Information gain

This houses investigation into 2 way information gain

- Two way IG feature selection

Two way IG feature selection

We're doing feature selection with two way IG like I did with mutual information for constructing my distantly supervised twitter sentiment lexicon in the sarcasm detection project (Comparison of two-side MI). However, in this case, we will just be trying the absolute version since that seemed to work best for sentiment.

One of the possibilities that we wanted to explore was whether you could bias against the majority class by doing feature selection that includes more information from the minority class.

Right now (03-17-2018) I am running balanced two way IG. After that, I will get results for 1.5 times more features from the minority class versus the majority class.

Revisiting 2-way IG

Sandra wants us to try out some different ways of calculating IG and then look at the top 100 results for each method to see what the quality of the results we're getting are.

I proposed the following methods to Danny

- IG(pos) - Abs(IG(neg)) grabbing from lowest and highest
- Norm(IG(pos)) - Norm(Abs(IG(neg))) grabbing from lowest and highest
- Abs(IG(pos) - Abs(IG(neg))) grabbing from highest
- Abs(Norm(IG(pos)) - Norm(Abs(IG(neg)))) grabbing from highest