Text Analysis for Social Scientists and Leaders

Assignment 4

You will be analysing a dataset of quarterly earnings calls. One dataset has one row per call. This includes the company ID (IBES_ID), fiscal year and quarter (FY, FQ), the text of their opening speech, the pre-announcement forecasted earnings per share (EPS_consensus) and the announced earnings per share (EPS_actual). Each call also has a unique ID, a date, and a count of the number of question and question-askers (each asker is allowed to ask multiple questions).

There is also a dataset of the question-and-answer session after each announcement. This dataset has the call ID (for merging), the name of question-askers, and their company, along with the name and role of the question-askers. Each row indicates a turn by one person in the conversation (askers and answerers combined, with "asker" indicating which role each person plays). Some summary counts - the order of turns, the order of questions, the number of questions each asker asks - are also included.

Before you begin: the datasets here are not very large, so always set s="lambda.min" when you make predictions or extract coefficients from the glmnet models. Otherwise you may end up with models that are too conservative.

1. Split your data into a training and a test set. The test set will be all calls during fiscal year 2012. What training set should you use, and why?

2. Train a LASSO model using only bigrams and trigrams from the opening speeches as features to predict the reported earnings per share (EPS_actual). Plot the coefficients of the ngram features that predict high and low earnings. What are the main patterns you notice?

3. In addition to the model from #2, train another LASSO model using word2vec embeddings from the opening speeches as features. Train a third model that combines these two feature sets. Calculate the accuracy of the three models on the test set, using the same features.

4. Choose two benchmarks you would like to use to compare to the trained models. Create a well-labelled plot showing the average accuracy scores of the models and benchmarks from questions 2 and 3. Give a description of the results that you found.

5. Find two examples of an opening speech where the word2vec-only model was accurate but the ngrams-only model was not accurate. One example should be when the EPS_actual was high, and the other when the EPS_actual was low. Read the full example yourself, but only paste the first 1000 characters of each speech into your homework (the whole thing will not fit), along with the EPS_actual and model predictions. Based on your reading, why do you think the vector model was more accurate? Do you think there is anything you can do in pre-processing to improve on this?

6. Use the Distributed Dictionary Representation technique to compute similarity of all the speeches to a dictionary - the Loughran & McDonald positive emotion dictionary. How well do the DDR similarity scores predict actual earnings per share? Compute the traditional dictionary score, again using the L&M positive emotion dictionary. How well does this compare to the DDR result? Make a new accuracy plot that includes both of these models, along with your benchmarks from question #2, and describe what you found.

7. There are 448 companies that have entries for all four quarters of both FY 2011 and FY 2012 (i.e. eight speeches total). On this question, we want to calculate the average similarity between a company's speech in 2011 Q1 to the seven other speeches by that same company (i.e. Q2-Q4 of 2011; and Q1-Q4 in 2012). You will need to find these 448 companies, extract their Q1-2011 speeches, and then left-join them into the speech dataset.

   With this dataset, calculate the similarity of each speech to its matching first speech, using word2vec. Then calculate the average similarity for each of the other seven quarters. Put these results in a plot. How does similarity vary over time? You can use the normal similarity

metrics, or you can use the Arora approach and remove the first principal component from each company's speeches before calculating similarity.

8. Let's now turn to the question-and-answer training data. First, create a plot showing the relationship between the asker order (for the first 20 askers in each call) and the number of questions they ask. Do askers early in the call ask more questions than askers late in the call?

9. What features of the companies' Q&A predict their actual earnings per share? Train a separate model using the text of the first ten questions from each call, and another one with the first ten answers from each call. For this question, you will have to carefully merge the turn-level data into the conversation-level data. Use unigrams and bigrams, and create ngram coefficient plots showing the features that predict EPS_actual from each model. You can use the normal stemming approach, or use lemmas to lump words with common roots together.

10. What is the difference between the questions and the answers? Create a politeness plot showing the feature difference in the question text and answer text (as extracted in question 9). Make sure to remove features that occur in less than 10% of calls. Make sure you are using the spacy-enhanced feature set. Train a lasso model to predict whether a turn is a question or an answer using the politeness features and produce a coefficient plot from that model.

11. What is the difference between the answers that companies give during their first quarter calls, compared to calls during other quarters? For this question, merge the conversation-level data into the turn-level data, and use only the first five questions in each call. Extract only unigram and bigram features, and show the resulting model in a coefficient plot.

12. Train a multinomial model, using the same data and features as in question 11, to predict which of the four quarters a call comes from (i.e. a four-category outcome). Apply this to the test data and create a confusion matrix to show the accuracy across all four quarters. What are the most common errors? Use the multinomial output to generate a binary result that mirror the test in question 12 - is this a call from the first quarter, or any other quarter? Using the test set, calculate the accuracy of the models in 12 & 13 in predicting the binary outcome (i.e. first vs. all other quarters) and plot those accuracy results together.