WG?

Text Mining & Natural Language Processing

Chibuzor John Amadi & Kristian Perriu

July 20, 2023

Summary

01. Intro

Introduction of what this project is about.

02. Dataset

Introduction and exploration of the dataset.

03. The model

A brief description about the model

04. Evaluation

Evaluation of the model.

05. Conclusion

Conclusion about the model.

Introduction

What is it about?

- The aim of this project is to predict what genere a song belongs to only by the lyrics of a song.
- The WG? is an acronym for What Genere? Which is exactly what this project is about.

Dataset

What to expect?

- Over 10k lyrics
- 4 different genres
- 120 different artists
- No remixes and no live versions of songs.

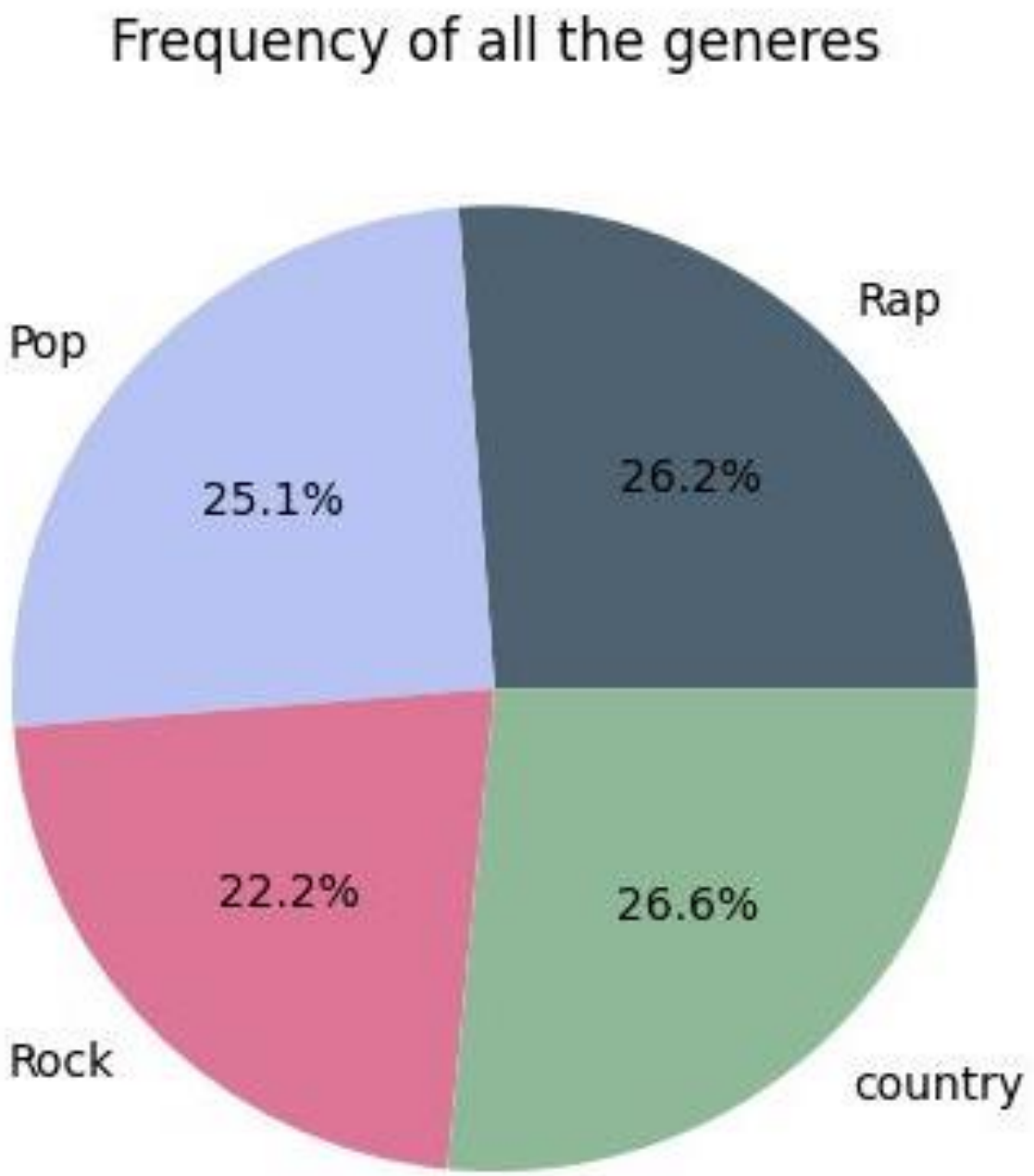
About our dataset

- Created by us using the Genius API.
- Cleaned from special characters.
- Structure of the songs preserved.

