Subreddit Prediction Model

Presented by Josh Robin

r/EDM

r/Rock

ACQUIRING THE DATA

EDM DATA

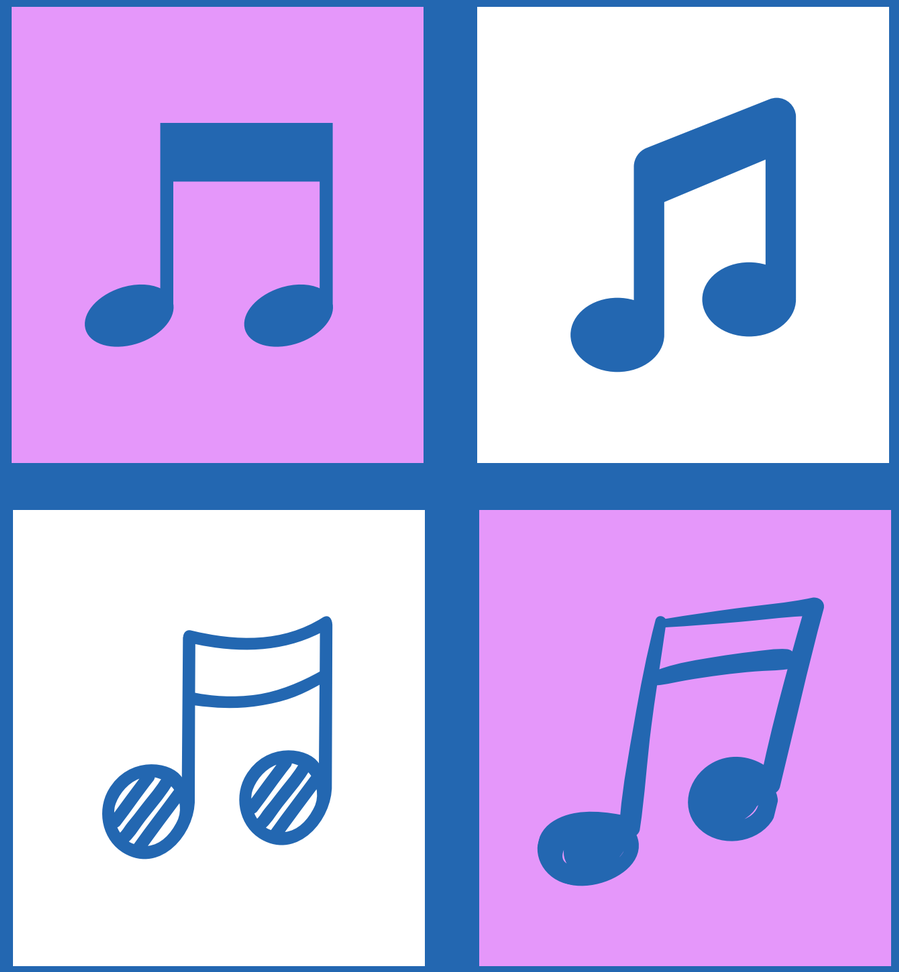
1,911 posts spread out over the past 510 days

ROCK DATA

1,922 posts spread out over the past 4,470 days

DATA MUNGING

- Dropping duplicates
- [removed], \[removed\], [deleted], and NaN
- Merging titles with posts content
- Tokenizing
- Removing URL's
- Removing punctuation
- Lemmatizing



COUNTVECTORIZER

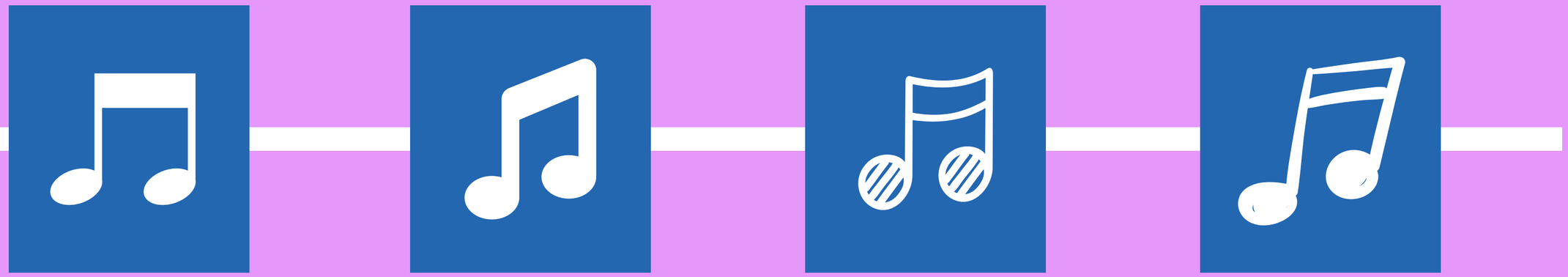
NGRAMS (1, 2)
MAX FEATURES 5000
STOP WORDS = ENGLISH

TF-DIF

NGRAMS (1, 2)
MAX FEATURES 5000
STOP WORDS = ENGLISH

MODELING

(CLASSIFICATION)



NAIVE BAYES

LOGISTIC REGRESSION

RANDOM FOREST

K NEAREST NEIGHBOR

Results

NAIVE BAYES	LOGISTIC REGRESSION	RANDOM FOREST	K NEAREST NEIGHBOR
BEST VECTORIZER:	BEST VECTORIZER:	BEST VECTORIZER:	BEST VECTORIZER:
TF-DIF	TF-DIF	CountVectorizer	CountVectorizer
TRAINING SCORE:	TRAINING SCORE:	TRAINING SCORE:	TRAINING SCORE:
90.4%	88.7%	87.3	73.2
TESTING SCORE:	TESTING SCORE:	TESTING SCORE:	TESTING SCORE:
88.5%	86.4%	84.6%	70.2

BEST PREDICTORS