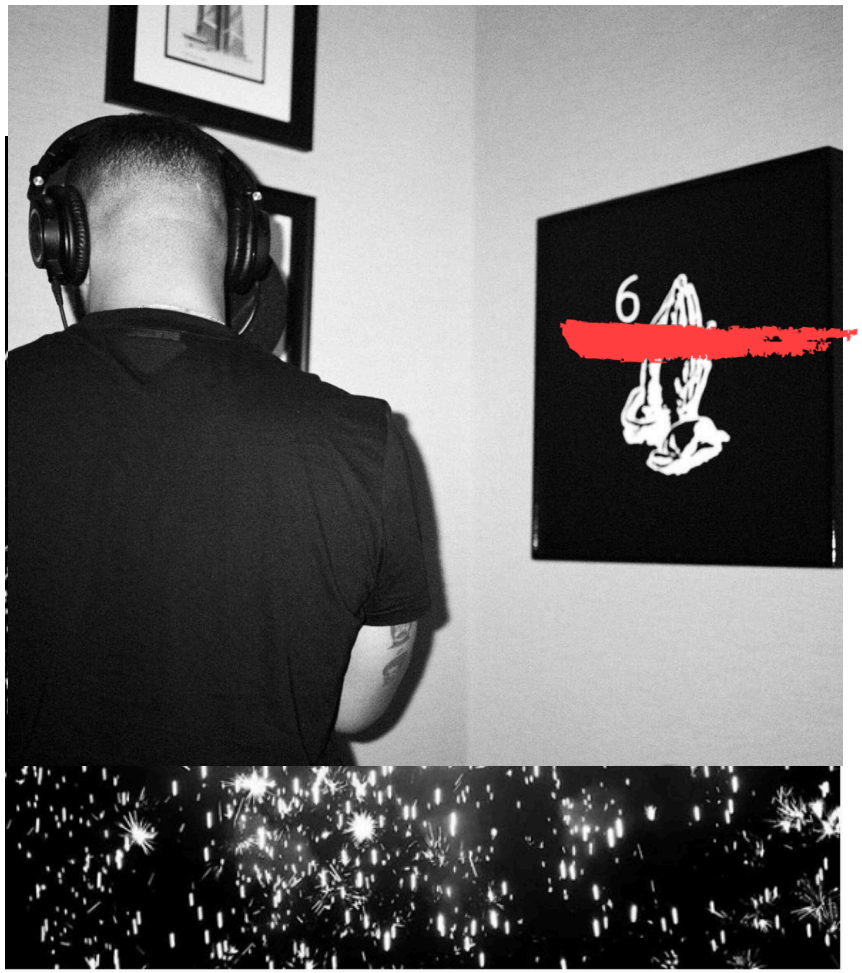
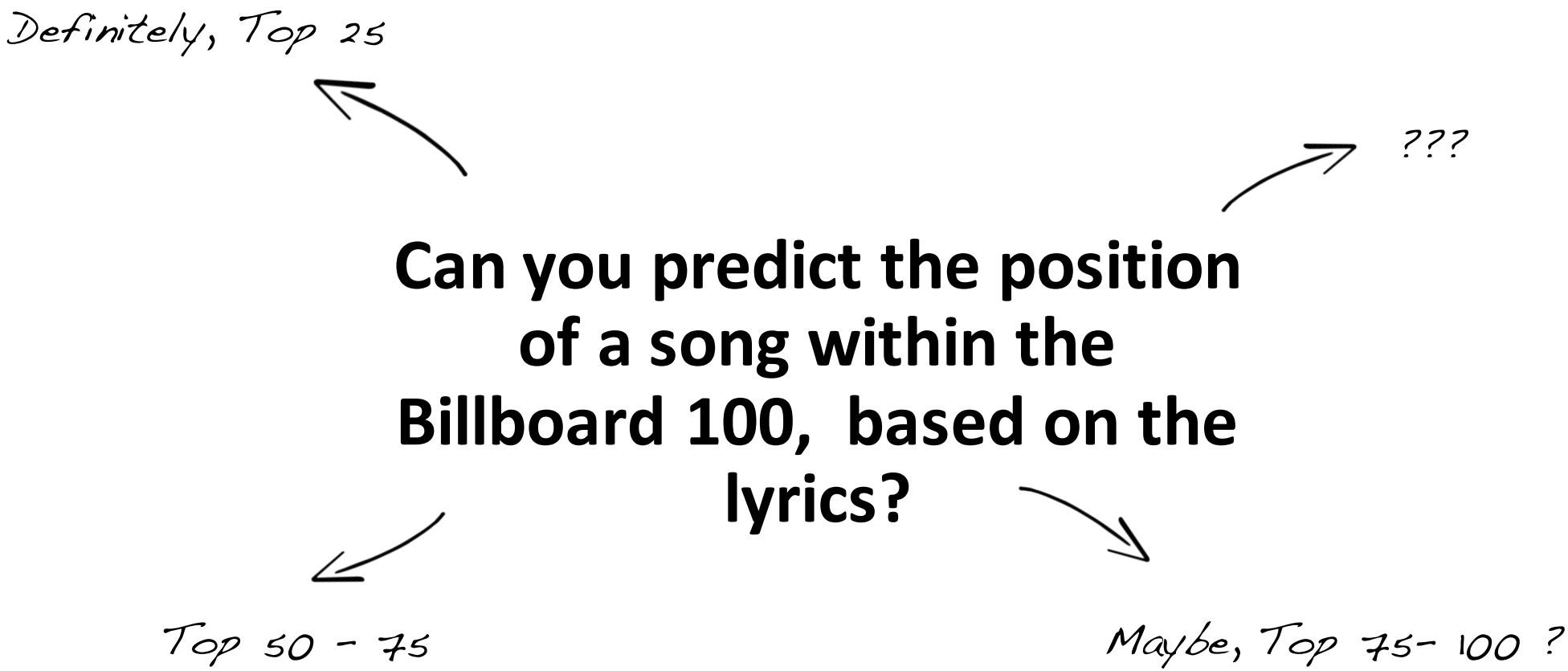


