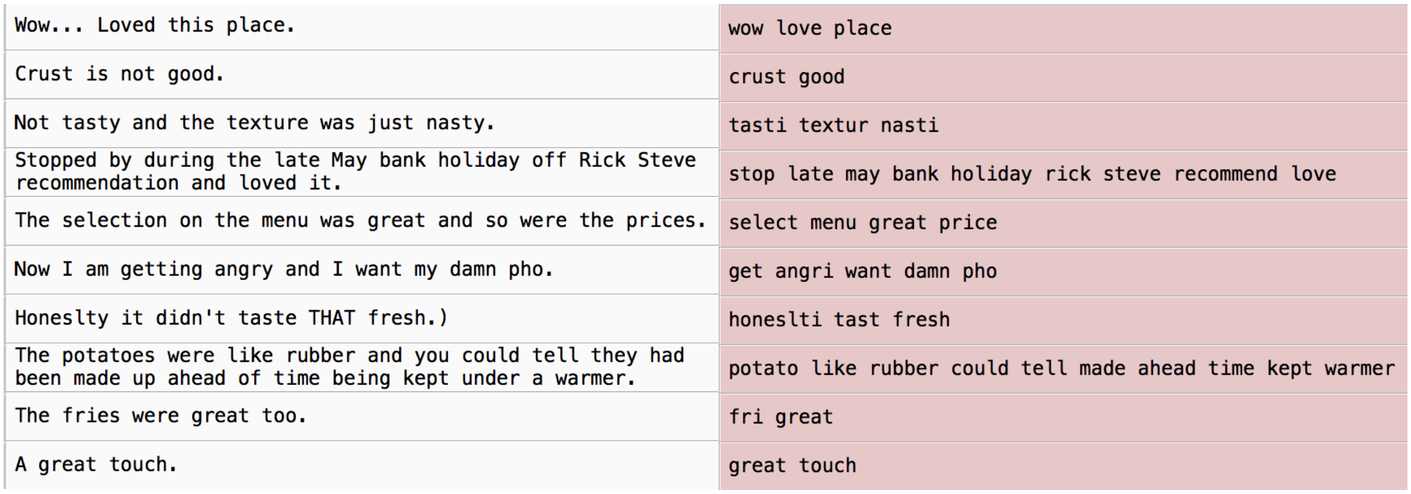
Natural Language Processing Findings

Natural Language Processing is used heavily today. One of the most recognizable uses is in programs like Siri or Google Home where you can speak to your computer and it understands your voice and translates it to text. Another common use is categorizing books by genre based upon the texted in the books.

I will be demonstrating this algorithm on a file containing reviews for a restaurant (*Restaurant_Reviews.tsv*). The machine learning algorithm will attempt to predict whether the reviewer left a positive or negative rating based upon the review.

The first step is to clean all the reviews in order to make the algorithm as efficient as possible. I have done this by removing all characters that are not letters, turning everything to lowercase, removing pointless words (such as *a*, *this, the* etc.), and lastly stemming. Stemming is the process by which we turn words into their root form. For example, *enjoyed* becomes *enjoy*. This is to minimize the total number of different words thus minimizing sparsity. Below are the first 10 reviews before and after they have been cleaned (left being from the *dataset*, right is cleaned words).



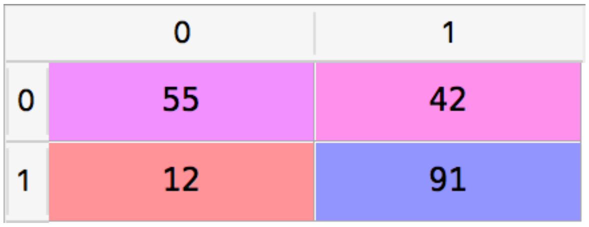
Wow... Loved this place.	wow love place
Crust is not good.	crust good
Not tasty and the texture was just nasty.	tasti textur nasti
Stopped by during the late May bank holiday off Rick Steve recommendation and loved it.	stop late may bank holiday rick steve recommend love
The selection on the menu was great and so were the prices.	select menu great price
Now I am getting angry and I want my damn pho.	get angri want damn pho
Honeslty it didn't taste THAT fresh.)	honeslti tast fresh
The potatoes were like rubber and you could tell they had been made up ahead of time being kept under a warmer.	potato like rubber could tell made ahead time kept warmer
The fries were great too.	fri great
A great touch.	great touch

We can then run the dataset against a machine learning algorithm. For this I am going to be using three different machine learning algorithms (Naive Bayes, Decision Tree Classification, Random Forest Classification) to determine which is best suited for this problem.

For each of these 80% of the dataset will be dedicated to the training set. Thus 800 will be reserved for training and 200 for testing. We have also only taken the most frequent 1500 used words in the dataset out of 1565.

Naive Bayes

After running the model and looking at the confusion matrix we get the following.



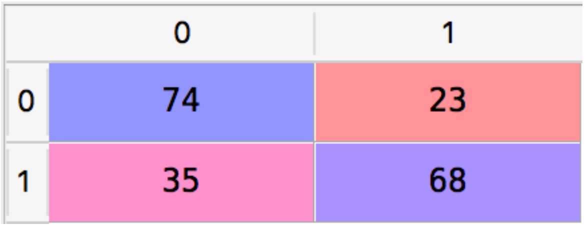
	0	1
0	55	42
1	12	91

- 55 negative reviews were predicted as negative reviews
- 42 negative reviews were predicted as positive reviews
- 12 positive reviews were predicted as negative reviews
- 91 positive reviews were predicted as positive reviews

Therefore, this machine learning model made 146 out of 200 correct predictions (73% accuracy).

Decision Tree Classification

After running the model and looking at the confusion matrix we get the following.



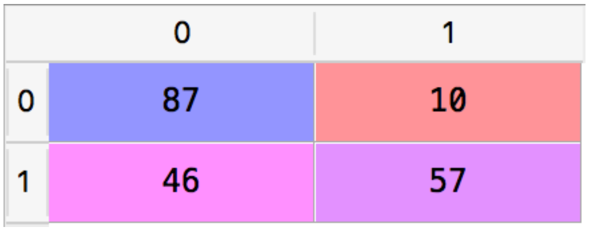
	0	1
0	74	23
1	35	68

- 74 negative reviews were predicted as negative reviews
- 23 negative reviews were predicted as positive reviews
- 35 positive reviews were predicted as negative reviews
- 68 positive reviews were predicted as positive reviews

Therefore, this machine learning model made 142 out of 200 correct predictions (71% accuracy).

Random Forest Classification

After running the model and looking at the confusion matrix we get the following.



	0	1
0	87	10
1	46	57

- 87 negative reviews were predicted as negative reviews
- 10 negative reviews were predicted as positive reviews
- 46 positive reviews were predicted as negative reviews
- 57 positive reviews were predicted as positive reviews

Therefore, this machine learning model made 144 out of 200 correct predictions (72% accuracy).

Summary

All three of these machine learning models are very close in accuracy. *Random Forest Classification* was the most accurate in predicting negative reviews and *Naive Bayes* was the most accurate regarding positive reviews. *Naive Bayes* is the winner with 73% accuracy. This isn’t the greatest, however considering the training set only consisted of 800 reviews the accuracy isn’t terrible. If we had a million reviews in the training set we would have a much more accurate prediction.