Data analysis

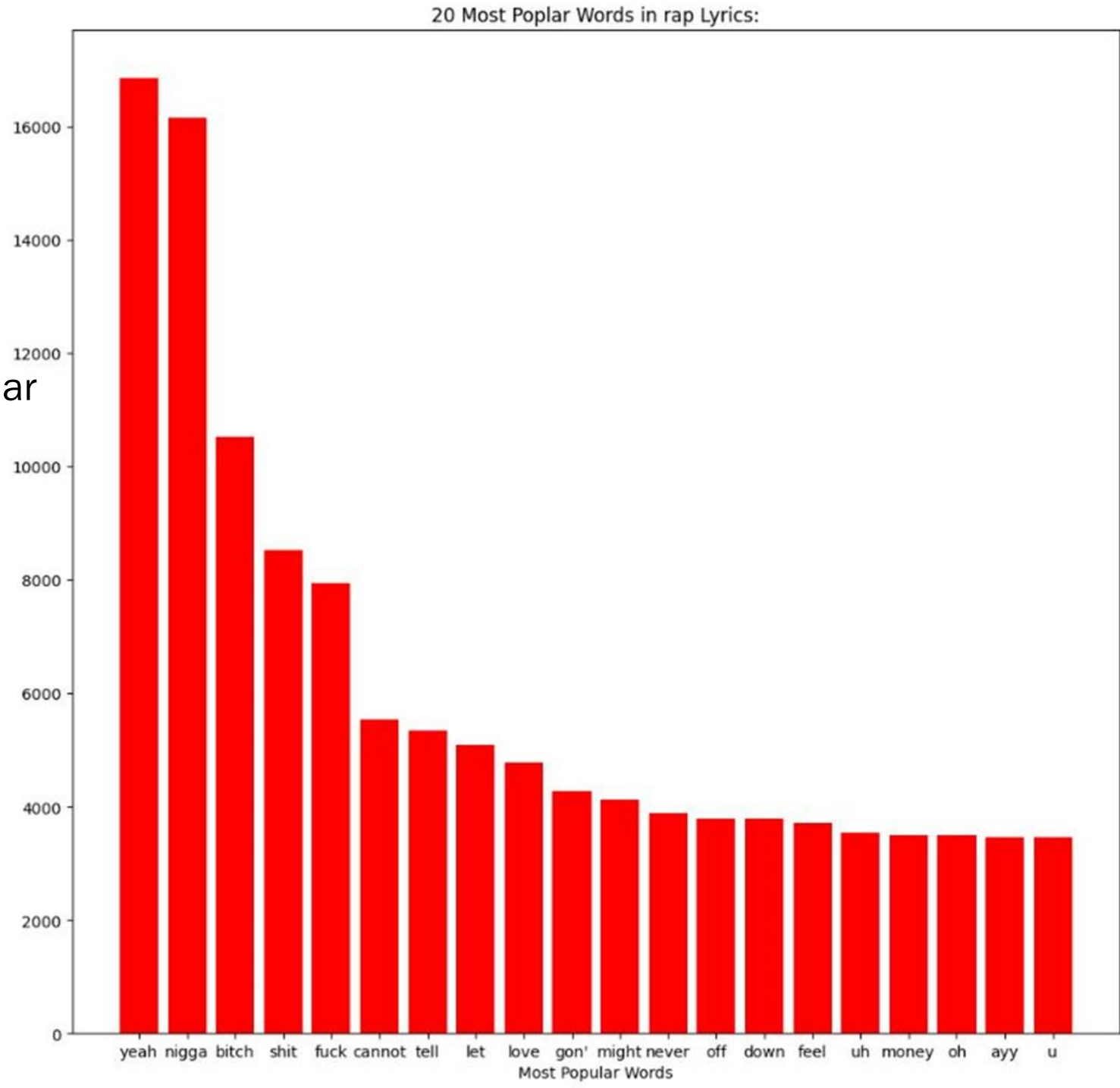
Target data

- 4 different categories
- Decently balanced.
- 2 categories slightly dominate over the others.



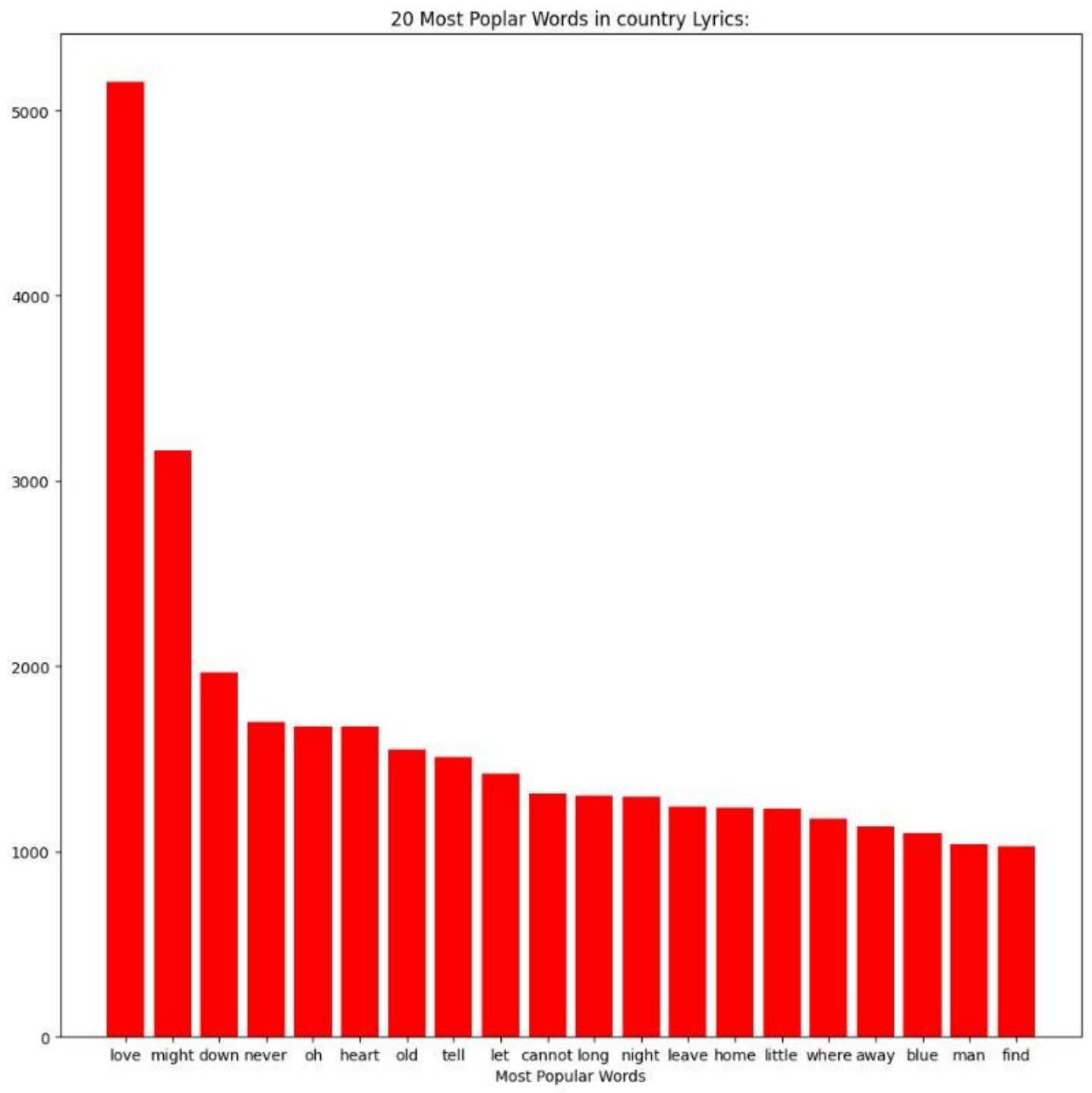
Top 20 rap

- Here we have the top 20 most popular words used In rap songs.
- Very special words that are not commonly used on the other songs.
- One of the reasons why the model performs exceptionally better in rap compared to other genres.



