Music Through Ages and Genres

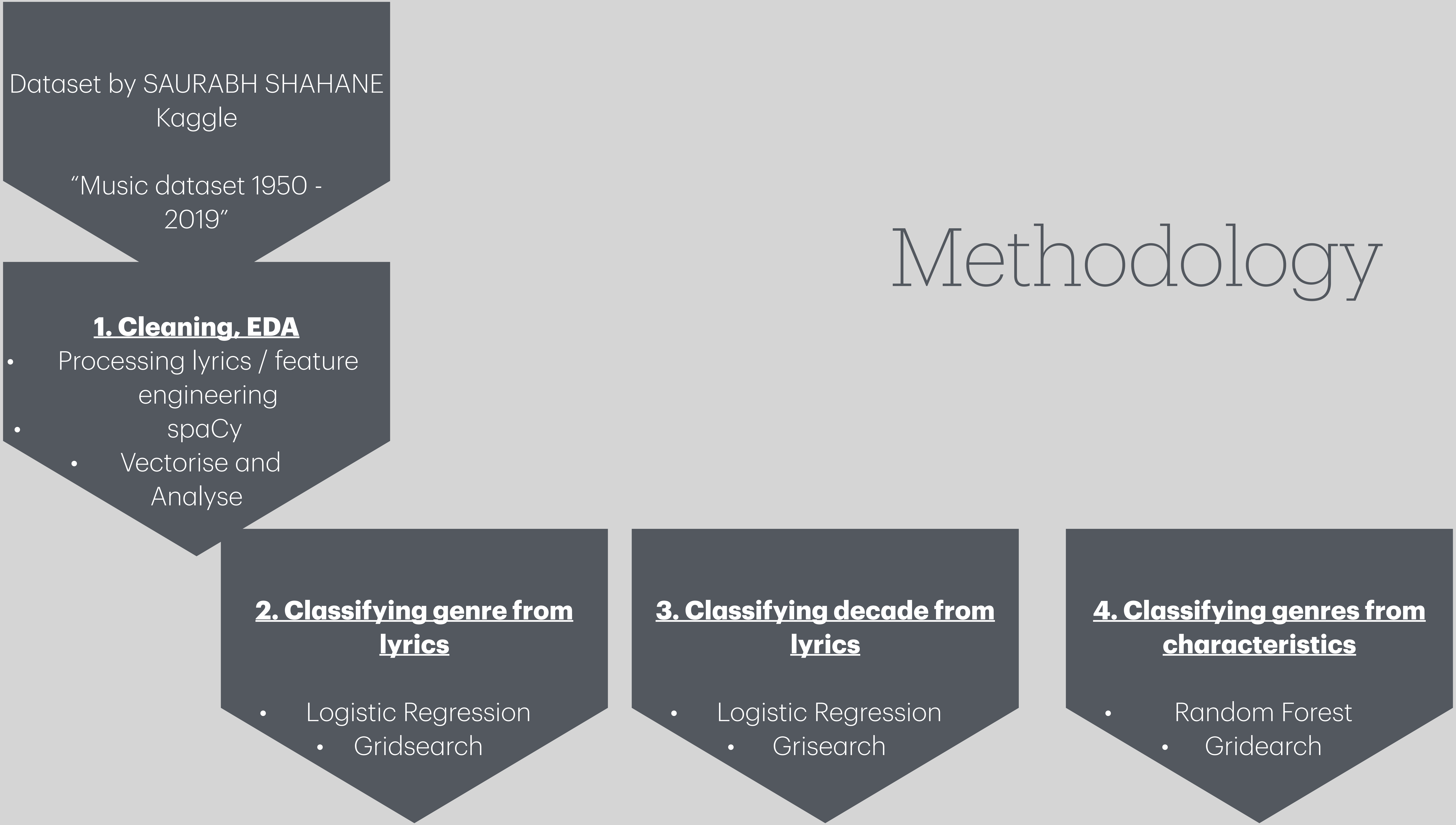
A project by Prab Jaswal

disclaimer: words of profanity contained

Problem, Scenario, Context

- Explore how the characteristics of music change over time and by genre
- 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



Dataset by SAURABH SHAHANE
Kaggle

“Music dataset 1950 - 2019”

