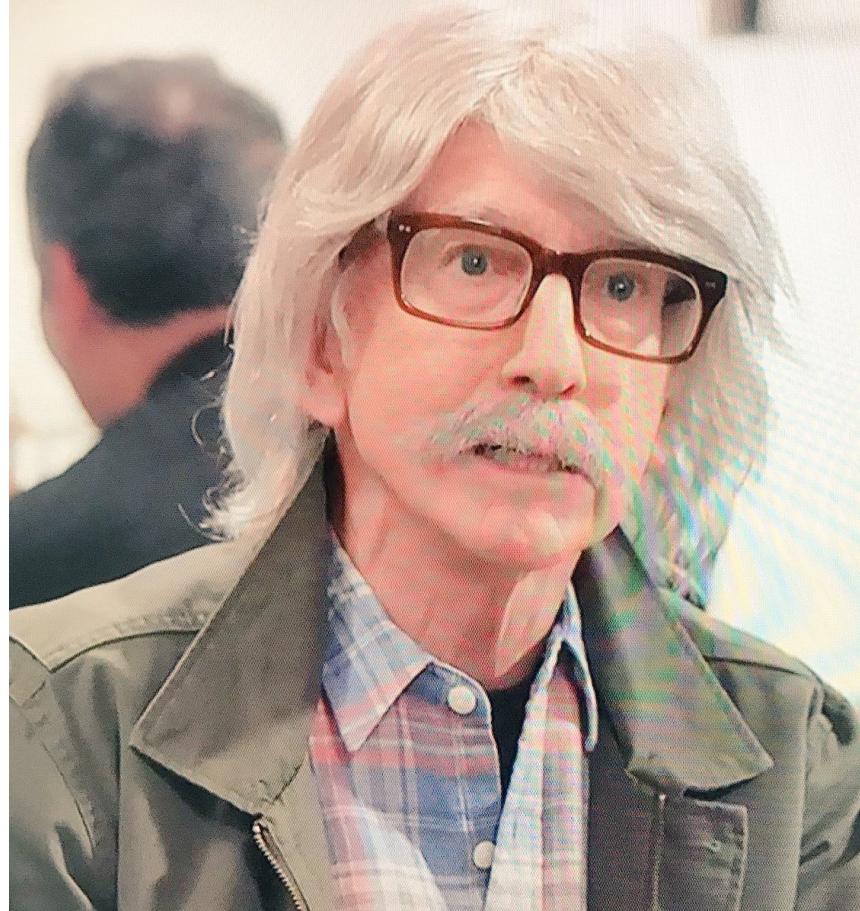


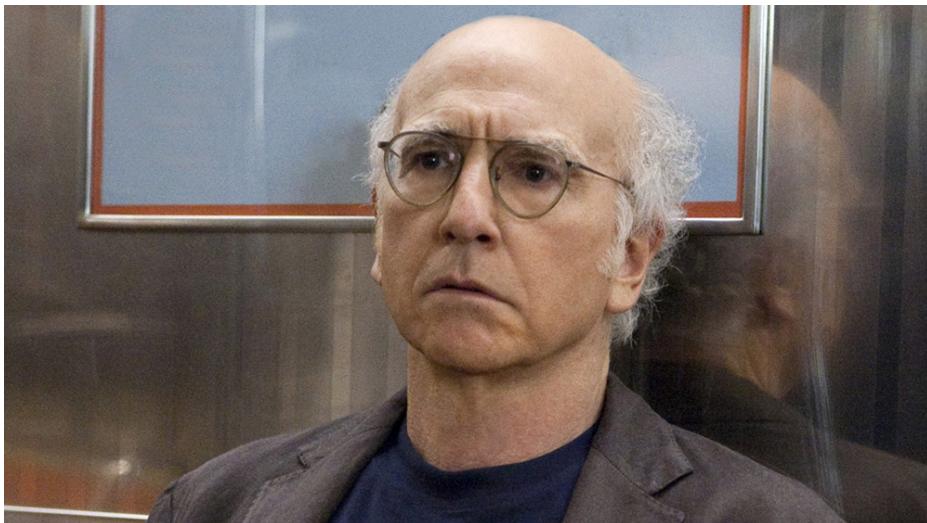
# Predicting TV shows by Chatrooms



Noah Monastersky | Data Scientist

# Problem Statement

- ▶ We are trying to see which show a post is talking about.



