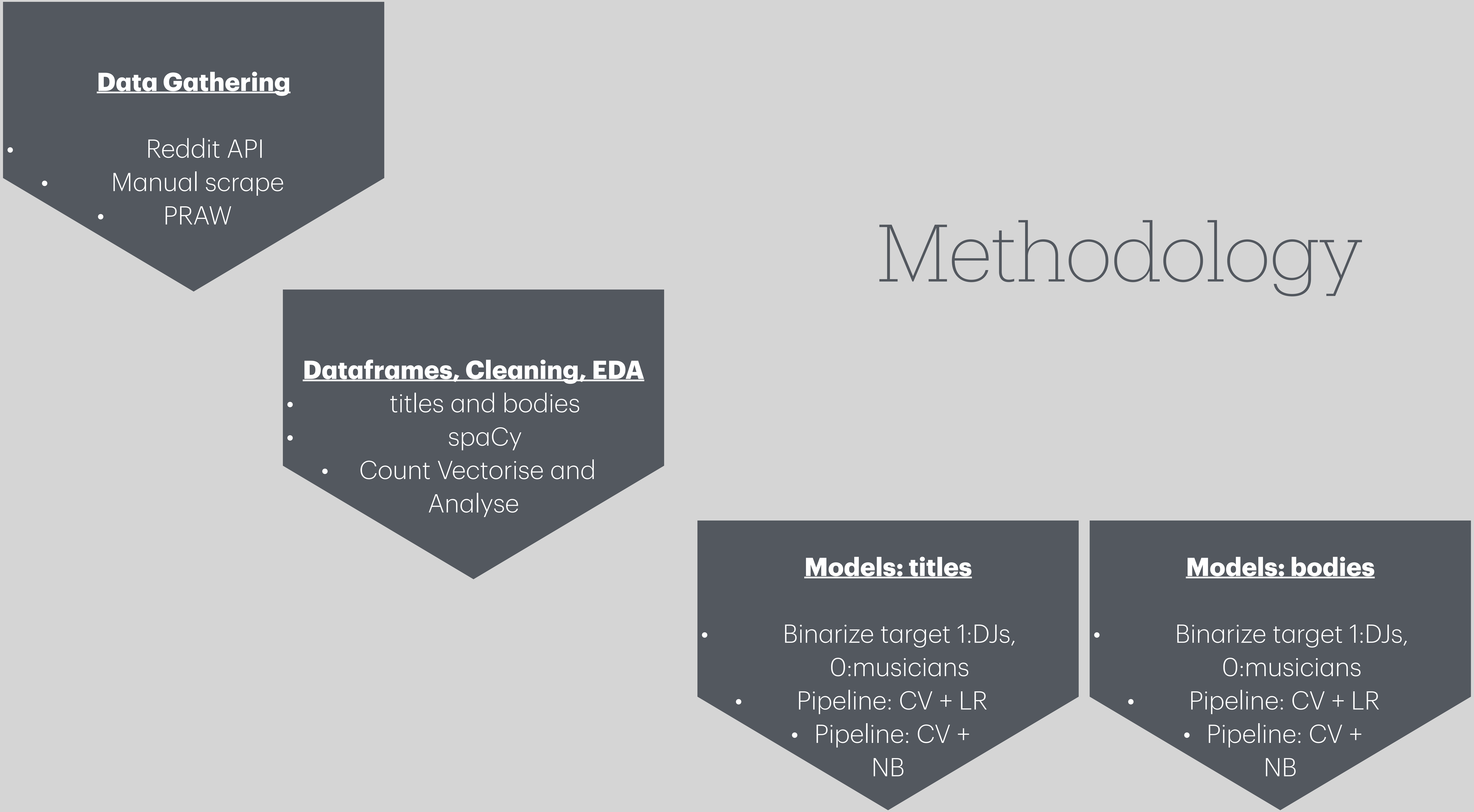
Classifying Reddit Posts : DJ's vs Musicians

A project by Prab Jaswal

Problem, Scenario, Context

- The growing divide between new entrants into the “live music” industry: DJ’s vs Musicians.
- Can we classify whether a post comes from either the r/DJs or r/musicians subreddit?
- Models: Logistic Regression vs Multinomial Naive Bayes
- Primary evaluation metric: ACCURACY score (consideration given to F1 score).