1. Cleaning, EDA

- Processing lyrics / feature engineering
- spaCy
- Vectorise and Analyse

2. Classifying genre from lyrics

- Logistic Regression
- Gridsearch

3. Classifying decade from lyrics

- Logistic Regression
- Grisearch

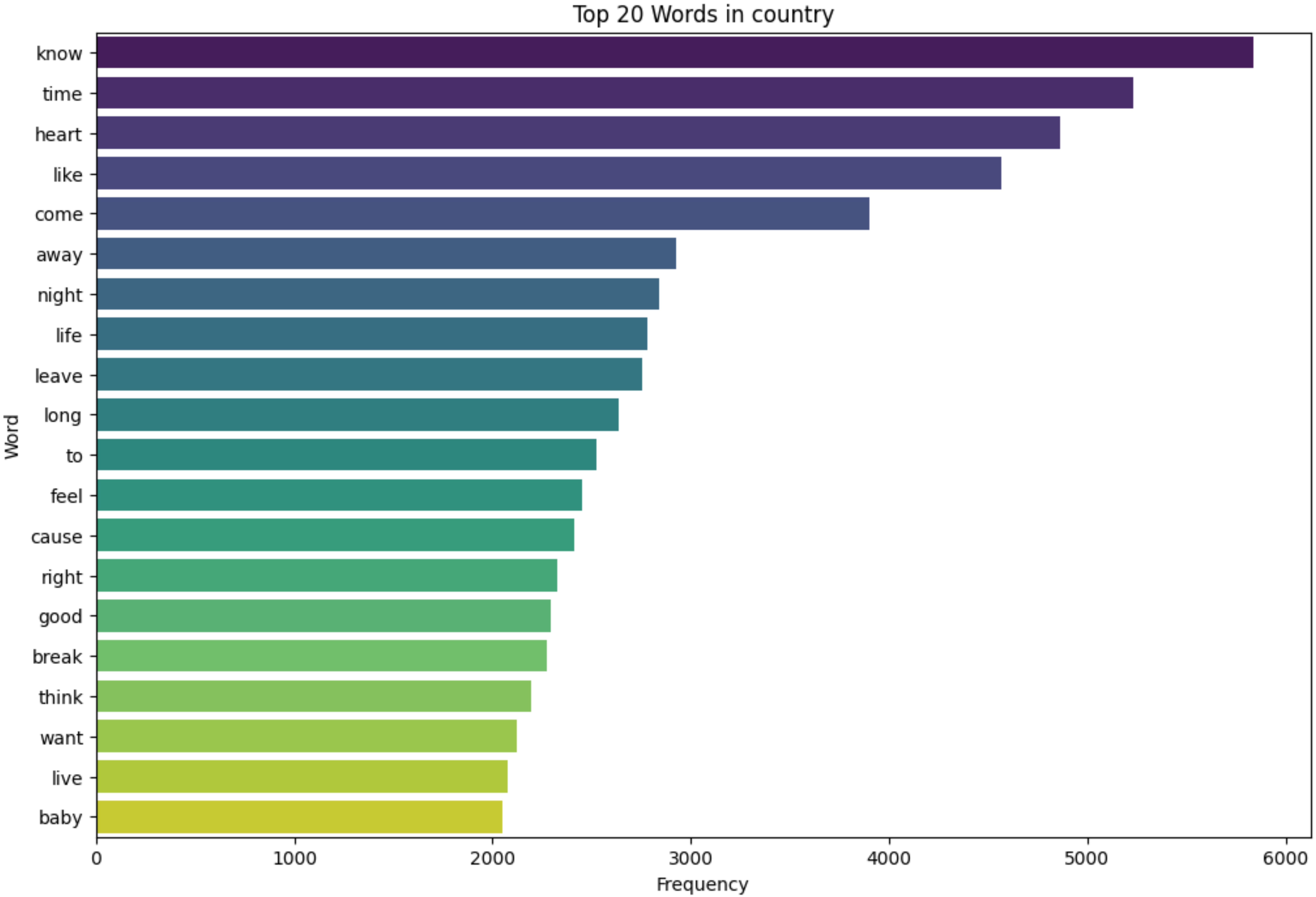
4. Classifying genres from characteristics

