Progress Report

Our group has made some reasonable progress on our project, which is reproducing the paper outlined in *Mining Causal Topics in Text Data: Iterative Topic Modeling with Time Series Feedback.* Right now, we have our response data: our time series data which we arbitrarily chose to be the closing price of Facebook, Apple, Microsoft, Tesla, and American Airlines stock from 1/2/2018 to 10/30/2020. We have also decided what we will be using as our text data: the most popular tweets surrounding a ticker symbol tag on twitter. In our case, $FB, $AAPL, $MSFT, $TSLA, and $AAL. To get our document collections, we have begun using Tweepy, a twitter API, to create a collection of top tweets containing each ticker symbol along with the day they were tweeted. We have also written out the pseudocode for the Iterative Topic Modeling with Time Series Feedback algorithm so that we can begin with our model soon. The major tasks that we still have to carry out are finalizing the document collection, writing the code for our topic modeling of the tweets, deciding what our causality measure will be and which testing strategy (Granger or Pearson) we will use to evaluate significance, deciding how strong of an effect we want our prior to have, and writing code to perform sentiment analysis on tweets and words.

One challenge we are working through is people who tag several ticker symbols in their tweet to try to make it more popular. These tweets often are not focused on the stock we are trying to observe and will create unnecessary noise. Another challenge is getting a complete understanding of how to use the prior in the iterative topic modeling and writing out the code for this process. A third problem we are facing comes in the presence of pictures. Oftentimes people will tweet a picture of a stock chart or the picture will contain essential information without which the tweets itself may seem out of context or to be missing information. Therefore, we are trying to think about what the best way to handle pictures will be. Right now we are thinking about filtering images out and ignoring all non-text data, but should we find that these tweets contain crucial text data, we may try to include them somehow.