Methodology



Data Gathering

- Reddit API
- Manual scrape
- PRAW

Dataframes, Cleaning, EDA

- titles and bodies
- spaCy
- Count Vectorise and Analyse

Models: titles

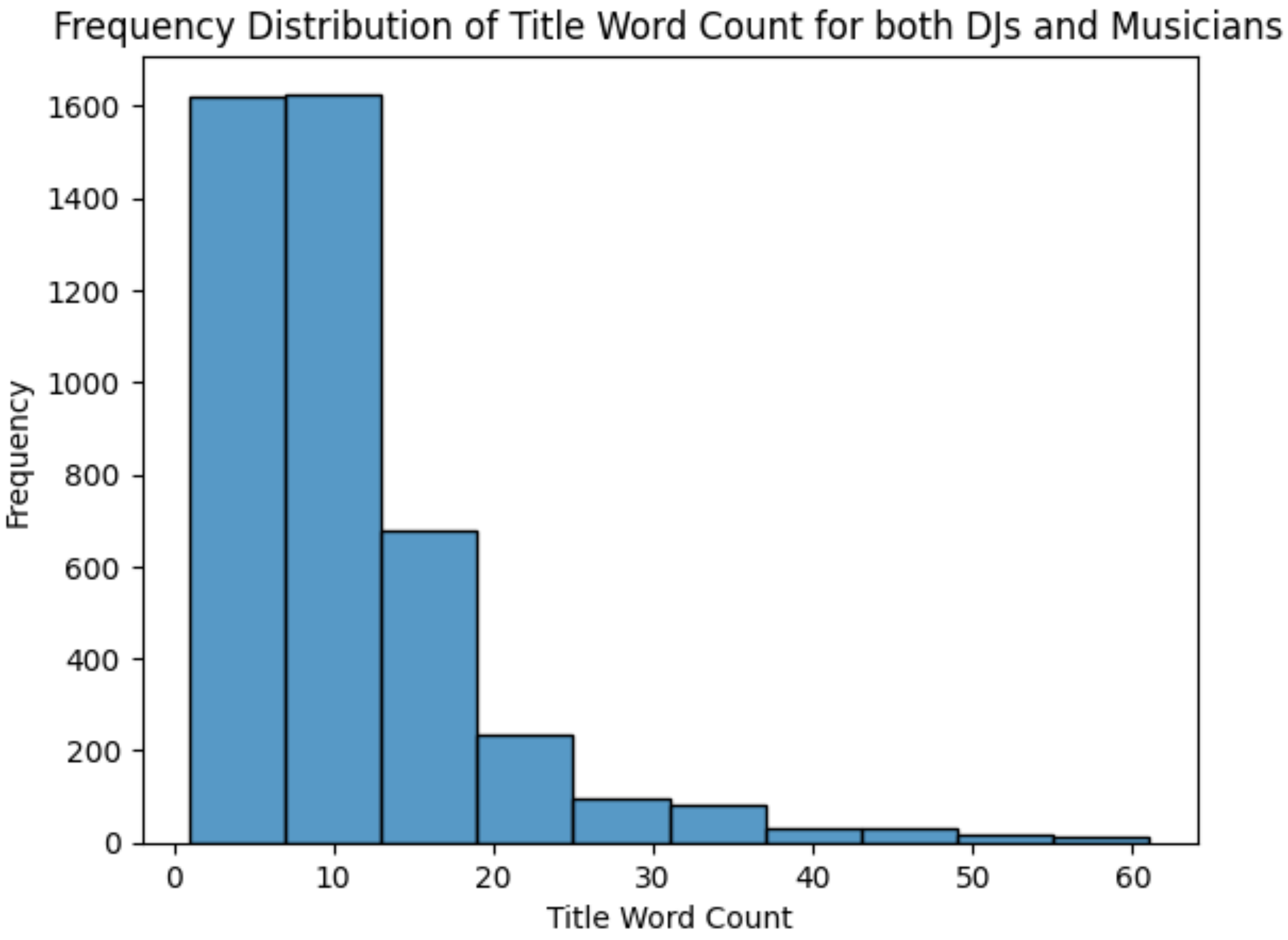
- Binarize target 1:DJs, 0:musicians
- Pipeline: CV + LR
- Pipeline: CV + NB