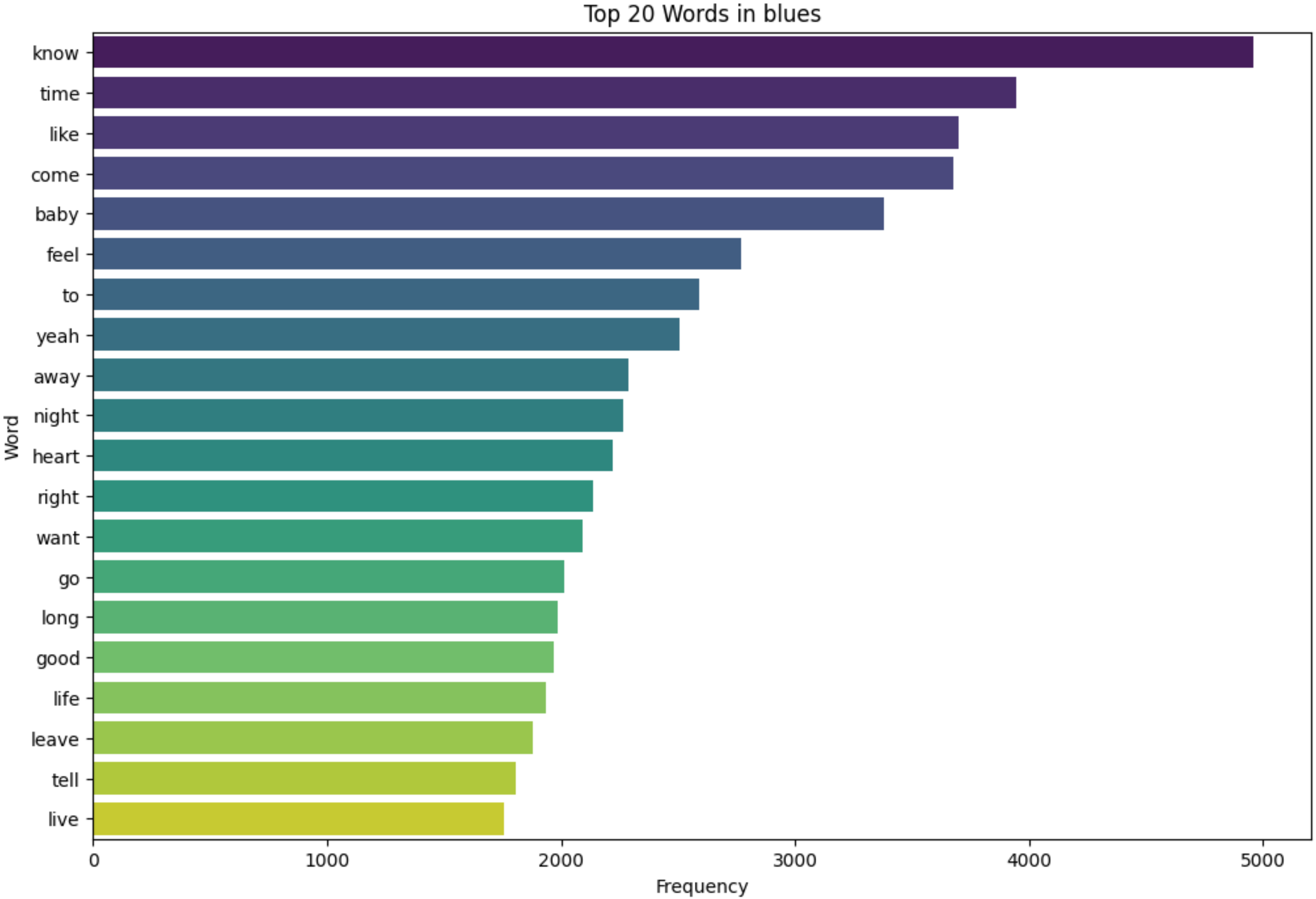
- Random Forest
- Gridearch

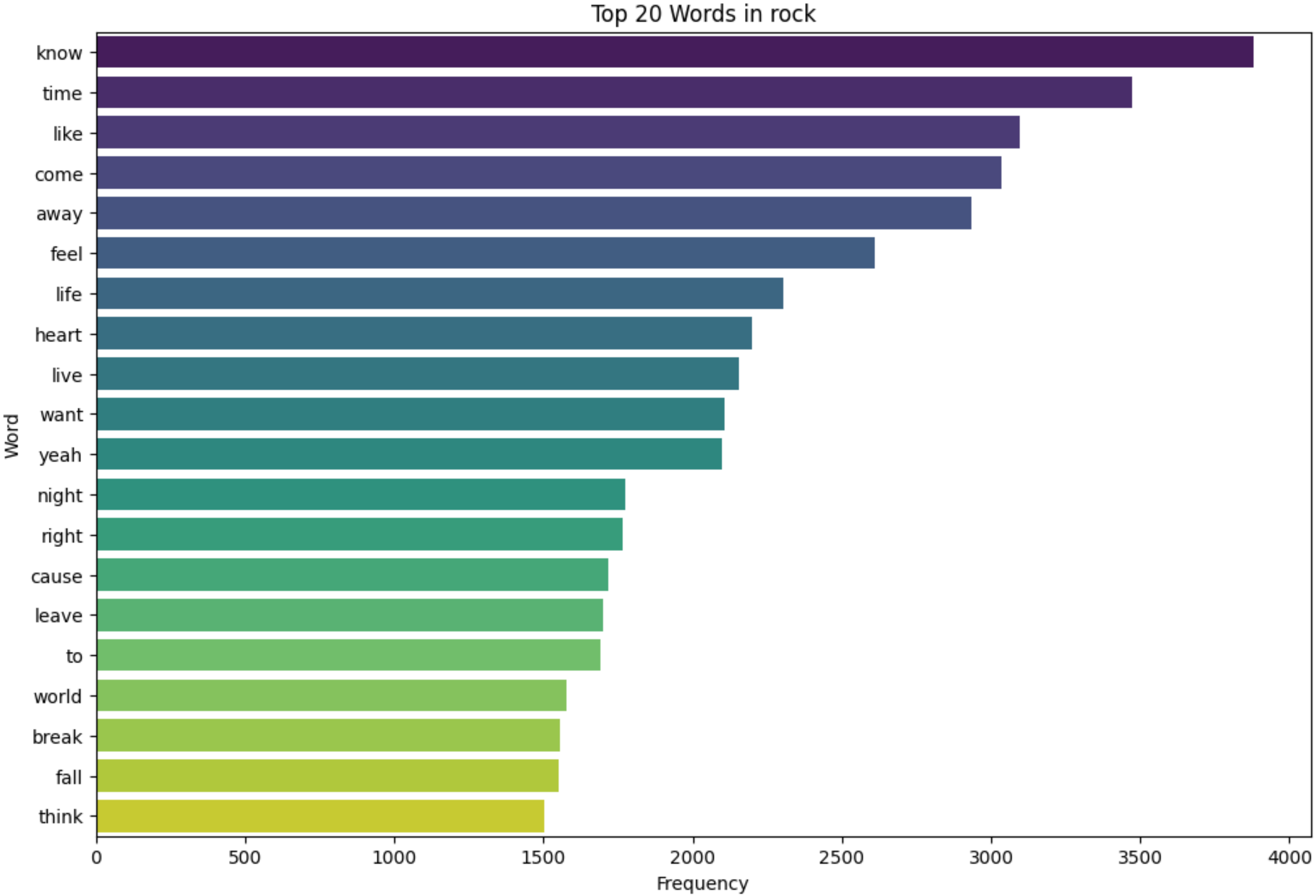
Exploring the data: Over Time