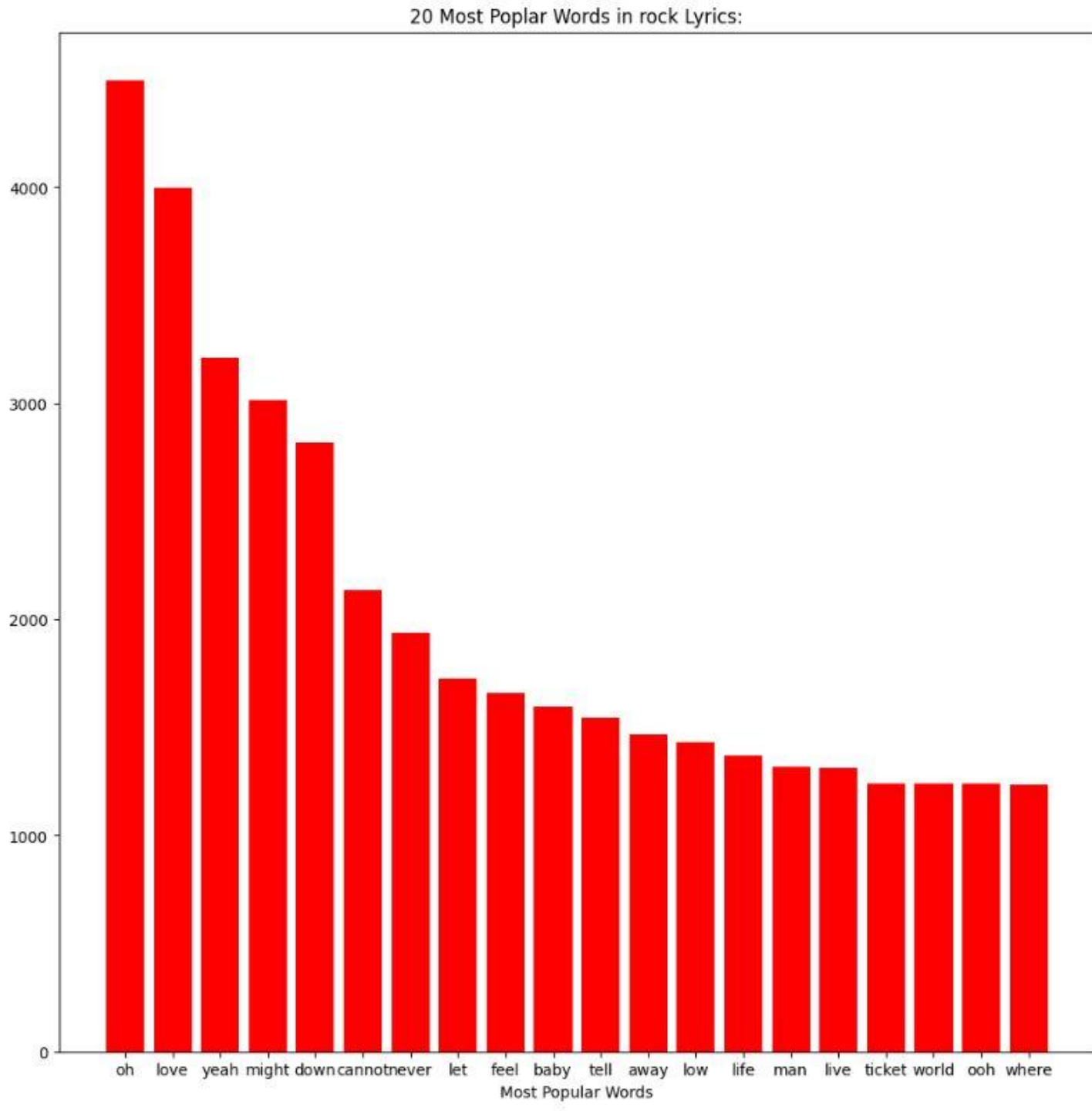
Top 20 country

- Top 20 most popular words in the country genere.
- Similar to the words we see in pop.