Models: bodies

- Binarize target 1:DJs, 0:musicians
- Pipeline: CV + LR
- Pipeline: CV + NB

Exploring the data: titles

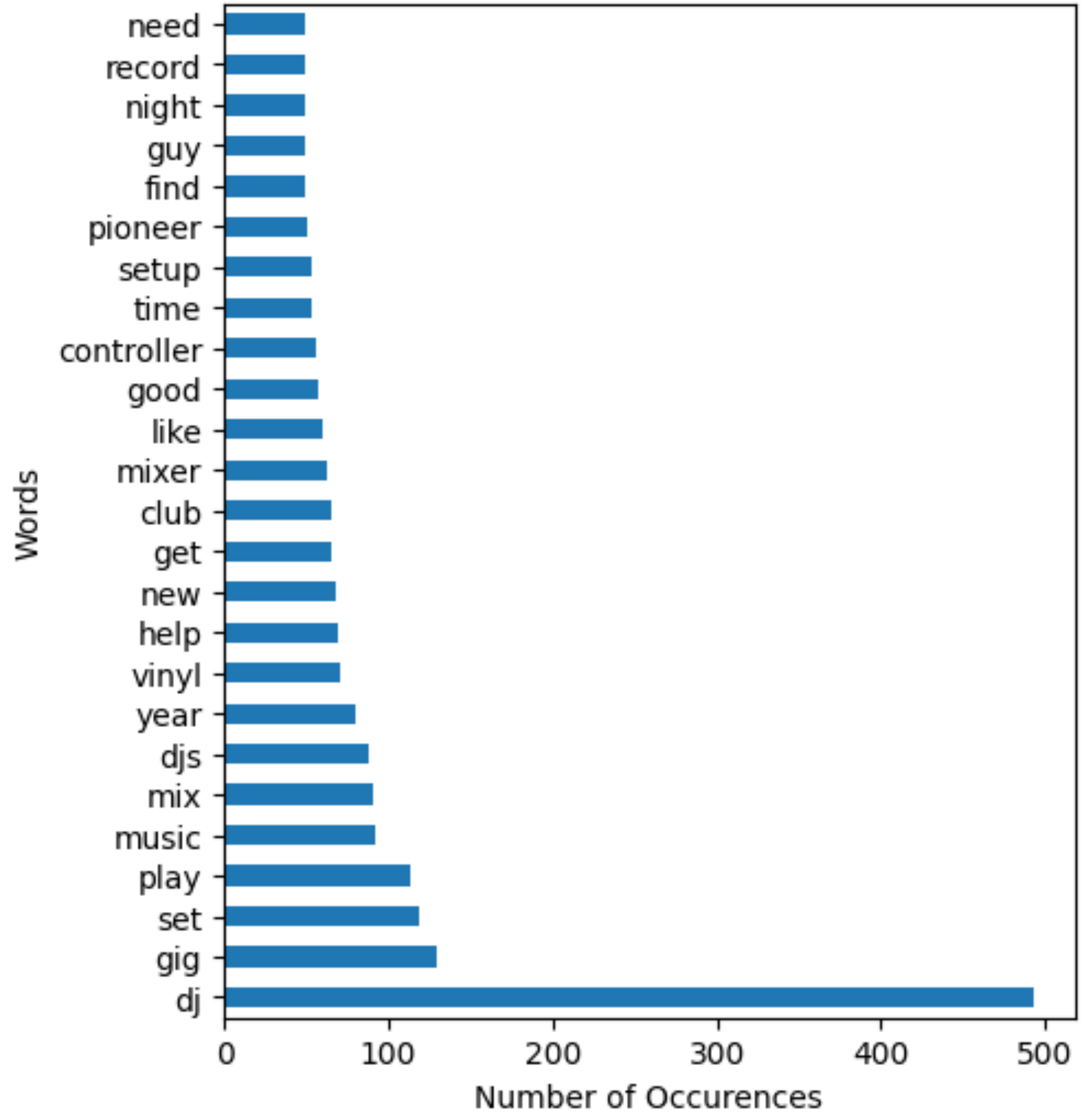
Word Count


Frequency Distribution of Title