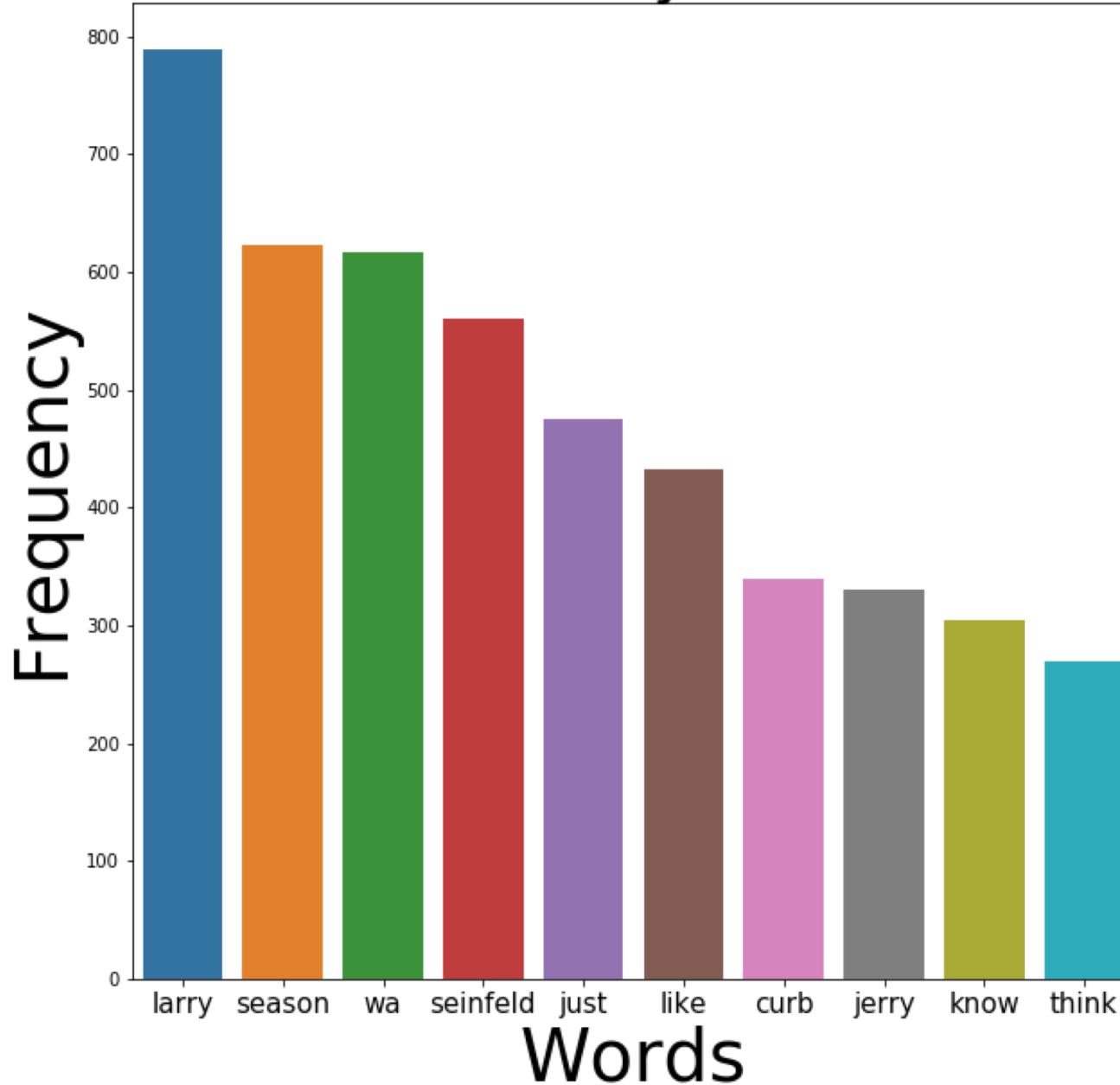
# Finding the Data

- ▶ 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



# 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 99.5% accuracy and a testing score of 86.5% 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 98.6% accuracy and a testing score of 89.5% 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 99.8% accuracy and a testing score of 88.3% accuracy

# Conclusions

- ▶ I would use my Naïve Bayes model since that one is best able to predict my unforeseen data
- ▶ Since all my testing scores for the dataset with comments were below 82 percent I can conclude that comments do not help in predicting a show.