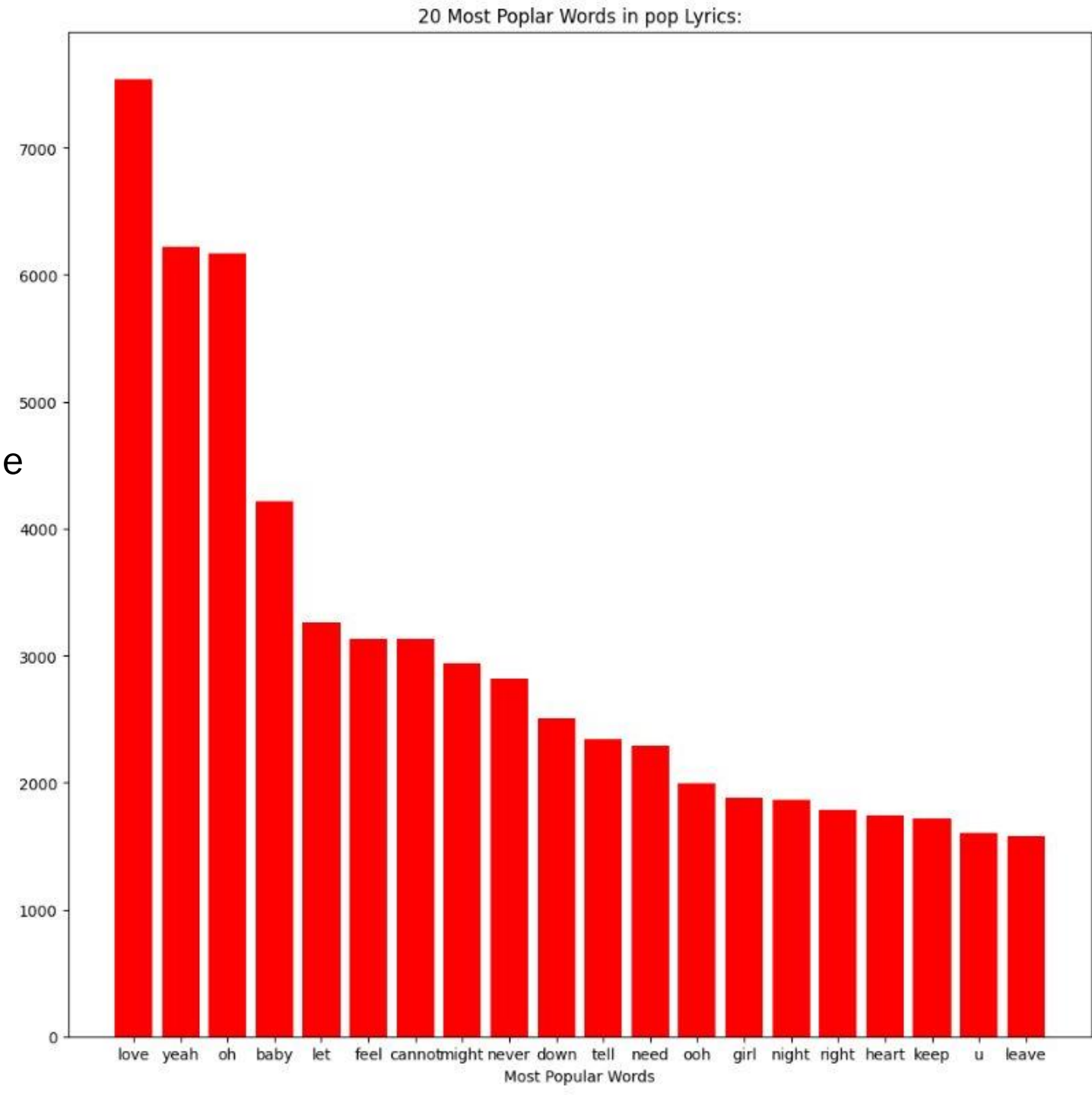
Top 20 rock

- Top 20 most popular words in the rock genere.
- Similar to the ones of rap and country to some extent.

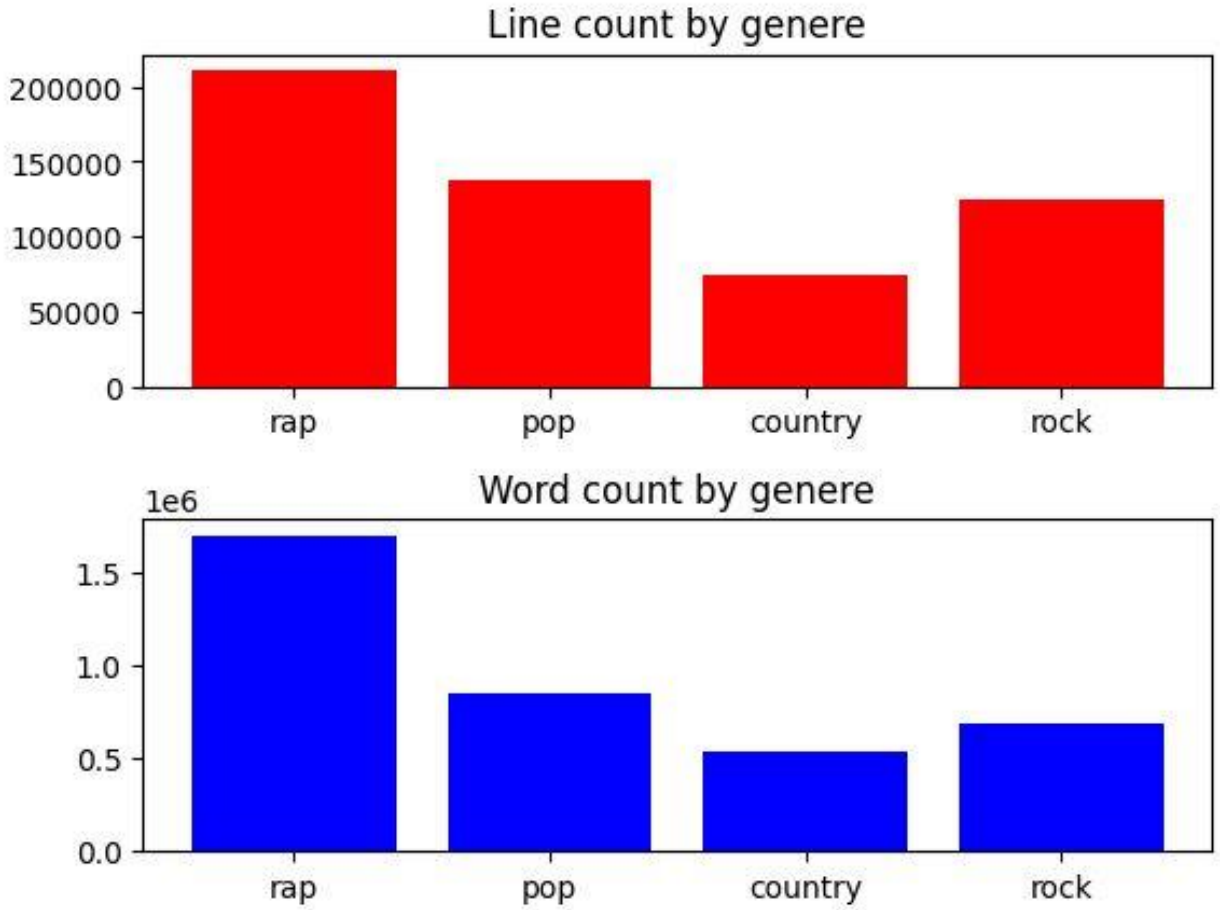


Top 20 pop

- Top 20 most popular words in pop.
- Similar to the rock and country to some extent.