Word Count for both DJs and Musicians

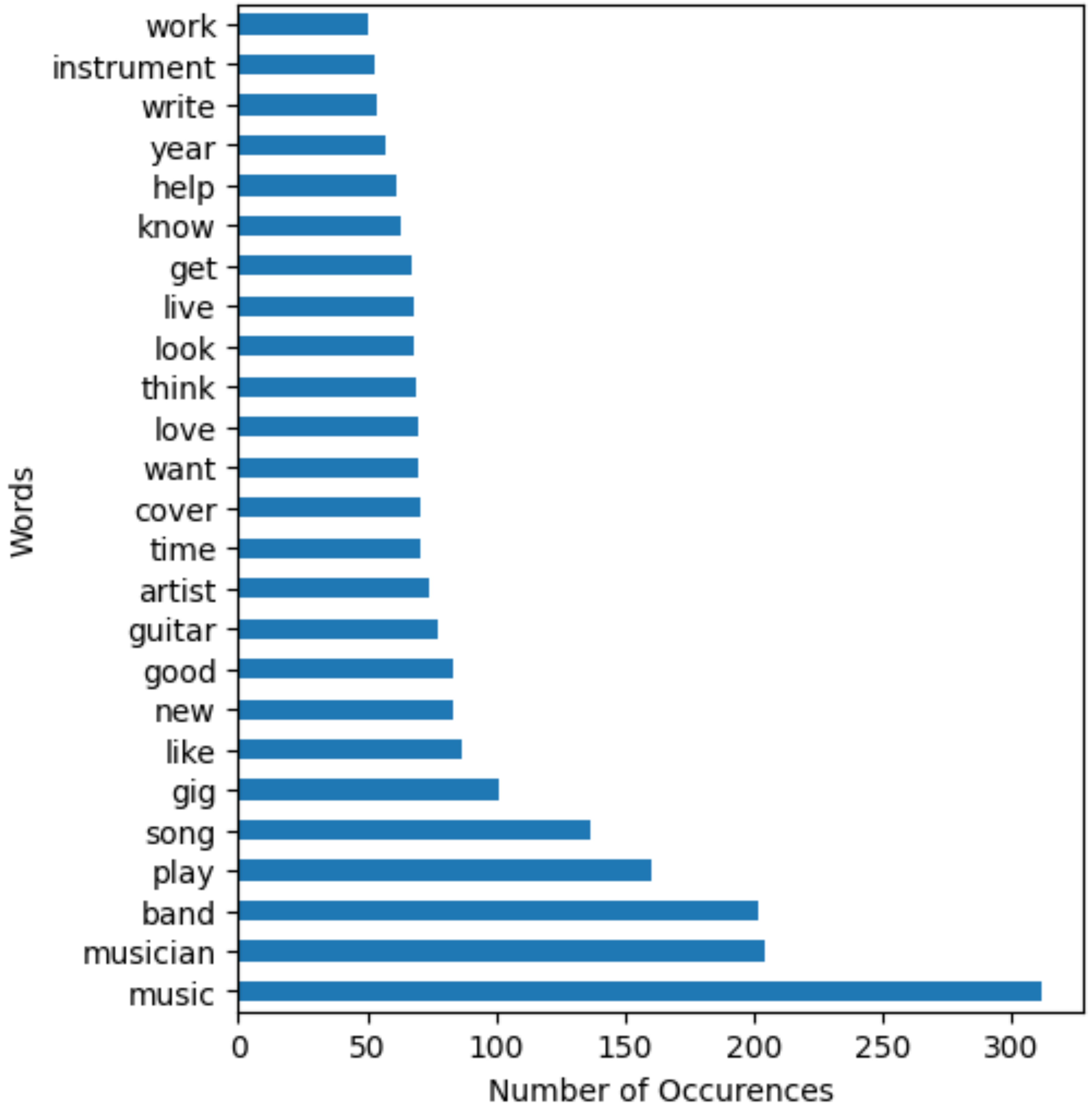
subreddit	Count	Mean word length	minimum word length	maximum word length
r/DJs (1)	2294	10.08	1	60
r/ musicians (0)	2136	10.10	1	61

