Introduction and summary

In this repository, I have conducted analysis of a dataset containing all available tweets from all 100 senators who were in office during May 5th, 2017. The analysis contains a comparison of the type of language used between different political parties. We also take a look at an internet created metric referred to as "The Twitter Ratio" as a means of gauging how good a tweet is. You will also see a comparison of how Republican and Democratic tweets are rated with this metric.

The outline of this repository is as follows:

1. **Exploratory Data Analysis (EDA) and Wordclouds** – Analyzing the data by generating statistics such as word frequencies and twitter ratio, over different party affiliations as well as plotting some wordclouds.
2. **Natural Language Processing (NLP) to build a model predicting tweet origin** – Using basic text processing methods such as tokenization, stop word removal, lemmatizing, and vectorization of text. Using the scikit-learn library to build a model allowing us to predict which party a tweet originated from.
3. **Using NLP again to build a model predicting The Twitter Ratio** – Using the same text processes described in section 3, but this time to predict how a tweet rates according to the Twitter ratio metric.
4. **A WIP to run a PYMC3 Bayesian estimation model** – This notebook is a work in process. It contains an attempt to show that Republican and Democratic tweets have a different Twitter ratio using Bayesian estimation.

1. EDA of the tweets

a. First things first, let's look at the entire dataset and see what "words" or "tokens" are most frequent. The top hits are "http" and "co" (it is safe to assume that co is the lemmatized product of com). Due to the character limitations of a twitter post, many senators post links to articles that they agree with.

    i. When looking at each individual party's top used words we see that both Democrats and Republicans use twitter as a legislative platform. Independents much more issue centric with their posts. One word of interest showed up in all three parties but at different ranks, Family. For Democrats it was #5, Republicans it was # 12, and Independents it was #14.

b. When we look at the most frequent bigrams in each party, some magic starts to happen. Health care is in the top two for all three parties.
    i. Democrats will tweet "President Trump", "sexual" assault", and "gun control". Republicans will tweet "President Obama", "happy birthday" and "thoughts and prayers". It is clear what each party thinks is important.

c. The Twitter Ratio! What is it? The ratio refers to an unofficial Twitter law which states that if the amount of replies to a tweet greatly outnumbers the amount of retweets, then the tweet is bad. If your ratio is greater 2:1, then you have messed up. For the sake of analysis I have divided the EDA here into three sections, Low(< 2), High(> 2), and Extreme(> 40).
    i. The majority of the tweets gathered live in in the lower spectrum of the ratio, with Democratic tweet frequency being the highest in this group.
    ii. Republicans tweets dominate the high spectrum of the ratio.
    iii. While there are very few cases in the extreme spectrum of the ratio, it is almost exclusively comprised of Republican tweets.

d. By dividing the ratio into quartiles, it becomes clear in which segment the majority of each party's tweets reside. Independent and Democratic tweets are mostly in the second quartile, where the majority of Republican tweets live in the fourth quartile.

2. Building a Bipartisan Model to Predict a Tweet's Origin

a. In this section I've built a model to analyze a tweet and make a prediction as to its origin. You'll notice that I've explicitly called this a bipartisan model. Although the dataset contains Republican, Democratic, and Independent parties, these classes are uneven. Of the 219K tweets analyzed, there are only ~9000 Independent tweets.
    i. Another note, Independent senators do not have a shared platform as to which topic trends are likely to be found.
b. A baseline accuracy was set up by looking at the most occurring class (Republican) and then dividing it by the total number of tweets.
    i. Baseline accuracy = 53.97%
c. I build both Random Forrest Classifier and Logistic Regression models. Single words were tested as well as bigrams. Ultimately the best model was a Logistic Regression using single words as input, with a custom stop word list, using an l2 penalty and a "lbfgs" solver.
    i. Model Accuracy = 83.31%

3. Building a model to predict a tweet's ratio

a. In this section I had built a model to predict the twitter ratio score of a tweet. This did not work. Initially I built a Random Forest Regression and a Linear Regression model but the accuracy on a continuous target was absolutely terrible. Preliminary accuracy scores were in the range of -9.0%. The solution was to engineer a new target: ratio quartiles. By lumping continuous values in to bins, this became a classification challenge.
b. To establish baseline, I once again looked at the most occurring class: the second quartile (25-50%)
    i. Baseline Accuracy = 25.01%

c. Out of all the Random forest and Logistic Regression models built, the best model was a Logistic regression with an L1 penalty using single words as input. This model was the least over fit.
    i. Model Accuracy = 39.07%

4. Conclusion

a. The goal of this project was to gain insights as to the difference between tweets from each party. I did see a clear distinction in the twitter ratio of each part as well as the topics each party found to be the most important.
b. I would to do further analysis of this dataset on what topics were most tweeted during the 2016 election and how these tweets rated on the ratio.