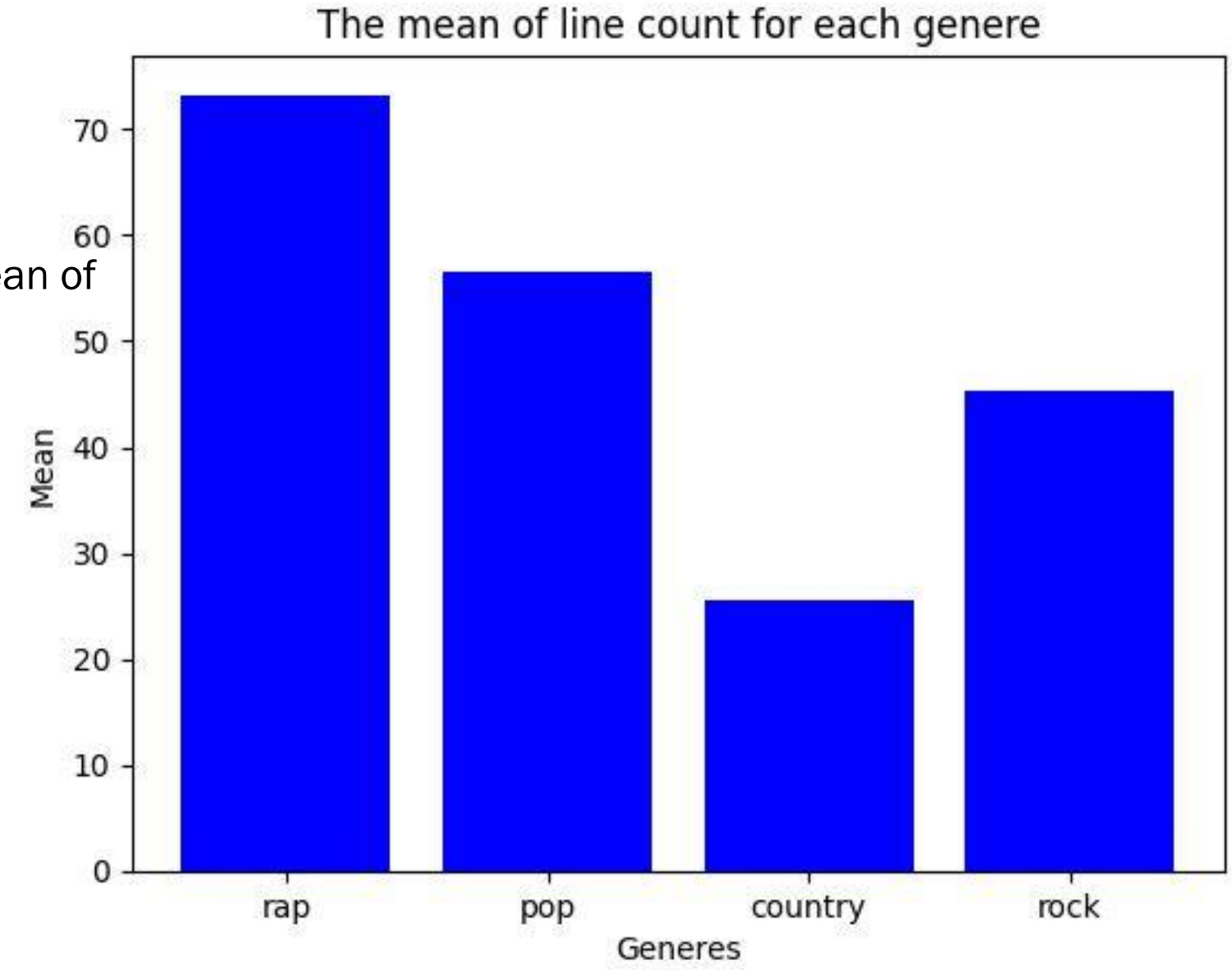


Line and word count



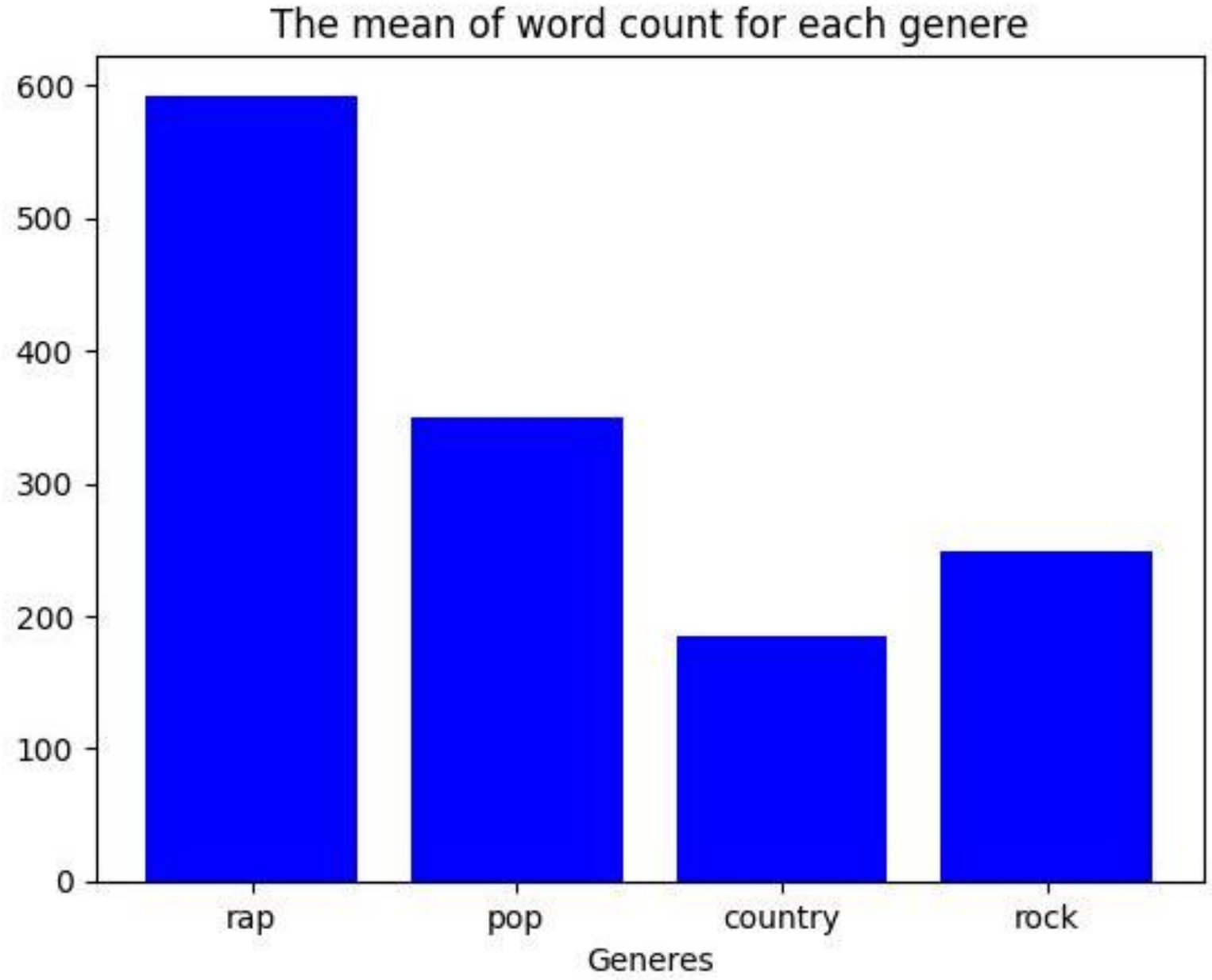
Line mean

This measurement represents the mean of the number of lines for each genere.



