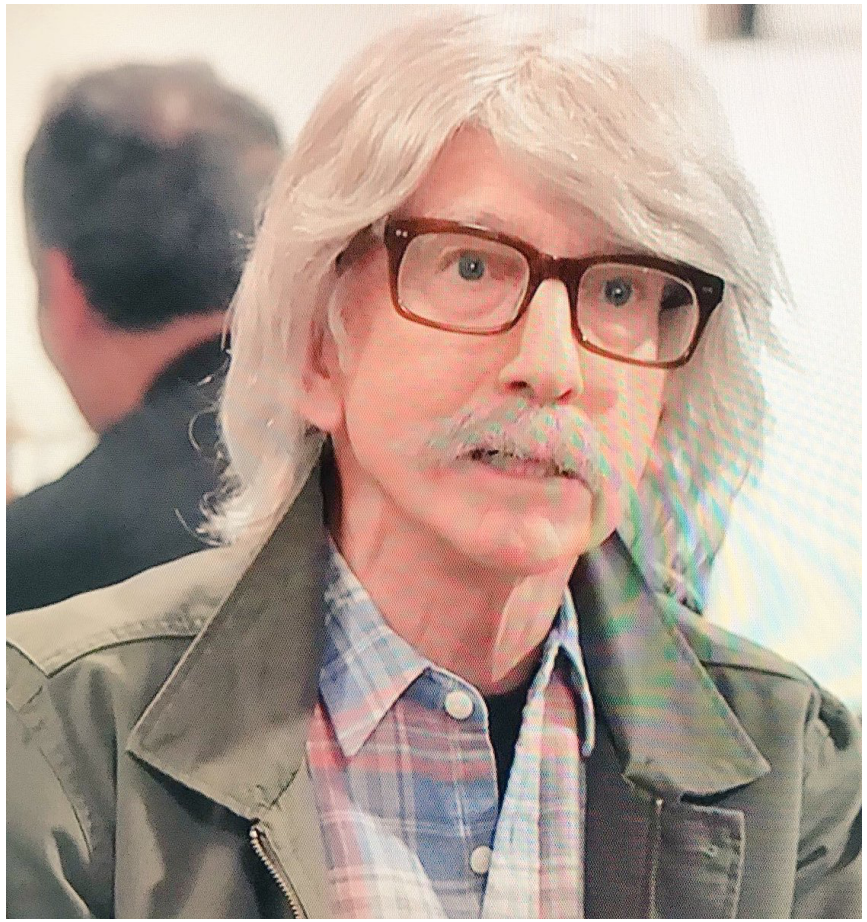
Predicting TV shows by Chatrooms



Noah Monastersky | Data Scientist

Problem Statement

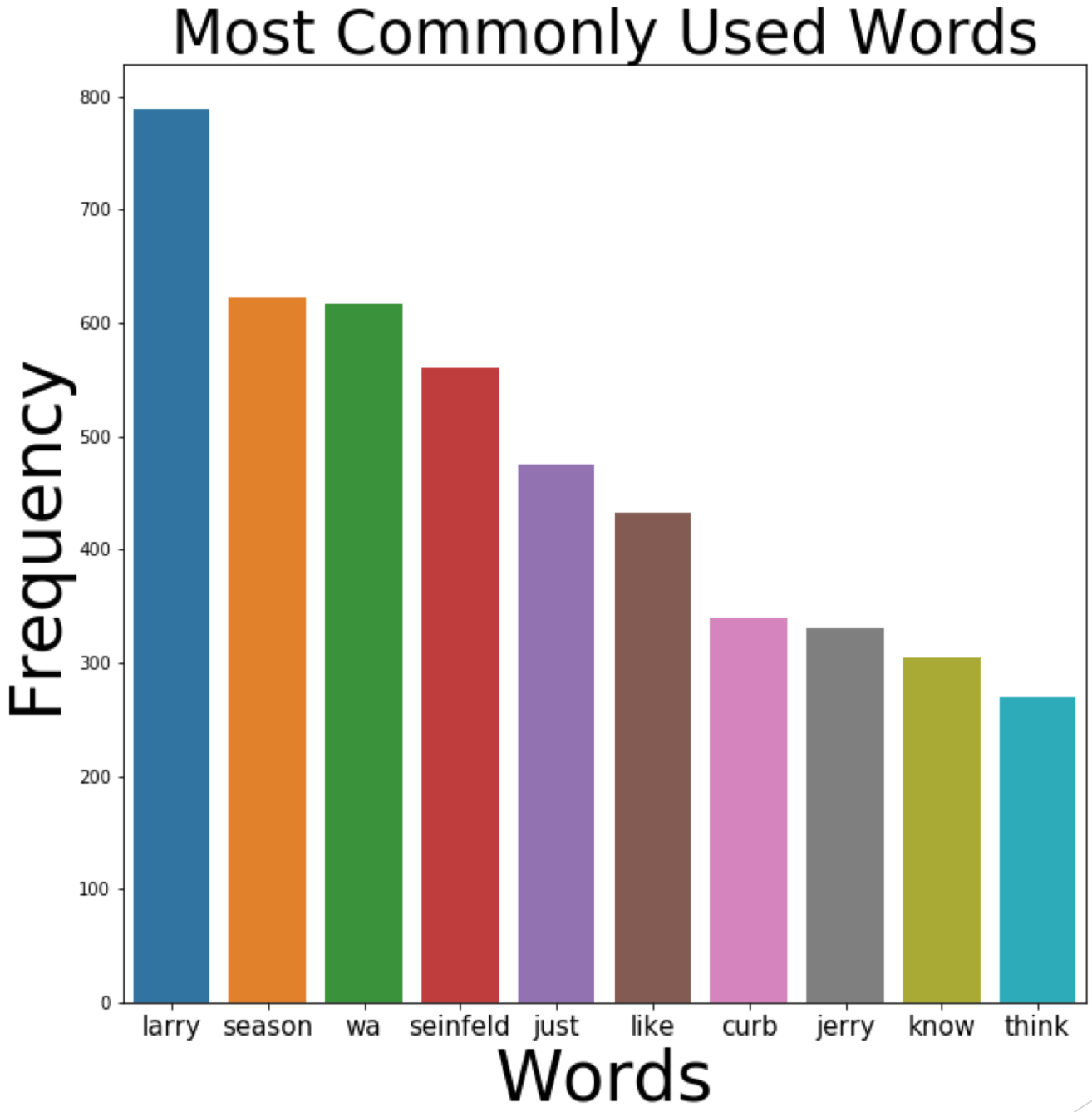
- NBC is trying to see if they can identify which of their shows online comments are talking to so that they can sort them accordingly

Finding the Data

- I used Reddit's API to pull posts made by reddit users on two shows
  - Curb your enthusiasm
  - Seinfeld
- I gathered one dataset that contained around 5000 posts and comments from Curb's reddit page and 5000 posts and comments from Seinfeld's page
- I also gathered another dataset that contained around 1000 posts from Curb's reddit page and 1000 posts from Seinfeld's page.

Data Cleaning

- Combined Title and text on posts
- Removed all website addresses
- Binarized my target variable
- Removed all non-alphabetical characters
- Removed the word “Episode”
- Tried removing the words “Larry”, “David” and “Seinfeld” but found that all these words had strong impact on predictability


Most Commonly Used Words
Frequency
800
700
600
500
400
300
200
100
0
larry
season
wa
seinfeld
just
like
curb
jerry
know
think
Words

Modelling with Logistic Regression

- Base Model
- Used a Count Vectorizer to break up the words which does not weight the words
- Used a Logistic Regression, which works like a linear regression for classification
- I was able to obtain a training score of 99.3% accuracy and a testing score of 87.6% accuracy

Modelling with Naive Bayes

- Used a TFIDF vectorizer to break up the words with English stop words. This weights unique words.
- Used Naïve Bayes which is a classification model that uses Bayes Theorem to calculate probabilities of multiple events. I used an alpha of 10
- I was able to obtain a training score of 95.8% accuracy and a testing score of 91.4% accuracy

Modelling with Random Forest

- Used a TFIDF vectorizer to break up the words with English stop words and set my max features at 1,100. This weights unique words.
- I used a Random Forest which is a classification model that uses trees to classify. I set my max features to be the square root of my features and my number of trees at 2000
- I was able to obtain a training score of 99.7% accuracy and a testing score of 89.2% accuracy

Conclusions

- My Naïve Bayes model worked best as that had the highest test score and the Training and test score were the closest.
- Since all my testing scores for the dataset with comments were in the 70's I can conclude that comments do not help in predicting a show.