Project 3: Reddit & Subreddit Posts Scrapings

By Matt Bildzok

What will I be doing?

I will be web scraping two Subreddit threads about space, titled r/NASA and r/SpaceX. I will take 1000 posts from each subreddit and clean them up into just the Subreddit and the titles. I will then create models to differentiate which posts come from which Subreddit.

EDA

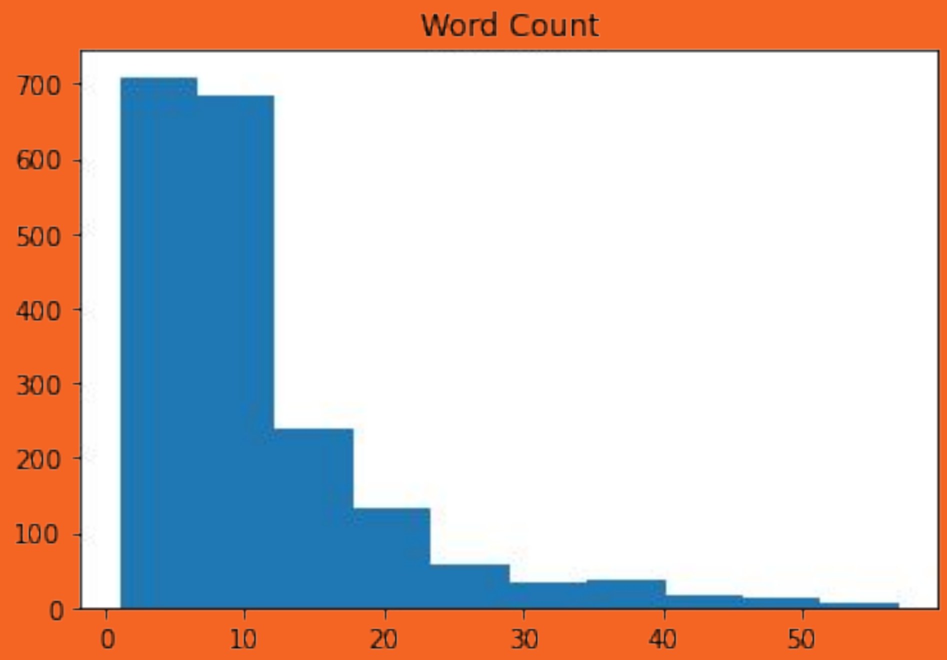
First, after I scraped all the data, I made sure to only include the Subreddit and the Titles, as those are the only two pieces of information that I am interested in.

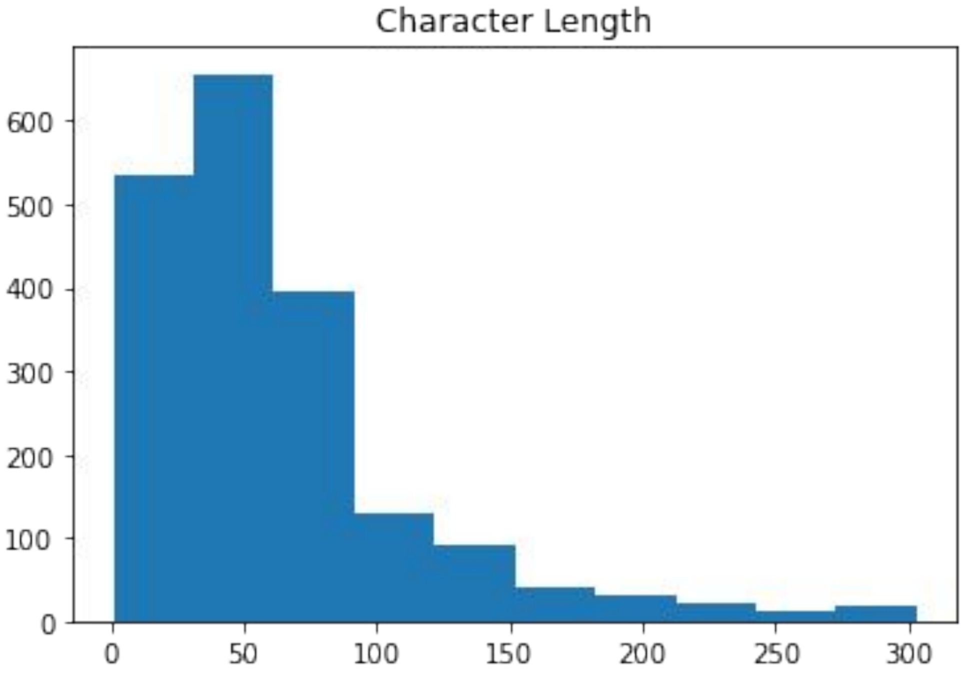
I then made sure that I was able to save my new Dataset into a Data Frame so that I can work with it much easier. After I did that, I exported it so that I can open it in a new notebook.

The next step is to clean the data further, I inspected all the data and made sure to delete the rows that were duplicated based on the titles. This took me from 2000 posts to 1930 posts, deleting 70 duplicates.