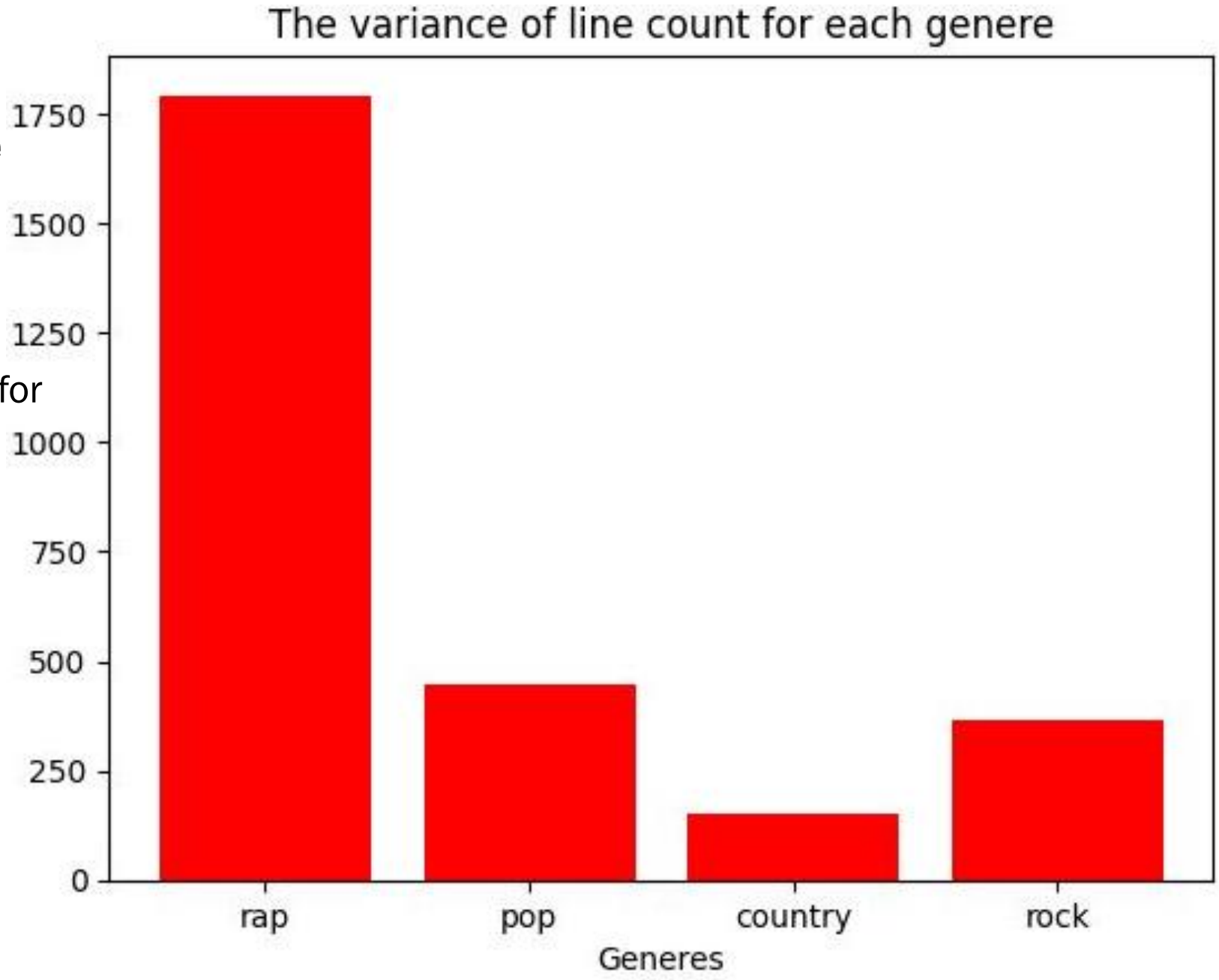
Word mean

This metric represents the mean of the number of words used in each genere.



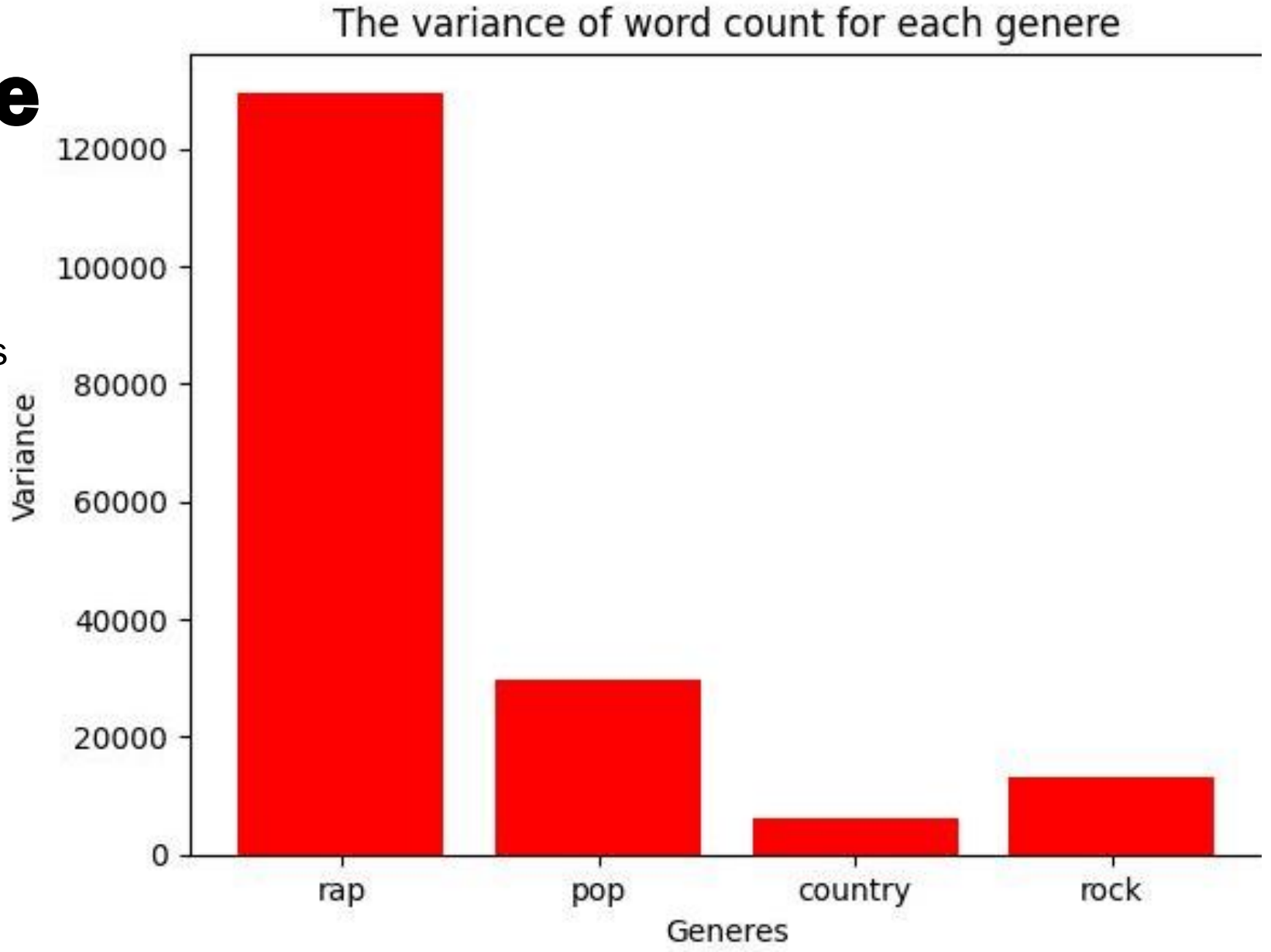
Line variance

This metric shows the variance between of the number of lines for each genere.