Top Occurring Words

Occurences of the top 25 most frequently occuring words in r/DJs subreddit titles

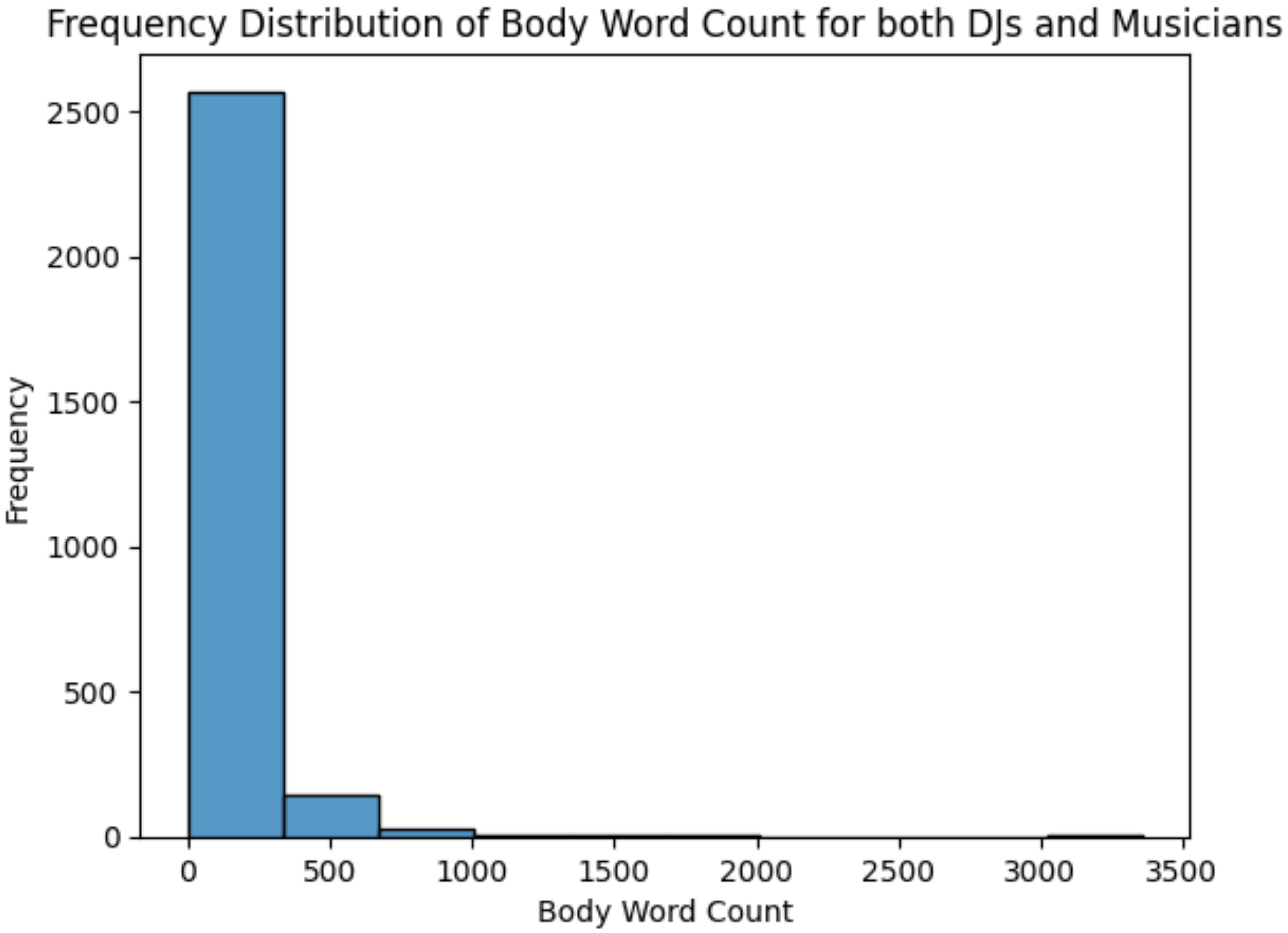


Occurences of the top 25 most frequently occuring words in r/musicians subreddit titles



