Making Music with Reddit

pairing beginner songwriters and recordists with the tools to grow

Adam Zucker

Problem

For the aspiring songwriter and recordist, there is a steep learning curve. The ability to pull from the experience of decades of recording and songwriting wisdom available online can be a huge help!

In this study, I will use Naive Bayes and Random Forest models to classify posts in the Audio Engineering and Songwriting Subreddits, with the goal of helping a beginner songwriter and recordist determine where to find the most relevant and helpful information for their learning process.

Contents

Preprocessing

1. Post Collection
2. Data Cleaning
3. Feature Engineering
4. Export for EDA and Modeling

Sentiment Analysis

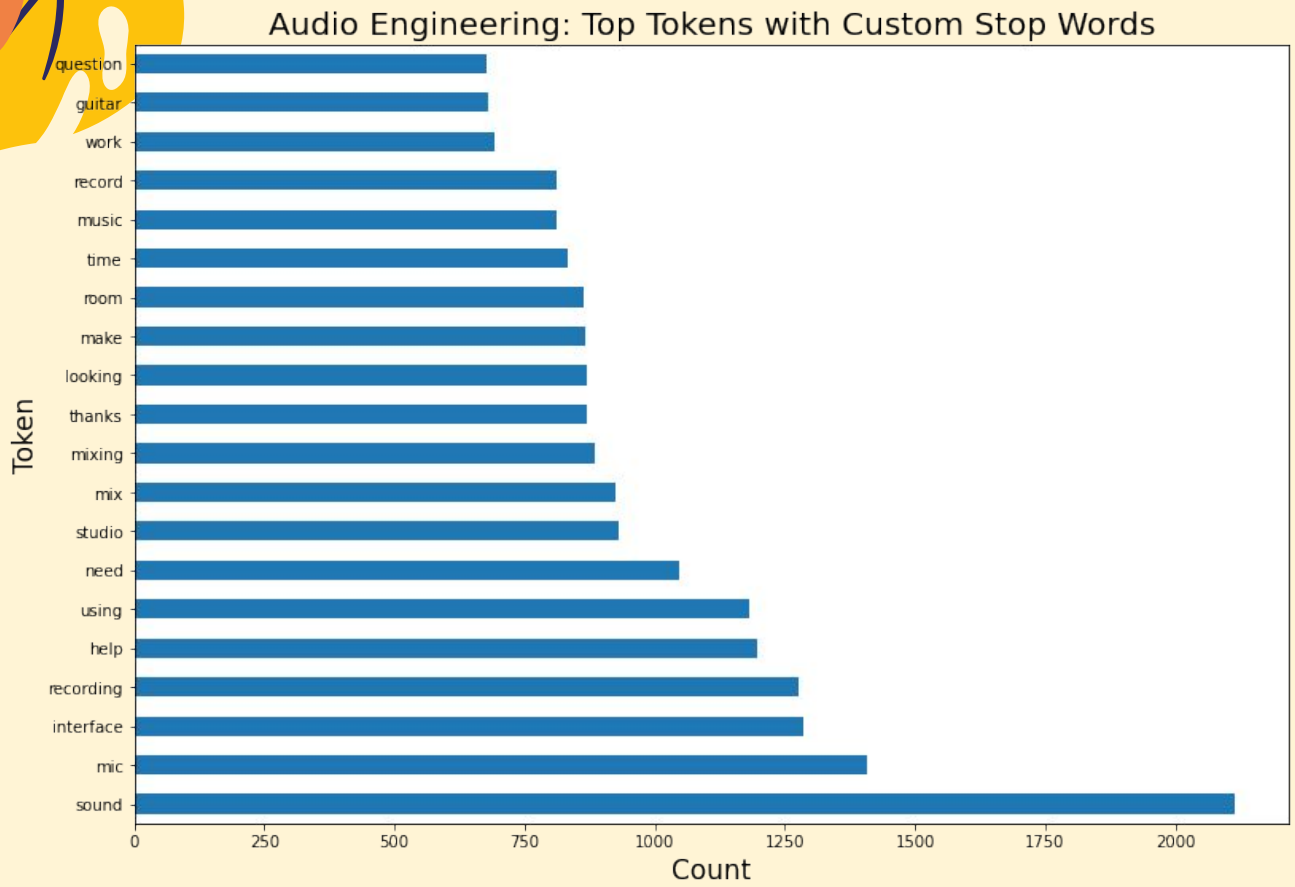
Audio Engineering

- A positive community!
- Sentiment intensity score of 0.488

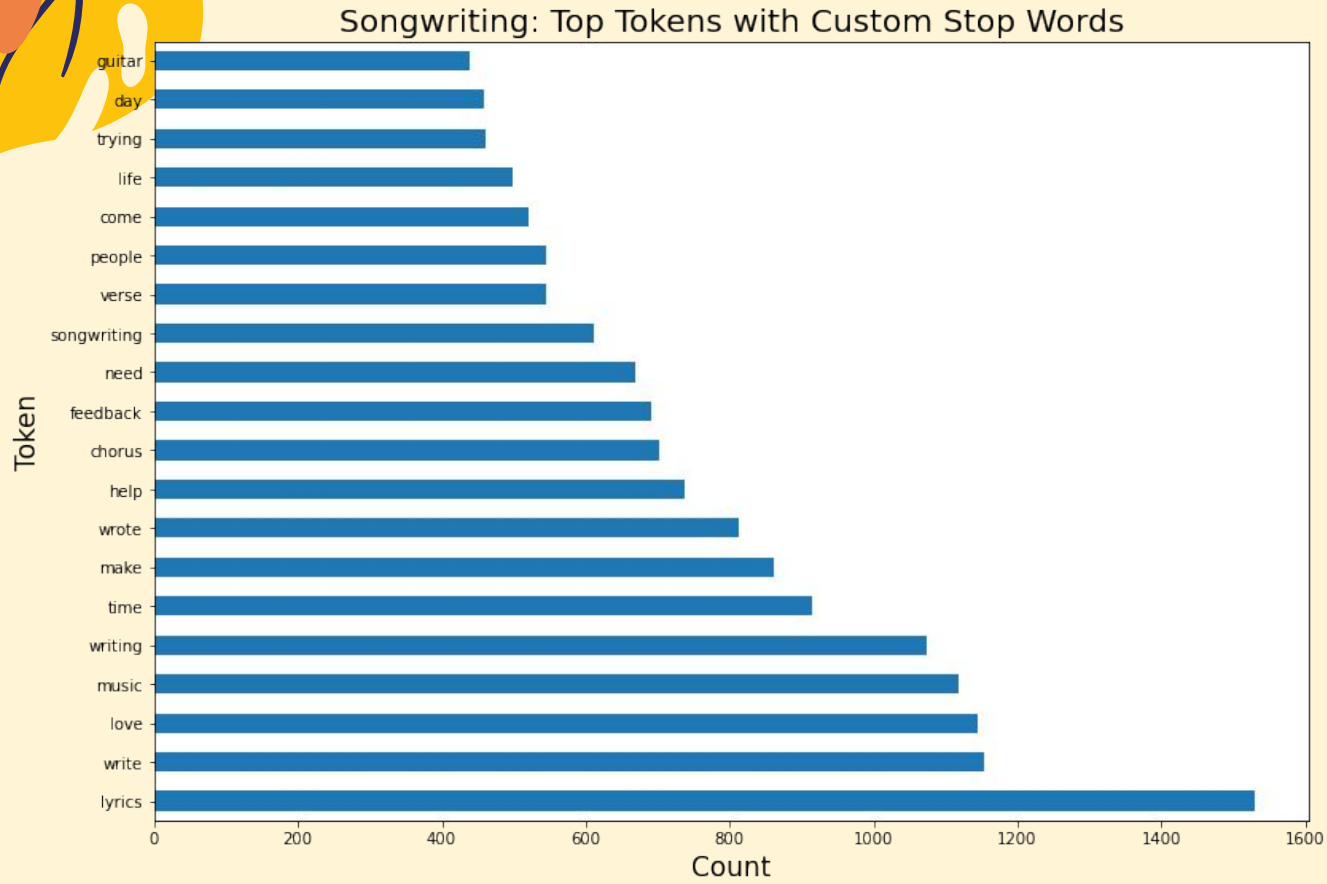
Songwriting

- Also positive!
- Sentiment intensity score of 0.327

Most Common Words



Most Common Words



Modeling

- Naive Bayes
- Random Forest Classifier
- Count Vectorizer & TFIDF Vectorizer
- Baseline Accuracy: about 51% for the majority class

Model Scores

Model	Training Score	Testing Score
Naive Bayes	Count Vectorizer: 97.3% TFIDF: 97.7%	Count Vectorizer: 95.6% TFIDF: 95.7%
Random Forest	Count Vectorizer: 97.3% TFIDF: 97.8%	Count Vectorizer: 92.8% TFIDF: 93.5%

Best Model

Naive Bayes with Count Vectorizer

- Training Score: 97.3% Accuracy
- Testing Score: 95.6% Accuracy
- Misclassification is minimal
