Text Analysis for Social Scientists and Leaders

Assignment 2

The code you were given in class should show you how to interpret various aspects of a language model, using some restaurant review data. You also trained and interpreted a topic model, and built a multinomial model. Your task this week is to take this workflow and apply it to job descriptions data. You will be predicting the salary of the job (using the "salaryNormalized" column), based on the text of the description (the "FullDescription" column).

First - load the job description data and read the different columns. Then divide your data in two: put 8,000 rows into the training data, and leave the other 2,000 rows for test data.

1. Train an ngram model to predict job salary from the text of the job description. Create a LASSO coefficient plot, and be sure label everything correctly and use a different set of colours beside the default ones we have been using. In writing, explain what the model found - what kind of features are predicting high salaries? Low salaries?

2. Select some examples from the data to explore what the model is learning about the data. Choose two texts that the model correctly gives high scores, and two that the model correctly gives low scores. Make sure they are not too long (ideally less than 200 words) so that you can paste them into your assignment. You should be using the filter() function.

3. Now find some examples of text where the model is getting it wrong. Choose two texts where the model thinks the salary should be high, but instead it is low. And choose two texts where the model thinks the salary should be low, but instead it is high. Again, make sure they are not too long (ideally less than 200 words) so you can paste them into your assignment).

4. Based on your analysis in 2 and 3, what could you do to improve the original model?

5. Let's now create two more models to apply to the test set. However, these models will be very simple. Don't use the LASSO at all - just use the word count and the sentiment to predict the salary. Plot the accuracy scores from your benchmarks and ngram model together.

6. Each job description is assigned a single "Category". In your training data, create a multinomial classifier to classify the five most common categories from the narrative text (you should ignore all the other categories for this question). Apply this multinomial classifier in the test data. What is the overall accuracy of this model? Create a confusion matrix - what mistakes is it most likely to make?

For the remaining questions, train a twenty-topic model in the training data, using the "narrative" text variable.

7. Use findThoughts and labelTopics to learn what each topic is about. Come up with some labels to describe eight of the topics. Put those labels on a labelTopics plot, which shows the five most distinctive words (by FREX) for each topic.

8. Choose one of those eight topics you labelled. Give me a word cloud of the words in that topic, and the two documents that are estimated to have the highest proportion of that topic.

9. Create a topic correlation plot. What are two of your labeled topics that seem like they correlate with each other?