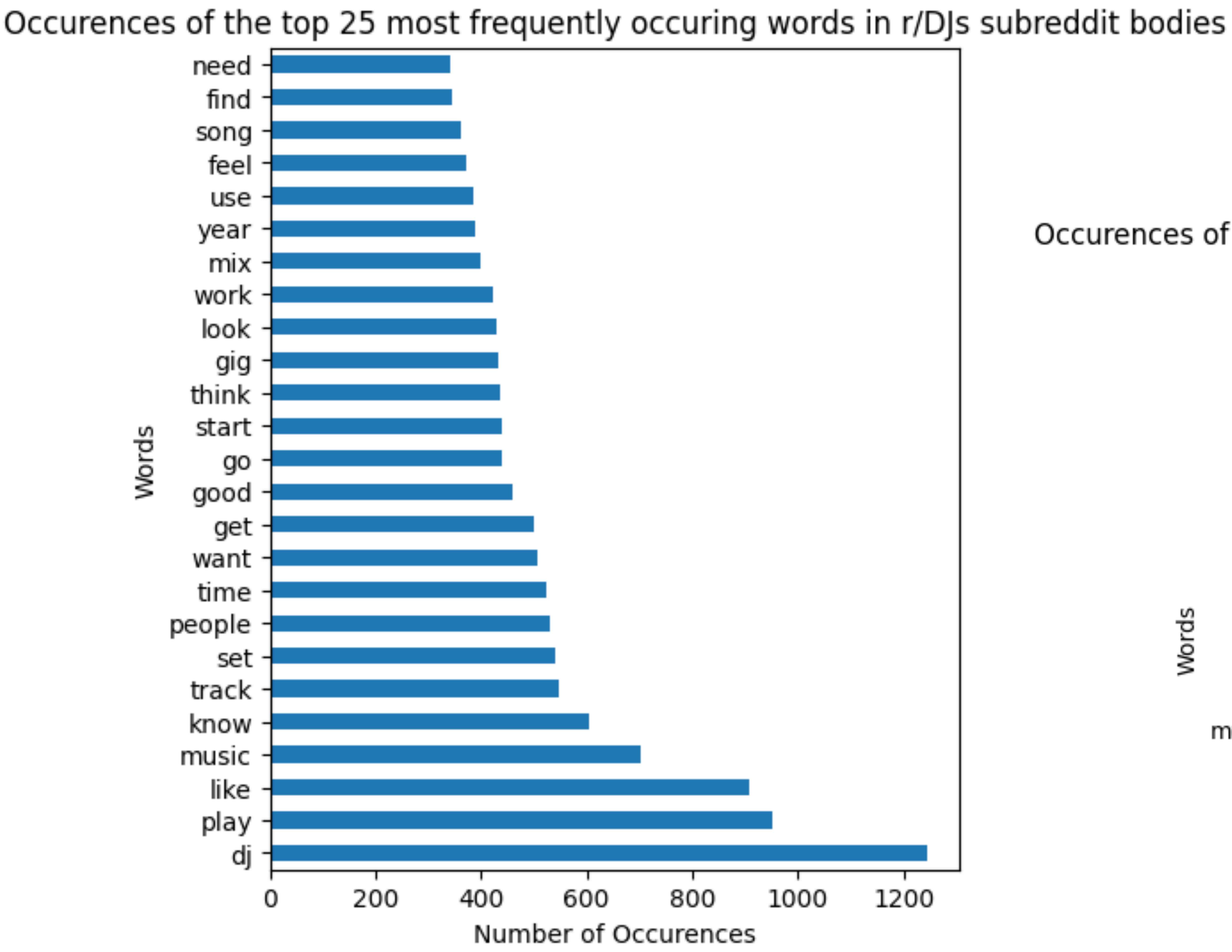
Exploring the data: bodies

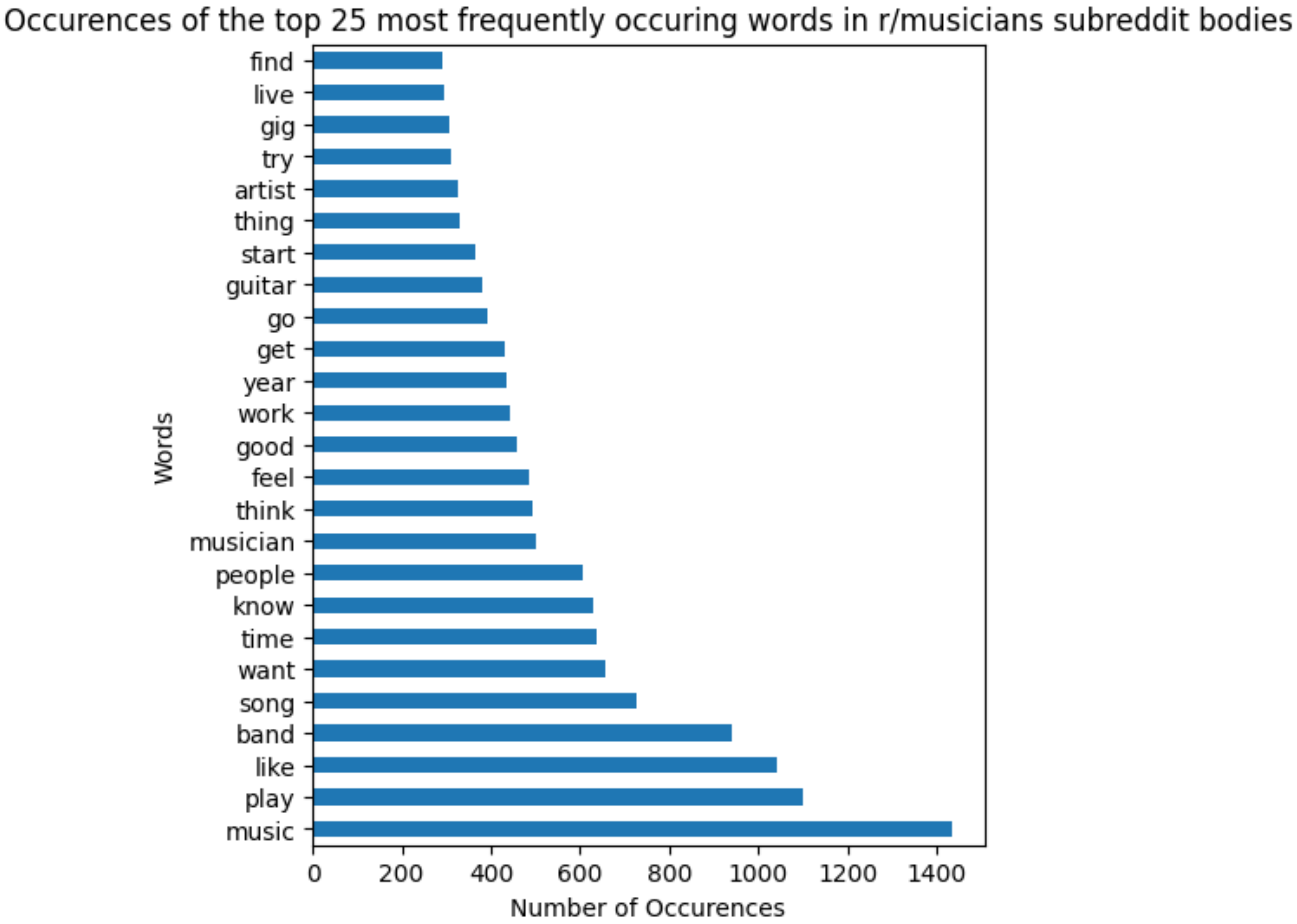
Word Count



subreddit	Count	Mean word length	minimum word length	maximum word length
r/DJs (1)	1412	135.41	1	1903
r/ musicians (0)	1342	135.37	0	3358