Word variance

This metric shows the variance between of the number of words used for each genere.



Conclusions

- Rap and country compared to each other have the greatest differences so far.
- Rock and rap have similarities in all of the metrics.
- We might expect the model to perform better in classifying rap and country.

The model

Star of the show

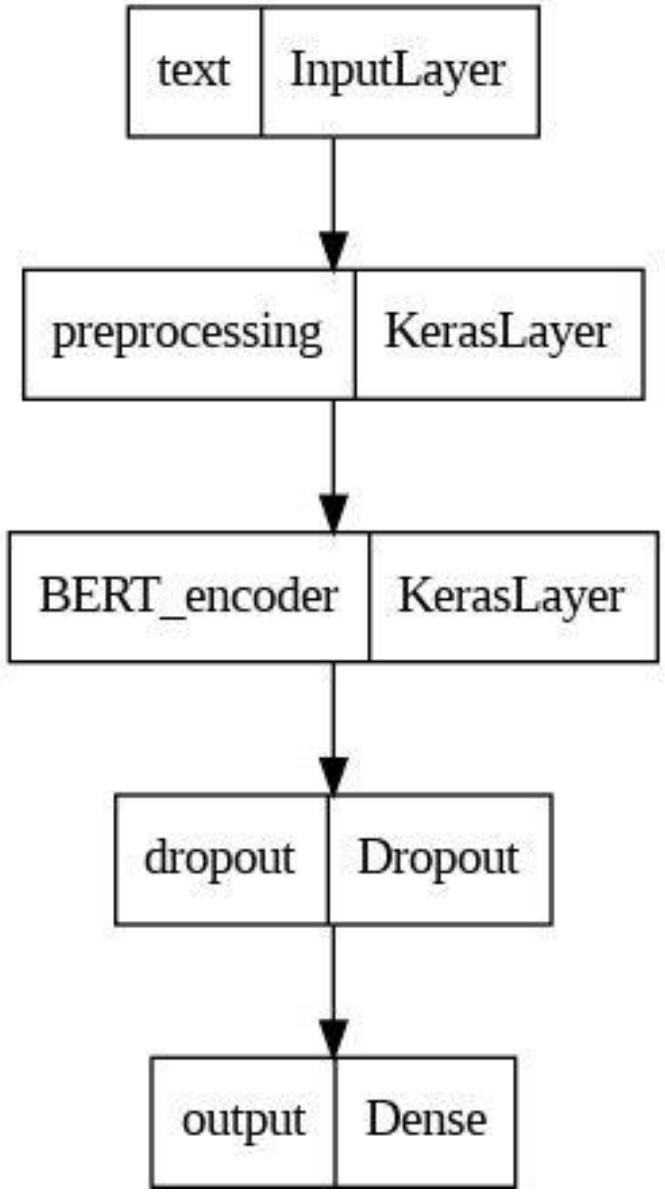
BERT or Bidirectional Encoder Representations from Transformers.

It’s role?

- Preprocesses the text from raw lyrics to numerical tokens.
- Does the embedding for the whole song.
- Produces a vector for each song which will be used for the prediction

Architecture

- We start with the unprocessed text.
- Text goes through the preprocessing layer that does the tokenization.
- The tokenized text passes through the encoding layer to produce the pooled_output.
- The pooled output passes through a regularization layer
- At last, it passes through the dense layer that does the classification.



