1. Overview:
   1. Following the word cloud, we wanted to delve deeper into the emotional sentiment of WSJ reporting and apply a more quantitative approach to analyze the text matrix and its impact on the financial market. To get a slight taste of latent semantic analysis for future research, we did some self-studying on different word lexicons and methods like principal component analysis and singular value decomposition and performed two kmeans clustering on all Wall Street Journal news titles. We also fitted a linear regression model using our text matrix to predict the relative position of the Industrials Index of the New York Stock Exchange, drawing from the hypothesis that media sentiment might impact people’s financial decisions.
2. Introduction:
   1. Text matrix: For our text matrix, each row of the wrangled text dataset represents all the news titles in that single day (June 1st to November 3rd), and there are two columns: the first column is just the date, while the second column contains all the news titles in one string.
   2. Lexicons: When we use the NRC lexicon, each word receives a ”yes”(1) or “no”(0) value for each of 8 categories (like trust, anger, surprise…). We represent every day with an 8-dimensional vector, each dimension representing an emotion. The value of each dimension is the number of words that day that contain the emotion. Similarly, for the Bing lexicon, each word is simply classified as either “positive”(1) or “negative”(0). We represent every day with a 2-dimensional vector. The first dimension is the number of words that are positive in that day. The second dimension is the number of words that are negative in that day. So .
   3. Models/Methodology:
      1. K-Means Clustering: Because the NRC text matrix represents every day with 8 emotions, we allow the user to choose 2 out of 8 emotions to perform K-Means clustering on. We also allow the user to choose
      2. Linear Regression: Our hypothesis is that WSJ’s reporting sentiment might not change the long-term fundamentals of an industry. However, the sentiment may impact people’s short-term perception of where the market is headed and therefore might be an effective indicator of short-term trading momentum (short-term movements of the index). Also, WSJ is a relatively conservative media, so its readership might include more investors of traditional industries like energy and industrials. To test out this hypothesis, we designed two variables:
         1. Percentage: A measure of the relative position(0% - 100%) of the industrials index of the New York Stock Exchange. We measure the closing price, high price, low price of every trading day within a user-selected date range. 0% would mean that the industrials index is at its lowest point during the selected time period while 100% would mean that the index is at its highest:
            1. Percentile = (Close\_Price - min(Low\_Price))/(max(High\_Price)-min(Low\_Price))
         2. Sentiment Ratio: To adjust for the absolute number of words, the ratio(0 - 1) is the number of positive words divided by the number of negative words.
            1. Ratio = Positive\_Num/Negative\_Num
3. Analysis/Discussion:
   1. K-Means:
      1. In our shiny app project, we discovered that Wall Street journal is a relatively conservative media that sees things in a more negative than positive light.
      2. Looking at the K-Means Clustering plot using the Bing lexicon, we further discovered that positive words and negative words seem to be negatively correlated in their frequency(in the bing clustering plot), which suggests that WSJ is a consistent media as well(the sentiments of their articles remain relatively stable in a given day, that positive news titles appear together and vice versa). A more dynamic/diverse reporting style would probably see a more random distribution of word embedding plots.
      3. Looking at the K-Means Clustering plot using the NRC lexicon, we
   2. Linear Regression:
      1. By fitting our linear regression model over many choices of date ranges, we generated a predictive model by including all dates from June 1st to November 3rd, and our following model has a significant linear coefficient with a p-value of , suggesting that more positive WSJ sentiment is positively correlated with an upward momentum for the industrials index.
4. Limitations:
   1. We experimented with but gave up on low-rank approximation as approximating the 8-dimensional NRC matrix to 2-dimensions killed off too much variance within the dataset. As a result, it became very hard to tell which dimensions explained the most variance within our text matrix. We instead gave the user the option to choose 2 dimension and perform K-means Clustering. Even though we displayed a sum of squared error plot on the side, we believed that more interpretations could be drawn from the dataset if we designed more intricate ways to represent our text matrix in 2 dimensions (by grouping emotions into 2 categories).
   2. We only included sources from the WSJ and performed a simple linear regression in the industrials index, and the industrials index was relatively stable during our chosen date range. To further our research, we could include a more comprehensive list of news sources (political spectrum, age of readership etc). We could also include longer date ranges to see if sudden market shocks are predictable.