Top Occurring Words

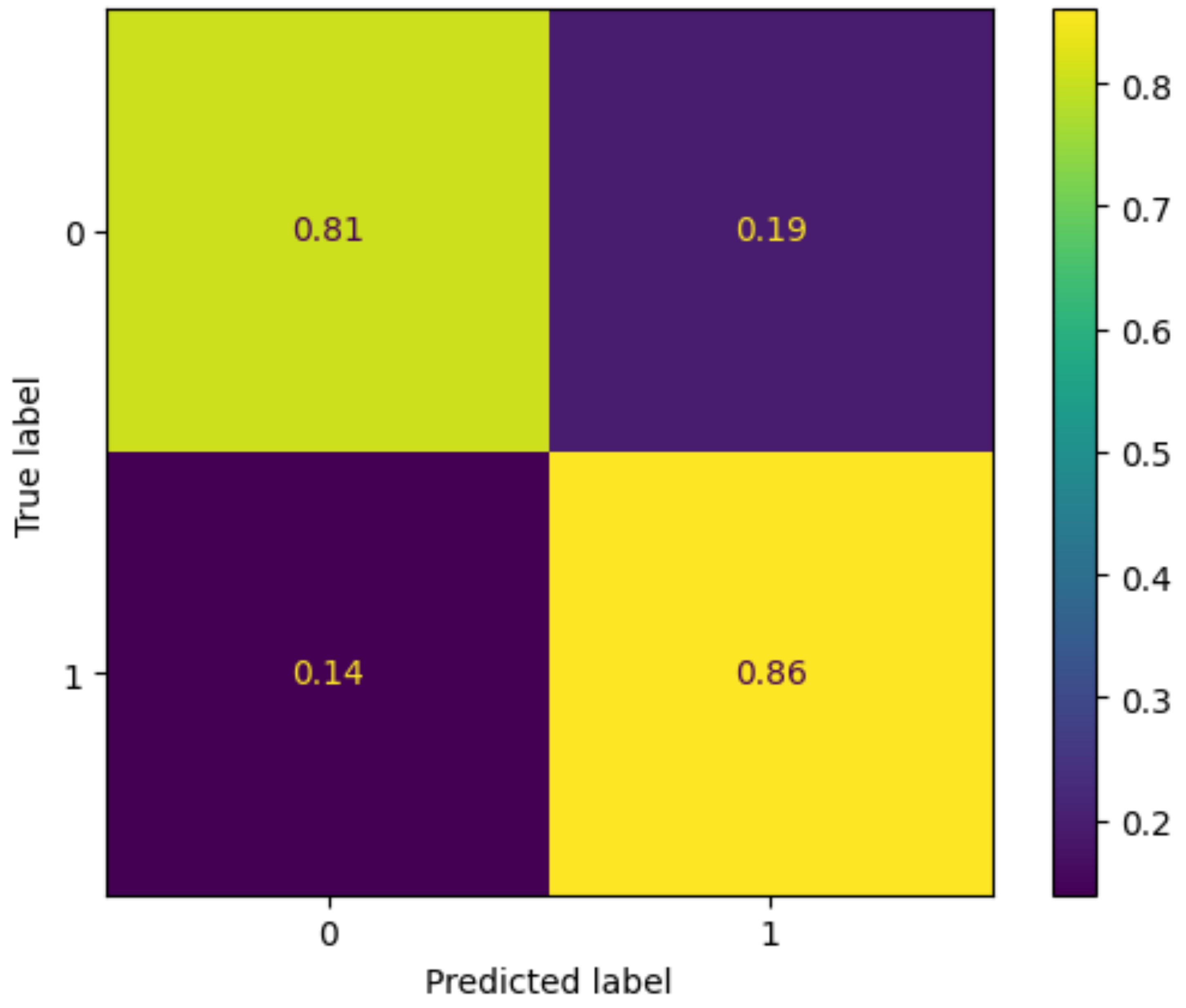




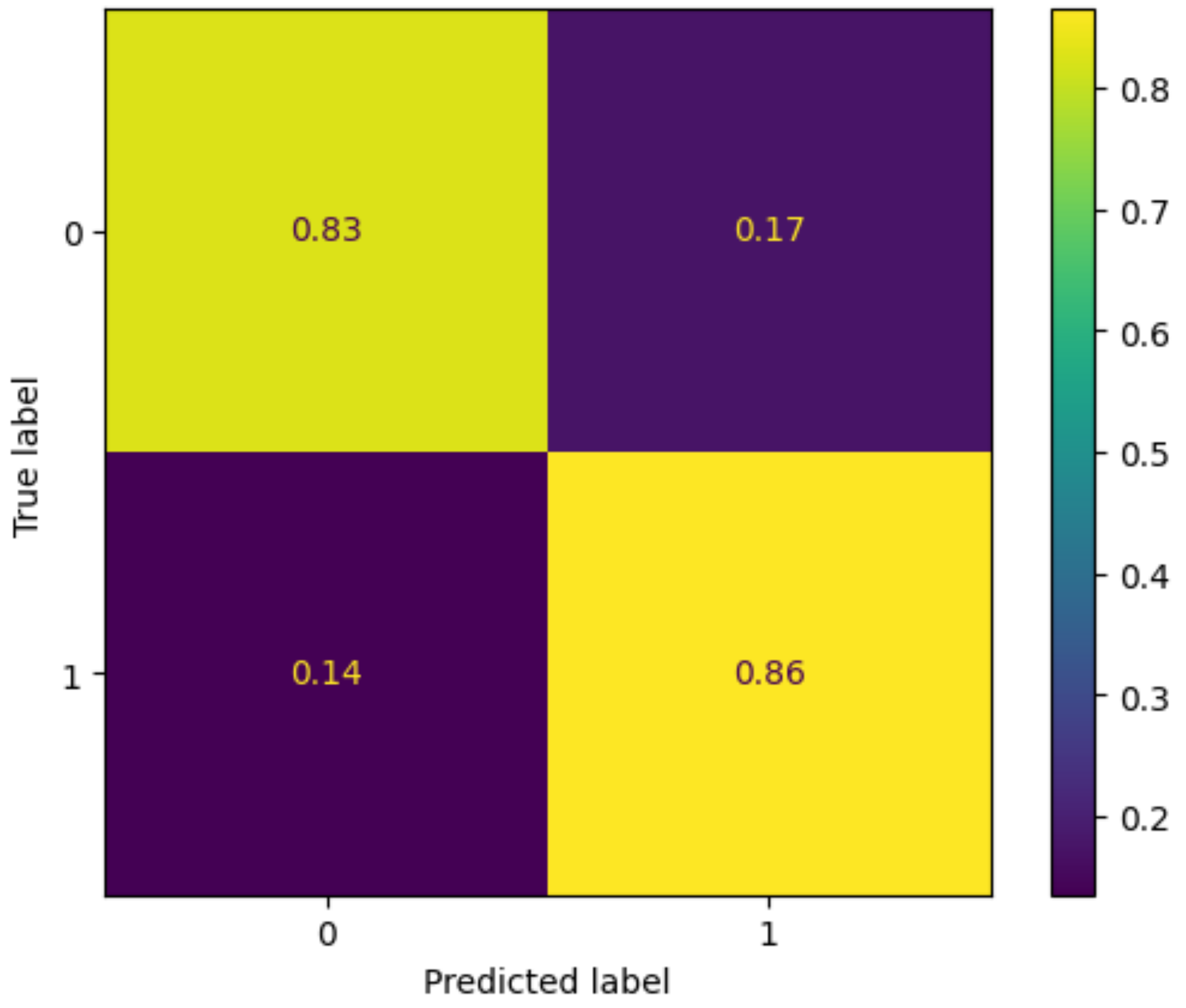
Models Titles

Logistic Regression vs Naive Bayes

- Logistic Regression Accuracy - 0.94 training, 0.83 testing (overfit).



- Naive Bayes Accuracy - 0.90 training, 0.85 testing (less overfit).