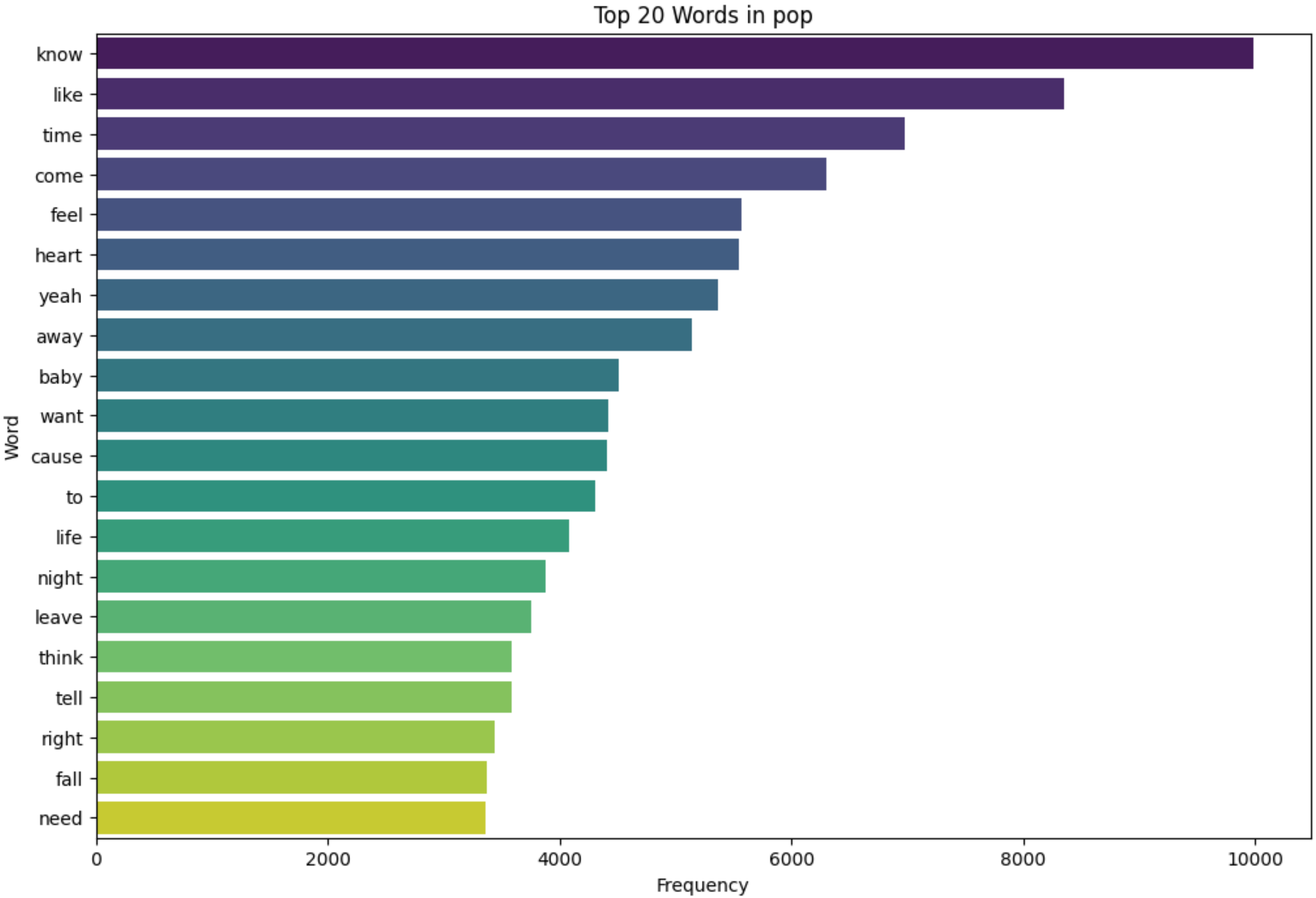
- Songs have gotten more electronic (less acoustic)
- Songs are getting shorter in length
- Songs are getting louder
- Songs are (as percieved at the point of data collection) getting more energetic as nightlife / club culture evolves.

