My process began with cleaning the data. After reading in the two text files, I converted them from list type to string type so that I could separate the reviews in each file. I removed all text that came before and after the reviews in each file and converted the reviews in each file from one string to a list of strings by splitting the files by the characters ‘ “,” ‘ that separated each review.

For step 1 of the assignment, using the random and math libraries, I shuffled the reviews in each file and split them into training data (70% of the reviews), development data (15% of the reviews), and test data (15% of the reviews). I also used set.seed to ensure that the data was split in the same group of reviews each time I ran the code. I printed the size of each data set to confirm that they were split correctly. This gave me a training set of 3,731 reviews, and a development set and a test set of 800 reviews each for the negative review data and for the positive review data.

For step 2, I defined a function called naïve_bayes_train to find the conditional probabilities of all the words in the negative and positive training data sets. In this function, I used a for loop to run through each review and split each review into individual words. The naïve_bayes_train function then used the Counter function from the collections library to keep a count of each word that appeared in the review. After running through all the reviews, a dictionary containing each word and the count of each word was created. I then used Laplace Smoothing when calculating the conditional probability of each word given each class. I ran the negative training data and the positive training data to get the probabilities of each word.

When tuning the Naïve Bayes classifier on the development set, I removed some stopwords and characters from all the reviews in the training data such as “the,” “an” and commas. This reduced my accuracy from 76.75% to 76.31%. Even though this reduced my accuracy, I felt that these words and characters should not be given significant influence in this model so I kept this adjustment in the final model. I then tried removing words from the training data that only occurred once in for the negative training data and for the positive training data. This reduced my accuracy on classifying the development set from 76.31% to 75.25%, so I also disregarded this method. I also attempted only including the 5,000 most common words in the training data, but this reduced my accuracy to 75%. When I reduced this to 1000 most common words, my accuracy reduced further to 71%, so I disregarded this method.

For step 3, I joined the training data and test data and trained the Naïve Bayes classifier on this new data set. I then evaluated the model on the test set and got an accuracy of 77.94%.

To evaluate the which reviews the classifier was most and least confident about the class that they belonged to, I created a function similar to the Naïve Bayes classifier, but the outcome was a sorted list of the absolute difference between the negative probability and positive probability of each review with the review from highest value (high confidence) to lowest (low confidence).

The classifier was very confident that these reviews were negative:

- "'you'll laugh for not quite and hour and a half , but come out feeling strangely unsatisfied . you'll feel like you ate a reeses without the peanut butter . . . '"
- "by the final whistle you're convinced that this mean machine was a decent tv outing that just doesn't have big screen magic . "
- “by halfway through this picture i was beginning to hate it , and , of course , feeling guilty for it . . . . then , miracle of miracles , the movie does a flip-flop .”
- "hypnotically dull , relentlessly downbeat , laughably predictable wail pitched to the cadence of a depressed fifteen-year-old's suicidal poetry ."
- “by turns numbingly dull-witted and disquietingly creepy .”

The classifier was very confident that these reviews were positive:

- “my wife is an actress has its moments in looking at the comic effects of jealousy . in the end , though , it is only mildly amusing when it could have been so much more .”
- "by and large this is mr . kilmer's movie , and it's his strongest performance since the doors . "
- “by not averting his eyes , solondz forces us to consider the unthinkable , the unacceptable , the unmentionable .”
- “ryosuke has created a wry , winning , if languidly paced , meditation on the meaning and value of family .”
- "my wife's plotting is nothing special ; it's the delivery that matters here . "

The classifier was least confident about the classification of these reviews that were actually negative:

- “a pretentious mess .”
- “a movie to forget”
- "i can take infantile humor . . . but this is the sort of infantile that makes you wonder about changing the director and writer's diapers .”
- “a yawn-provoking little farm melodrama .”
- “i felt sad for lise not so much because of what happens as because she was captured by this movie when she obviously belongs in something lighter and sunnier , by rohmer , for example .

The classifier was least confident about the classification of these reviews that were actually positive:

- “a literate presentation that wonderfully weaves a murderous event in 1873 with murderous rage in 2002 .”
- "a tender and touching drama , based on the true story of a troubled african-american's quest to come to terms with his origins , reveals the yearning we all have in our hearts for acceptance within the family circle .”
- “a gracious , eloquent film that by its end offers a ray of hope to the refugees able to look ahead and resist living in a past forever lost .”
- “a penetrating , potent exploration of sanctimony , self-awareness , self-hatred and self-determination .”
- “a compelling yarn , but not quite a ripping one .”

Many of negative reviews that the classifier was confident about include some very strong words that I would expect people to use to describe something negative in everyday life such as “dull” and “unsatisfied”. Because these words are expected, perhaps they appeared frequently in the data that the final data was trained on. Two of these reviews contain the word “dull” which may be a signal that that word was given a high probability of being associated in negative review.

The positive reviews that the classifier was confident about do not seem as obvious as the negative reviews. Some of these reviews include a mix of neutral language and language that I would think would be associated as negative such as “mildly.” Perhaps the positive words that appear in these reviews held more of a weighting when it came to training the model. One of the reviews contains the word “special” which seems like a positive word, but is used in this review as “nothing special.” In this context, the word “special” is not being used to convey positive sentiment.

I noticed that two of the reviews that the classifier was unsure of had significantly less words than the other reviews listed above. These may have been more difficult to classify because there were few words for the classifier to work with. If all or many of the words in these short reviews did not appear in the data that the final model was trained on, this would make it especially difficult.

When evaluating the most useful features for each class, I looked at the words that had the top 20 probabilities for each class. The features with the highest probabilities for each class were actually similar for both classes. Words such as “not,” “like”, “no.” I may have needed to remove more stopwords when training the data. I would assume that the words with lower probabilities on each list that differed between the lists had a larger influence on the classification.