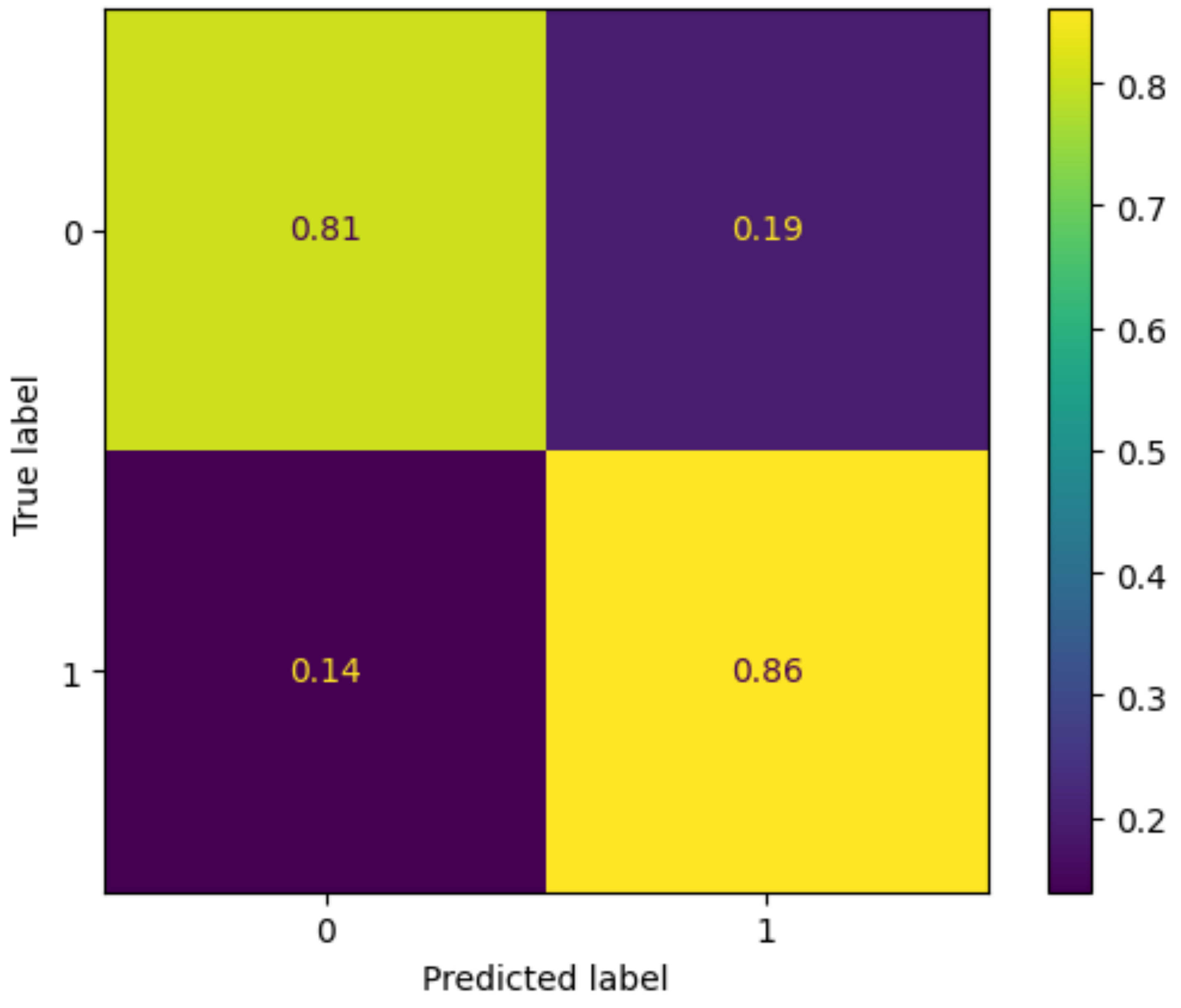


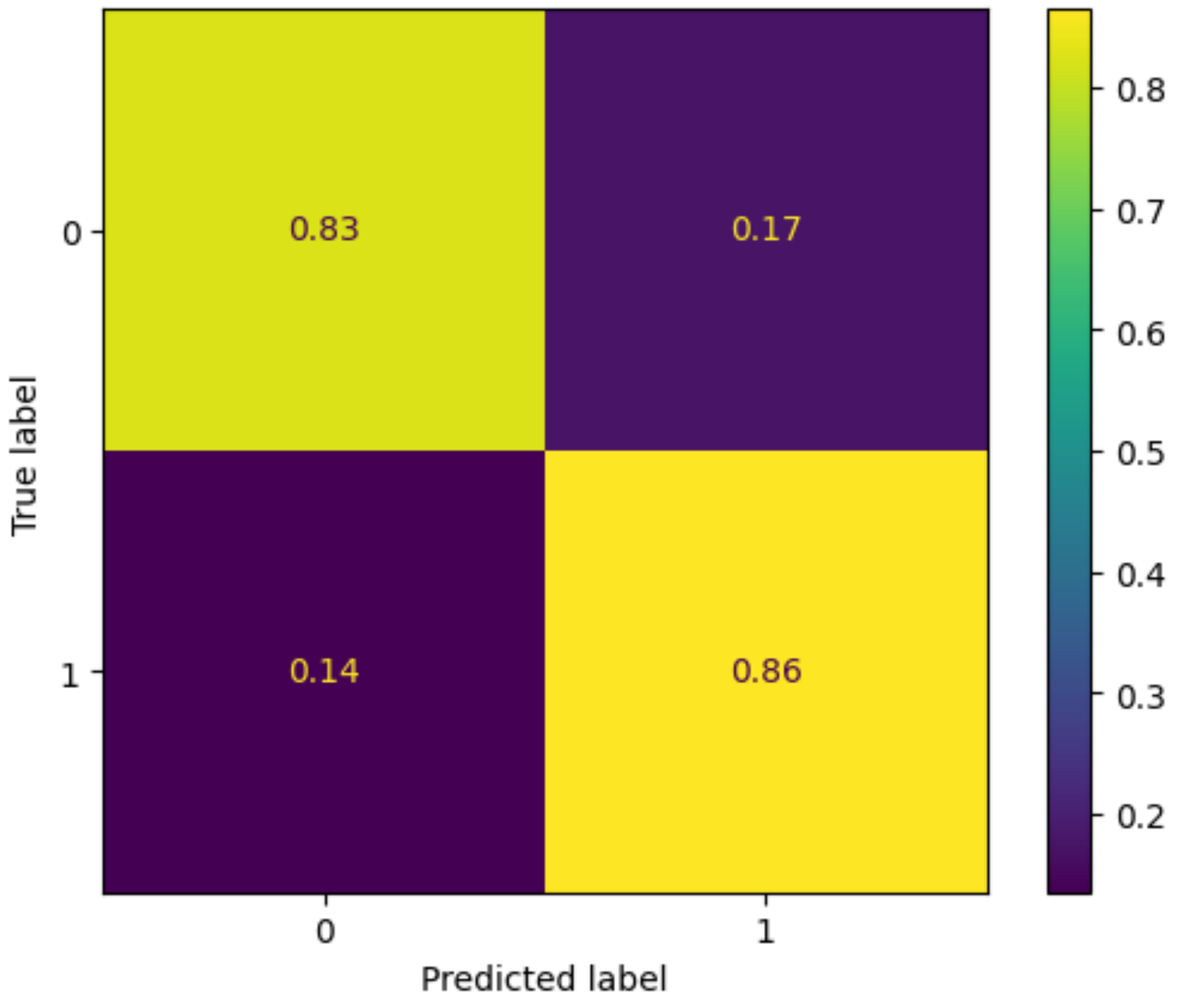
Models Bodies

Logistic Regression vs Naive Bayes

- Logistic Regression Accuracy - 1.00 training, 0.92 testing (overfit).

- Naive Bayes Accuracy - 0.96 training, 0.92 testing (less overfit).





Questions