Springboard Project 11.3 (Naïve Bayes)

The aim of this project is to find words most commonly associated with positive (“fresh,” in the words of of Rotten Tomatoes) reviews. Specifically, I use Naïve Bayes to find the probability of a fresh rating given the presence of various words in movie reviews. Conversely, I also find the probabilities for words most likely to appear in negative reviews.

I started with brief exploratory analysis, averaging the ratings (effectively 0 or 1) per critic and then putting them into a histogram. The ratings look to be almost normally distributed, with a steep drop from averages above 0.6 to averages below that. The distribution looks more "normal" above 0.6. Subsequently, I complete a few practice exercises as part of the Springboard assignment.

From there, the data are split into test and training sets, and document frequency is evaluated to determine the best value for alpha, a parameter that in this case can be a small, decimal value that will avoid the use of zeroes in probability computations. Next, cross validation is performed, and parameters are selected to optimize model accuracy.

Finally, a model is produced and used to generate the words with the highest probabilities of use in positive reviews. Additionally, the words with lowest such probabilities are produced.