Text Analysis

Due Date: 4/5/2015

Last week we went through an overview on a variety of tools and approaches to handling and processing text.

Objective

Using the dataset from Lecture 13, build the best model to predicting if a movie will be "fresh" or "not fresh".

Put together a notebook that addresses:

- additional data cleanup
- approach you took and how you determined that approach
- end result: showing off the classifier

As mentioned in class, steps you should be considering:

1. Consider feature selection here: we haven't dropped a single feature! How could we evaluate this process? (check out `f_classif`)
2. How might we modify CountVectorizer to achieve better performance?
3. What else is in the quote that we could identify and replace more easily?

Grading

This is an open-ended objective, do the best you can to build a model that improves the score! We are primarily observing:

How well can you code through your problems?

1. Starting by reviewing in detail the content we covered in class. Objectively you're showing us that you at least understand how the code we used in class worked.
2. exploring elements of text analysis and understand results. Objectively, you've tried a few things and can explain why they worked or didn't work (or rather, why they either contributed or did not contribute) to your end model. You may have toyed with the base variations of classifier learners.
3. full exploration of the material: everything from this week you picked up well, applied to a previous knowledge, and created reasonable work. We're expecting to see more creative features and processes to model optimization

How well do you understand the theory?

1. Akin to [1] above, we'd be expecting work to build off of what was done in class. Include educated guesses on why things work the way they work.

2. you've attempted to explain why your model works the way it does rationally. There may be some error to the logic.

3. Some kind of deep dive into explaining why your model is the best model (or why your model didn't succeed) shows you understand the theory behind the last few lessons in class. In particular, paying attention to a few particular features and their effects overall help a lot. The vernacular coincides with material from class.