I then looked at the longest titles based on words and the longest titles based on characters, just to see how long they were

I noticed that the majority of titles had a small number of words in both Subreddits



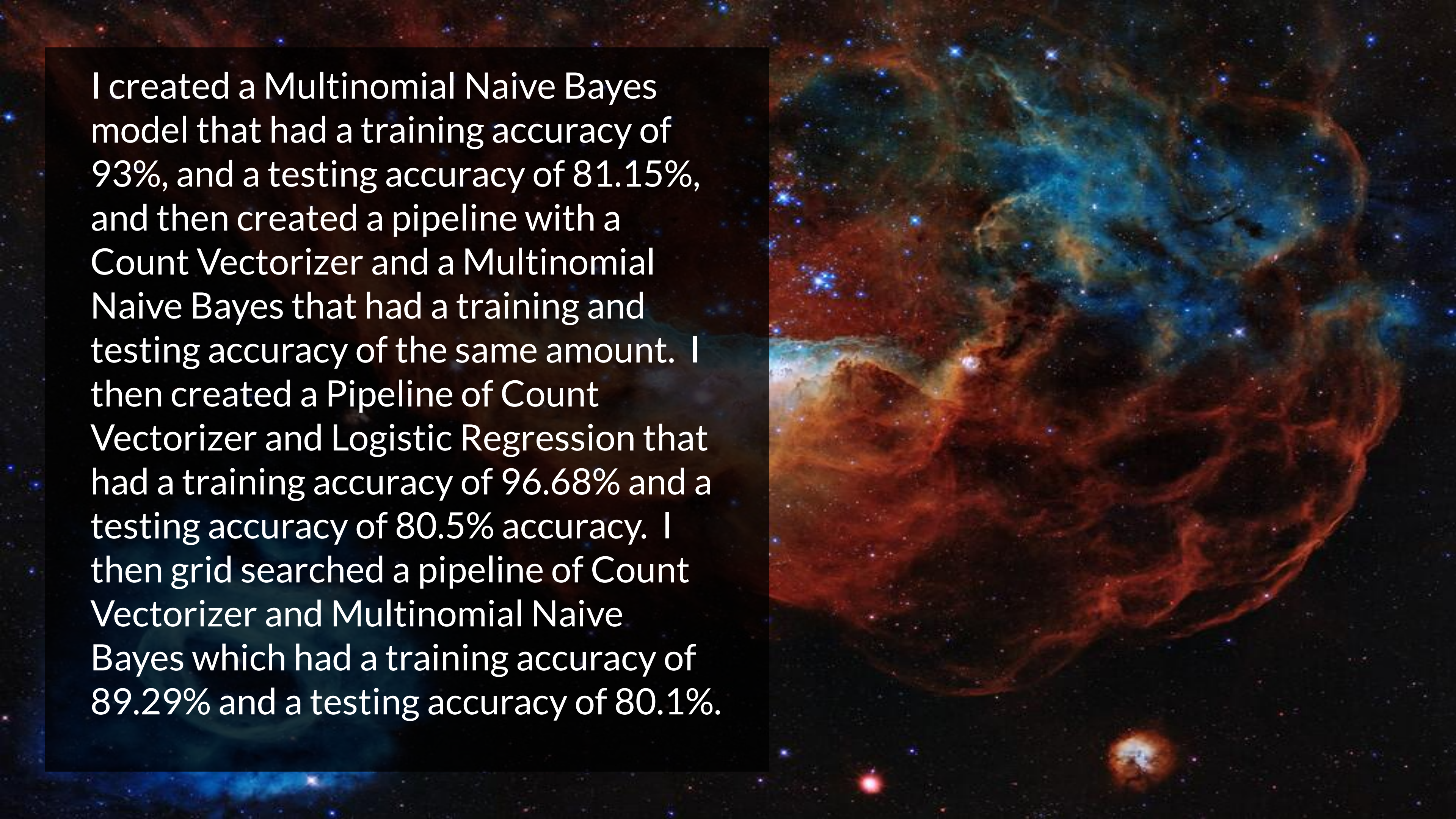


I also noticed that the majority of posts had fewer than 100-150 characters, which was interesting.

Story for illustration purposes only

When building my models, I decided not to take out stop words, or add a single word, such as NASA, as I thought that it would be a bit of a bias and influence the model by doing so. I figured that it may be easier in some light, to include NASA at first, because it is the name of one of the Subreddits, but upon further investigation, NASA was present in both of them, and in a large amount, so I thought that it would add to the complexity of the model.

This turned out to be a good decision, as in the training data alone, the word NASA appeared 0 times in 1212 posts, 1 time in 221 posts, 2 times in 13 posts, and 3 times in 1 post, out of the 1970 total posts.



I created a Multinomial Naive Bayes model that had a training accuracy of 93%, and a testing accuracy of 81.15%, and then created a pipeline with a Count Vectorizer and a Multinomial Naive Bayes that had a training and testing accuracy of the same amount. I then created a Pipeline of Count Vectorizer and Logistic Regression that had a training accuracy of 96.68% and a testing accuracy of 80.5% accuracy. I then grid searched a pipeline of Count Vectorizer and Multinomial Naive Bayes which had a training accuracy of 89.29% and a testing accuracy of 80.1%.

I decided to go with the grid search results as that had the closest proportion of accuracy and still had a very high accuracy compared to what I expected for such two topics, as I figured that not removing the word “NASA” might skew the results quite drastically.

Conclusion

In conclusion, I found that the model that I have created far surpassed what I had anticipated, and can very accurately predict which Subreddit a post comes from , even though both are quite similar and the literal name of one Subreddit appears in both. This will be useful to to predict whether or not a post will pertain to a specific topic or not, and whether or not it can help in people finding things that interest them and branch off from that and explore similar topics.