Text Analysis for Social Scientists and Leaders

Assignment 1

For the class assignments, you will need to create an RStudio project from this repository, which has the data you need.

https://github.com/myeomans/TASSL

In Part 2 of the provided code, you learned how to build an ngram model to predict the price of different restaurants. You split your data in two. You trained a model to predict restaurant price in one half, and tested the model's accuracy in the other half. You also plotted the coefficients from the model. For the next three questions, complete the exact same workflow for one other quantity of interest. You are allowed to build a model to detect either gender (look for the "male" variable) or star rating (look for the "stars" variable). The choice is up to you!

1. Produce a plot showing the frequencies and coefficients of the features in your model. Make sure it is easy to read!

2. Report the percentage accuracy of the model you trained, using the held-out data, and write a few sentences interpreting the results. What features seemed to be the best predictors? How do you think you could improve the model?

In Part 3 of the provided code, you learned about a new dataset of glassdoor reviews. As with any new dataset, we did some important exploration of the text! Then we created a random split, and built an ngram model to predict star rating ("overall") from the text in the pros box. We applied that model to the pros box and the cons box from the test data

3. Now build a new model - train it on the cons text from amazon. Test it twice - on the pros text from amazon and on the cons text from amazon. Create some plots that compare these four accuracy scores (with error bars!). Make sure you label the plots correctly. You can create two separate plots, or one big plot. What do you think explains the differences in accuracy that you see?

4. Create the ngram coefficient plots for the two ngram models (you can re-use the plotting code from week 2). What do you think of the features the model is paying attention to? Why do you think the models have such different results?

5. Change the preprocessing steps of your two models, and do the whole thing again. You have several options: you could add stop words; you could turn off word stemming; you could include the rarest words; you could use bigrams instead of unigrams. Make some choices about how to change your workflow, and justify why you think they would matter. Then show me the coefficient plot and the accuracy score for your new models.

6. So far you have only used data from amazon employees. Now, I'd like you to do some transfer learning. Use one of the models you trained on amazon data and see how well it can predict star ratings in data from Microsoft. Then build a model using Microsoft data, test it on Microsoft data, and report the accuracy of both tests in a plot or two (with error bars!). Generate the ngram coefficient plot for the Microsoft model, as well.