- Best Parameters
  - Max features: 25,000
  - N_Gram Range: 1 & 2 grams
  - Stop Words: English
- Perhaps a slight overfitting, but I'm willing to live with that

A Note on Random Forests

	model	score	n_estimators	max_depth	max_features	ccp_alpha
0	Random Forest & TVEC 1	0.925671	100	25	auto	0
1	Random Forest & TVEC 2	0.926819	150	30	auto	0
2	Random Forest & TVEC 3	0.930550	200	35	auto	0
3	Random Forest & TVEC 4	0.933850	200	35	auto	0
4	Random Forest & CVEC 1	0.875592	50	6	NaN	0
5	Random Forest & CVEC 2	0.890658	100	8	NaN	0
6	Random Forest & CVEC 3	0.910890	75	15	auto	0
7	Random Forest & CVEC 4	0.925528	150	25	auto	0
8	Random Forest & CVEC 5	0.926962	100	35	auto	0

Conclusions

- My models were surprisingly effective predictors! I thought that the terminology would be much more similar.
- This distinction is a good thing for aspiring songwriters and recordists! It's easier to find the right forum for your needs.
- My misclassification rate was very low!
- Given more time, I'd like to scrape more posts, including the comments on these posts, and see how that changes my model.

Conclusions

- Let's expand this to more Subreddits! The more communities I pull from, the more my model can grow, and the better it can help direct beginners to the best Subreddit for their needs and goals.

- I'd like to scrape, and perform similar analysis on more "official" sources for recording and songwriting, such as TapeOp and Sound On Sound.

Thanks!

Do you have any questions?

awzucker@gmail.com