
After a trial of different numbers of components in LDA model, I finally landed on 20 topics and got some meaningful results. The Viz of LDA shows topic bubbles scatter all over the principle component feature space without much overlap between each other (see the left side of Viz below). The top 20 words ranked by their probability of showing up in the context given a topic are listed on the right side of the LDA Viz. A concept was assigned to each topic with respect to the top 20 keywords. Interestingly enough, the topics are related to either an industry or a country involved in or affected by the trade war. The three Viz below represents topics as Agriculture, Energy Resources, and Taiwan.



## Sentiment Analysis

The sentiment analysis was completed by the TextBlob tokenizer’s pre-trained model of which the output is a polarity score ranging from -1.0 to 1.0 and a subjectivity score ranging from -1.0 to 1.0. After measuring the 15% quantile and 85% quantile of the polarity, I saw 15% of the documents have polarity below 0.0 and 15% have polarity above 0.124. I labelled the 15% low in polarity as negative, 15% high in polarity as positive and the remaining as neutral. An adjusted polarity score was then given to each topic: -1.0 for negative, 0.0 for neutral and 1.0 for positive.

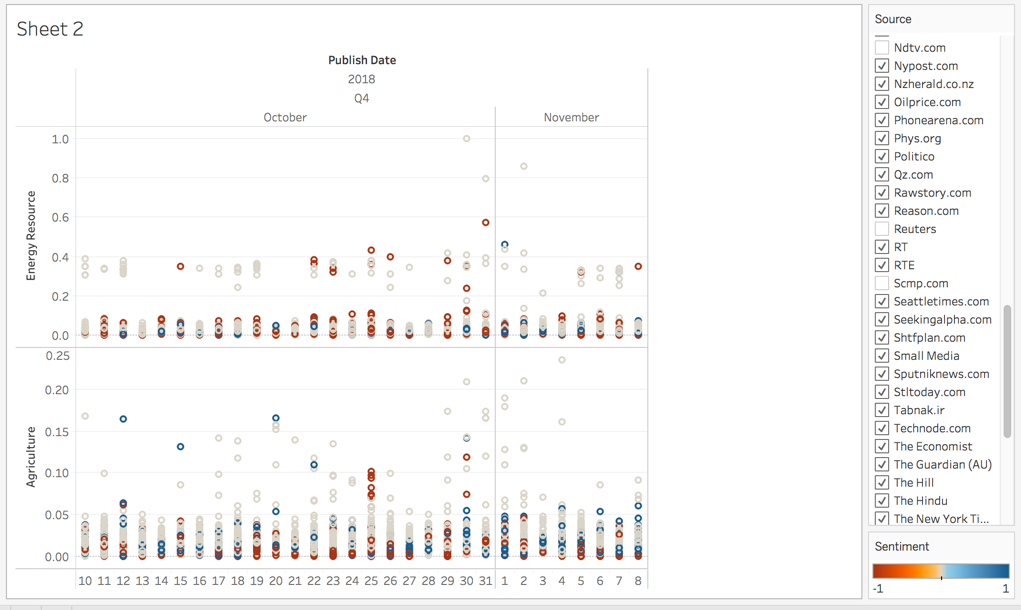
The analysis was conducted on each document or each news article, and then aggregated by topics. I chose two metrics to evaluate sentiment of topic: percentage of negative articles and mean adjusted polarity score. The results are intuitively visualized as heatmaps below.

As shown in the heatmap, agriculture and energy resource ranked high in percentage of negative articles and low in mean adjusted polarity score. It indicates that trade war may have some negative impacts on the agriculture and energy resources. After viewing typical articles of those two industries (topics), I realized that China imposed additional tariffs against US exported agriculture product which hurts the agriculture of US. Also, crude oil prices slumped because the biggest oil consumer, China, reduces its import of oil due to slower growth in economy incurred by trade war. According to the sentiment analysis of news, it is not recommended to invest in agriculture and energy stocks in the trade war period.

## Similarity Analysis

The cosine similarity of Term-frequency Inversed-document-frequency (TF-IDF) vectorizer was computed pair-wisely between articles. The final product is an interactive tableau dashboard. The end user can pick up a topic, and then the system will choose a typical article from that topic and plot its similarity with all other articles in a time plot (see figure below). The user can interact with the time plot by hovering mouse over the bubbles and the metadata of the corresponding article will be displayed. By clicking on the URL link, the user can go directly to the webpage and read through the full article. Plus, options are available for users to apply filters on news sources and color code the sentiment. This interactive dashboard offers the user not only a global view of a topic’s hotness and sentiment trend over time, but also a detective’s magnifier to pinpoint the root cause.



# Future Work

If granted some funded to get a license of well-developed API, I would collect articles in a longer time span with a higher quality. Next, in the Tableau demo, typical articles from each topic shall be selected after examining the full content of the article instead of examining just the probability of documents (articles) over topics. Also, fed with more data points, my model may discover some intriguing clusters with visual assistance of Tableau interactive dashboard. Eventually, I would build up a flask application which recommends a reader with more articles similar to what he/she is currently reading.