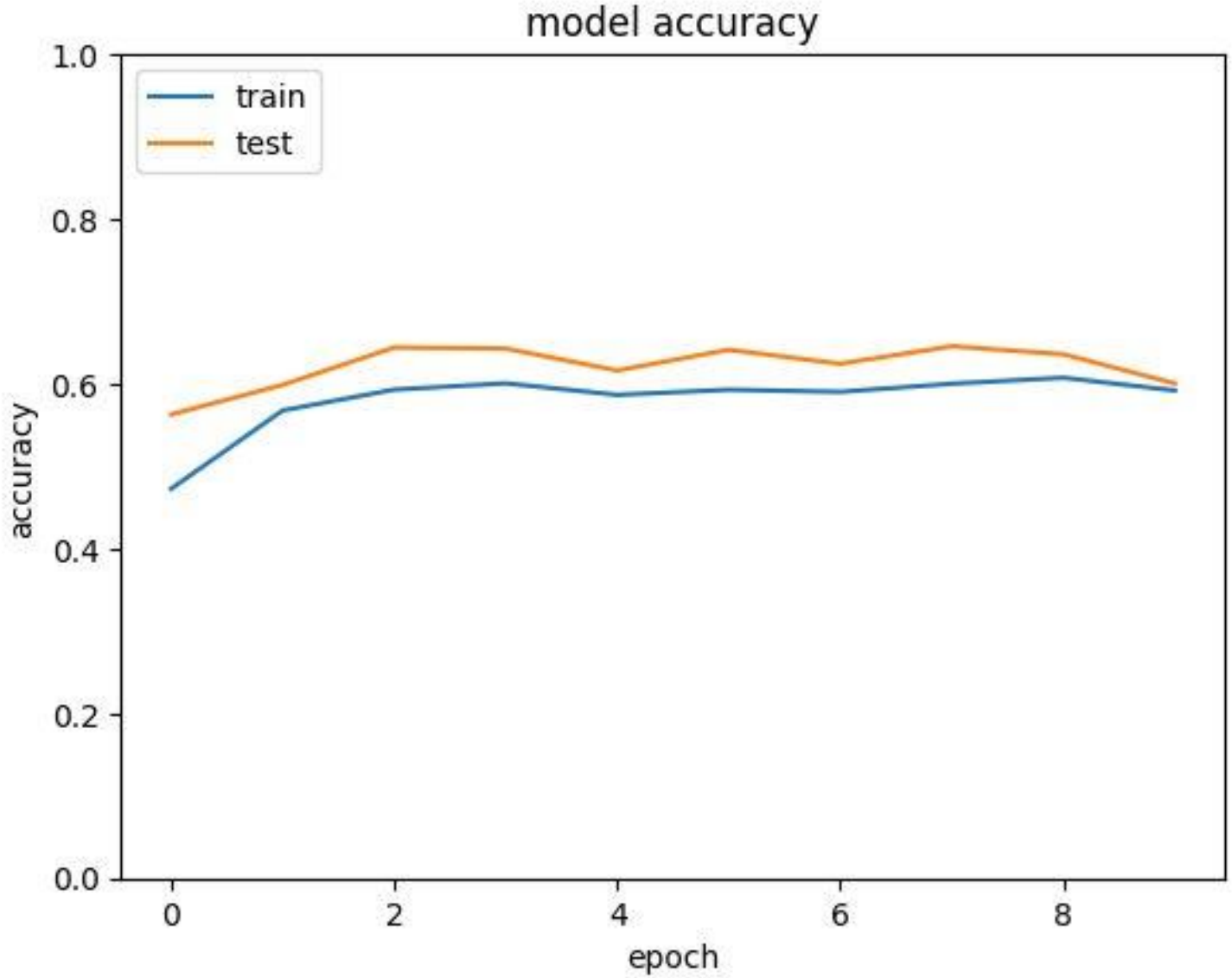
Evaluation

Performance metrics

	precision	recall	f1-score	support
0	0.90	0.70	0.79	884
1	0.53	0.45	0.49	736
2	0.83	0.41	0.55	857
3	0.43	0.83	0.57	822
accuracy			0.60	3299
macro avg	0.67	0.60	0.60	3299
weighted avg	0.68	0.60	0.60	3299

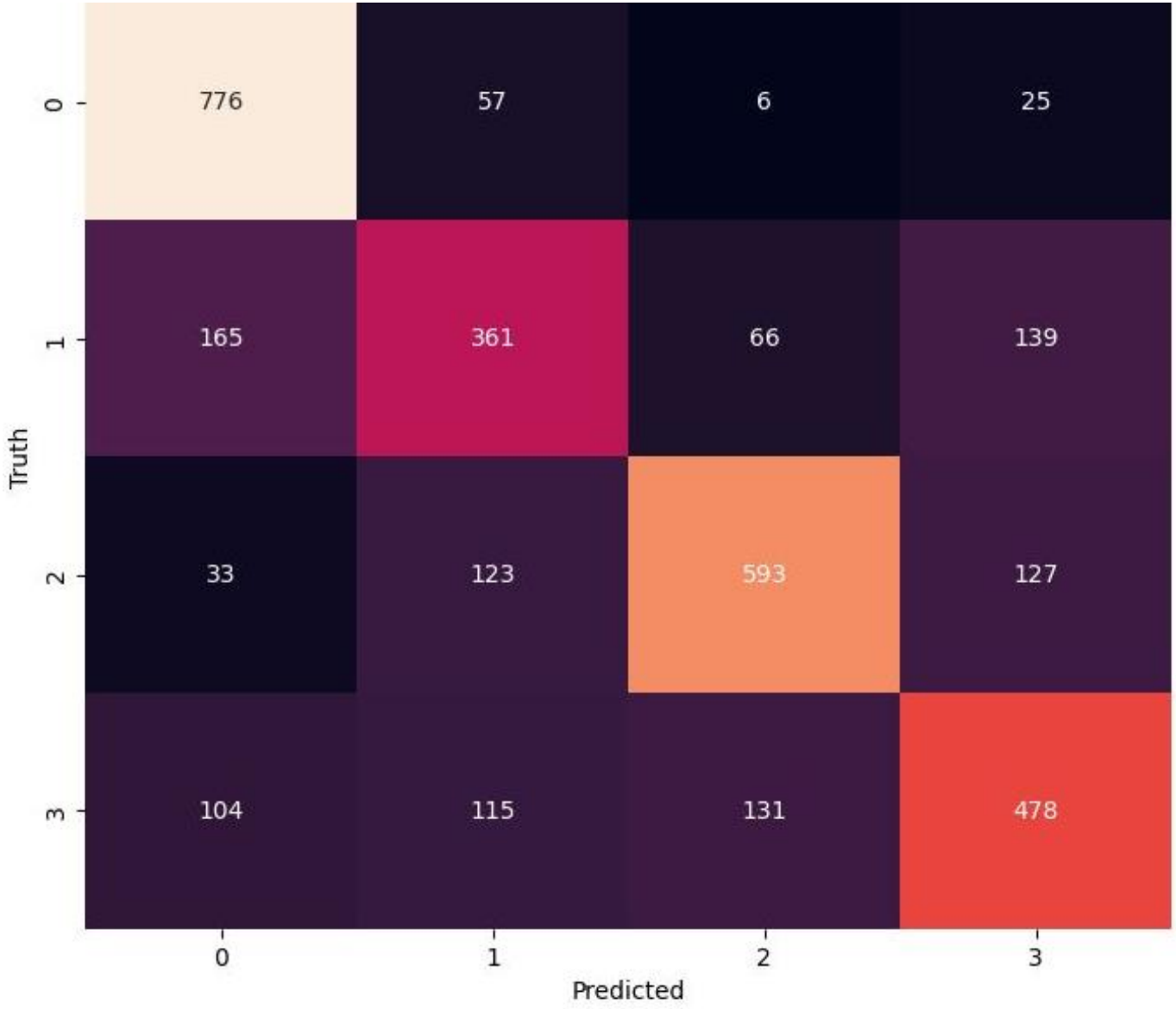
Learning curve

- Ran for only 10 epochs.
- Starts to converge after 10 epochs.
- Not overfitting.
- Decent result.



Confusion matrix

- Rap performing way better than the other genres.
- Struggling with pop, rock and country.
- Clear separation between rap-country and rap-rock.
- Misclassification between rap and pop.



Conclusion

What went wrong?

- Too much similarity between 3 different genres.
- Too many classes for a model this simple.
- Not enough information introduced.
- Different songs might belong to different genres

How to improve?

- Use of songs that belongs only to one genere.
- More tunning and training.
- Better architecture.

Overall

- Some misclassifications between different genres.
- Clear distinction between particular genres.
- Decent accuracy of 61%.
- Could be better but its still a decent model.

Thank you