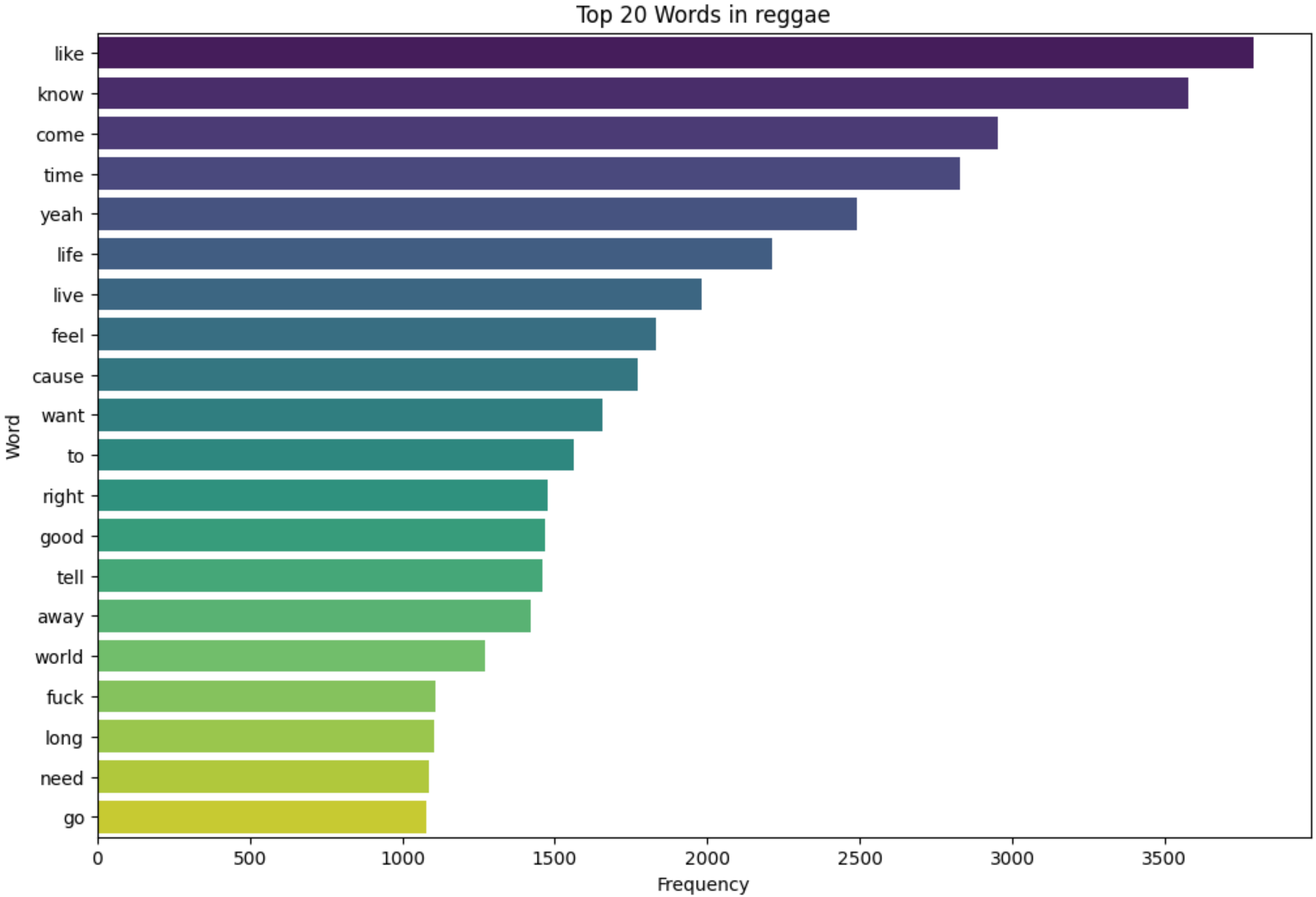
Exploring the data: by genre

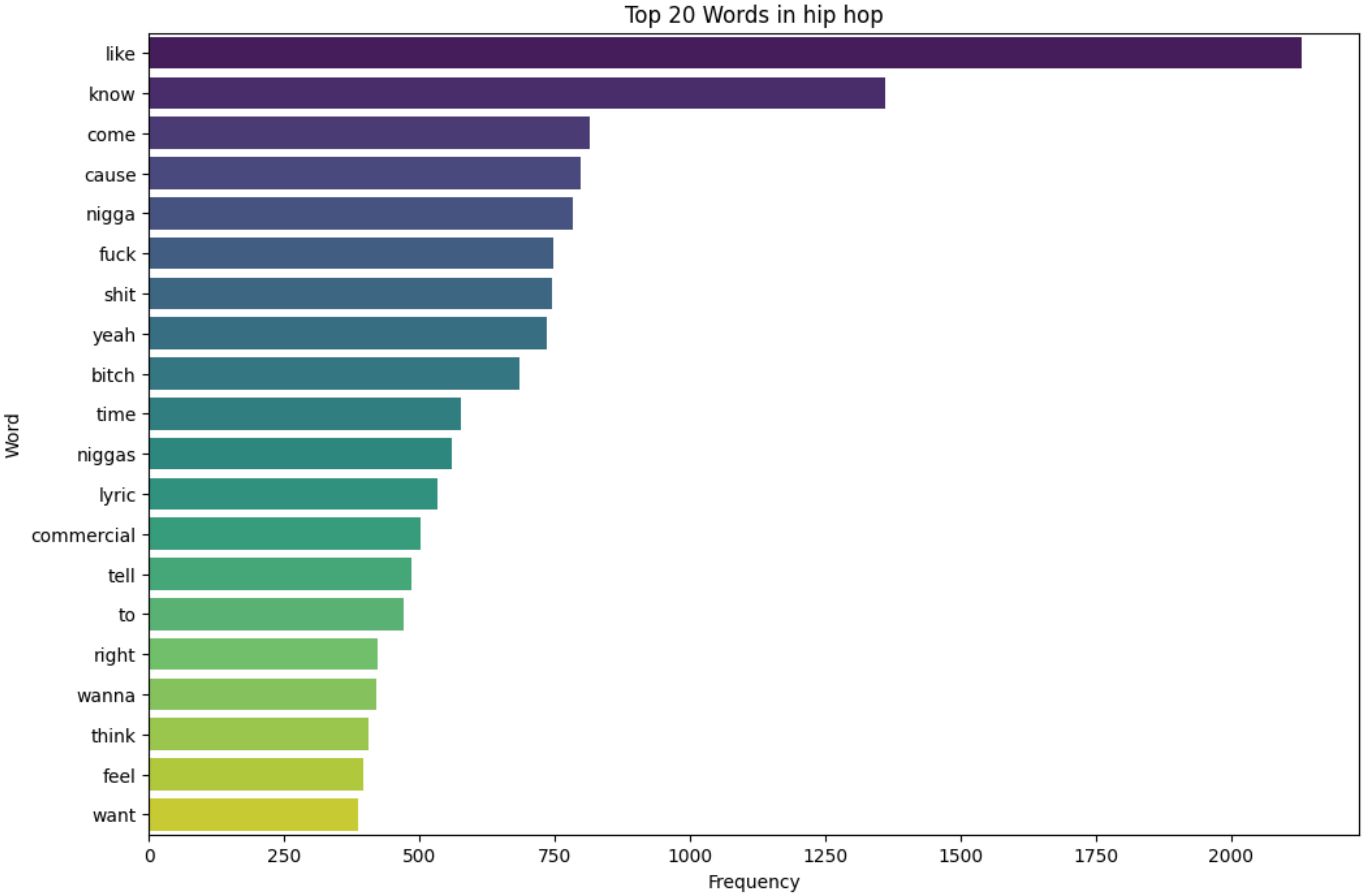

Top 20 Words in country
know
time
heart
like
come
away
night
life
leave
long
to
feel
cause
right
good
break
think
want
live
baby
Word
0
1000
2000
3000
4000
5000
6000
Frequency


Top 20 Words in blues
Word
know
time
like
come
baby
feel
to
yeah
away
night
heart
right
want
go
long
good
life
leave
tell
live
0
1000
2000
3000
4000
5000
Frequency


Top 20 Words in rock
know
time
like
come
away
feel
life
heart
live
want
yeah
night
right
cause
leave
to
world
break
fall
think
Word
0
500
1000
1500
2000
2500
3000
3500
4000
Frequency


Top 20 Words in pop
know
like
time
come
feel
heart
yeah
away
baby
want
cause
to
life
night
leave
think
tell
right
fall
need
Word
0
2000
4000
6000
8000
10000
Frequency


Top 20 Words in reggae
like
know
come
time
yeah
life
live
feel
cause
want
to
right
good
tell
away
world
fuck
long
need
go
Word
0
500
1000
1500
2000
2500
3000
3500
Frequency


Top 20 Words in hip hop
like
know
come
cause
nigga
fuck
shit
yeah
bitch
time
niggas
lyric
commercial
tell
to
right
wanna
think
feel
want
Word
0
250
500
750
1000
1250
1500
1750
2000
Frequency

Exploring the data: genres by characteristics

- Hip-hop has highest scores in characteristics of Obscenity and Danceability
- Jazz leads in acousticness and instrumentalness with hip-hop being the lowest of these
- Country leads in sadness
- Rock leads in energy and violence.

Models - Classification

- Classifying genres from lyrics
- Classifying decade from lyrics
- Classifying genres from characteristics
- Imbalanced Classes (challenge) - F1 Score

Next Steps

- Trying different models (e.g., Random Forest, SVM, Neural Networks).
- Engineering new features or using different text vectorization methods (e.g., Word2Vec, GloVe).
- Increasing the dataset size or improving the quality of the data through webscraping lyrics for songs.
- Given further time to undestand ROC AUC curve, I would also incorporate this as a metric for model evaluation however my current understanding at this time is too rocky for it to be a useful metric in this project.

I welcome a discussion if you choose!