NLP on Billboard 100

Justine Williams


Definitely, Top 25
???
Can you predict the position of a song within the Billboard 100, based on the lyrics?
Top 50 - 75
Maybe, Top 75- 100 ?

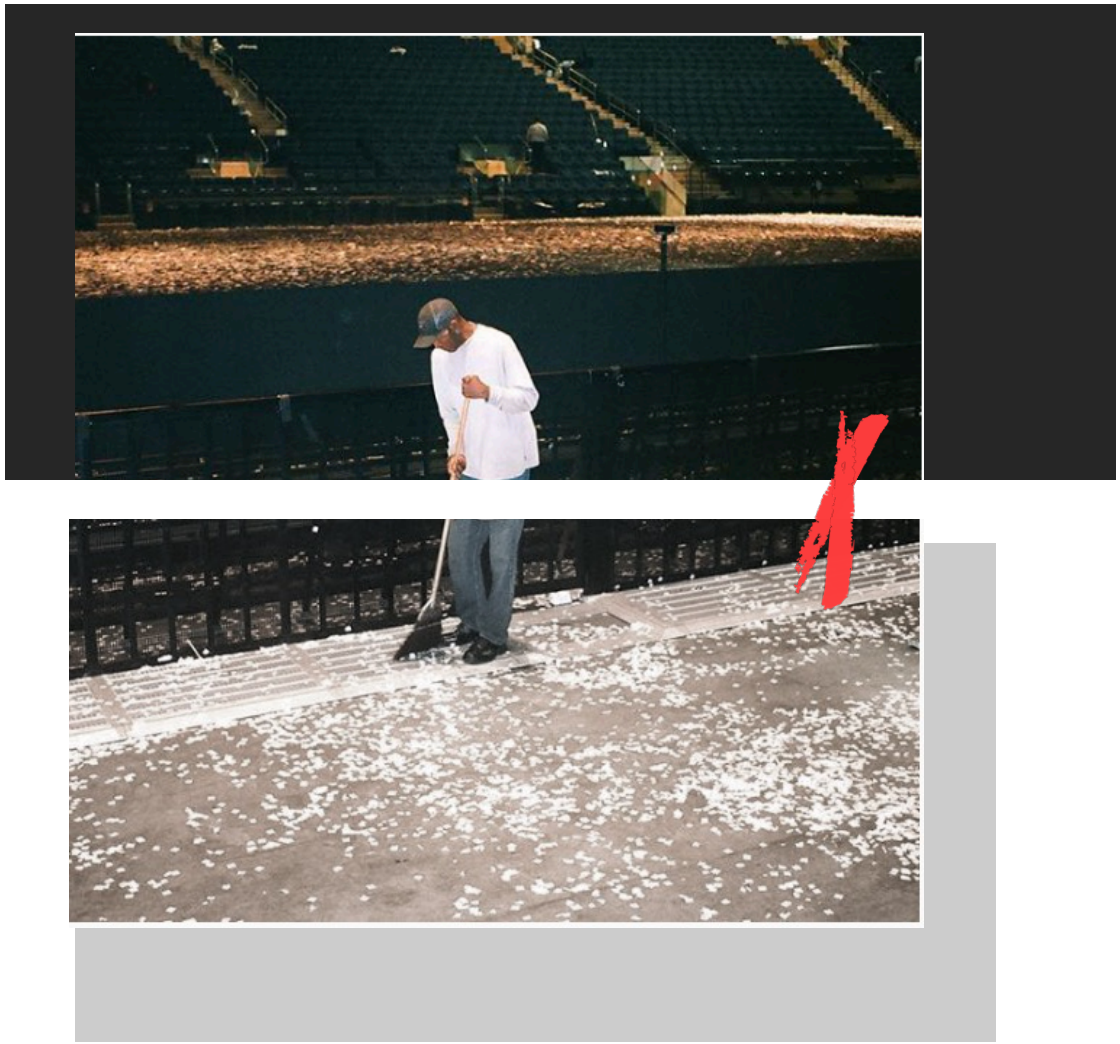
Agenda

Data Acquisition

EDA

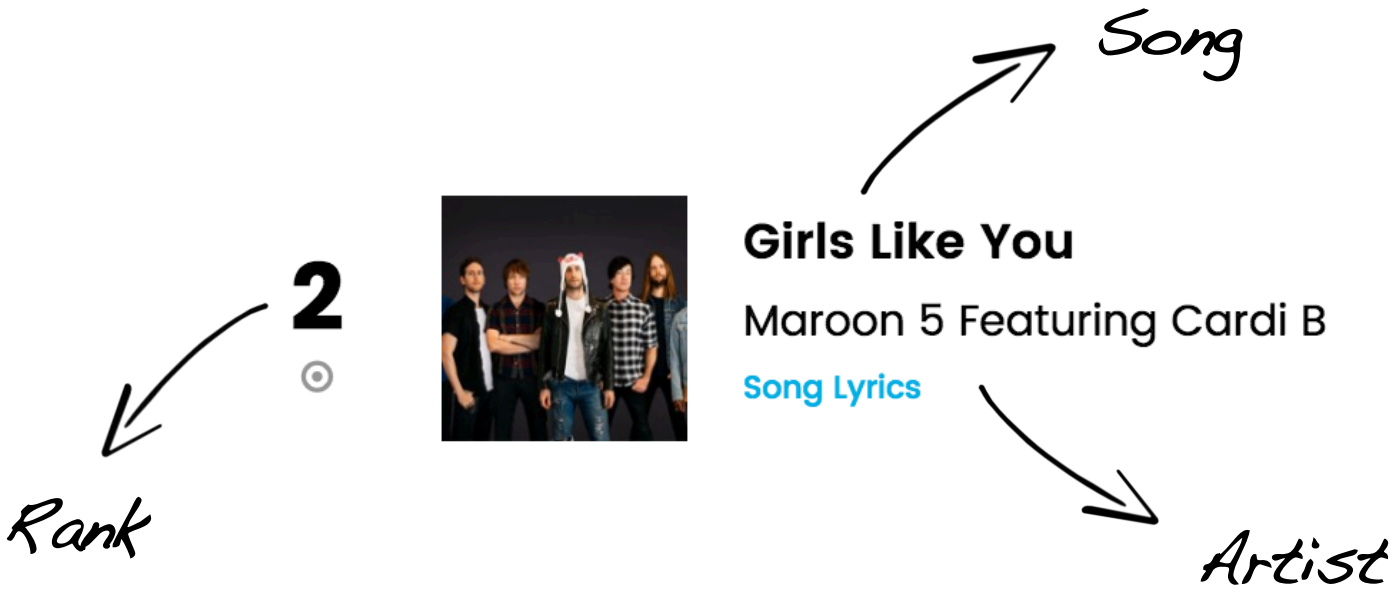
Modelling

Next Steps

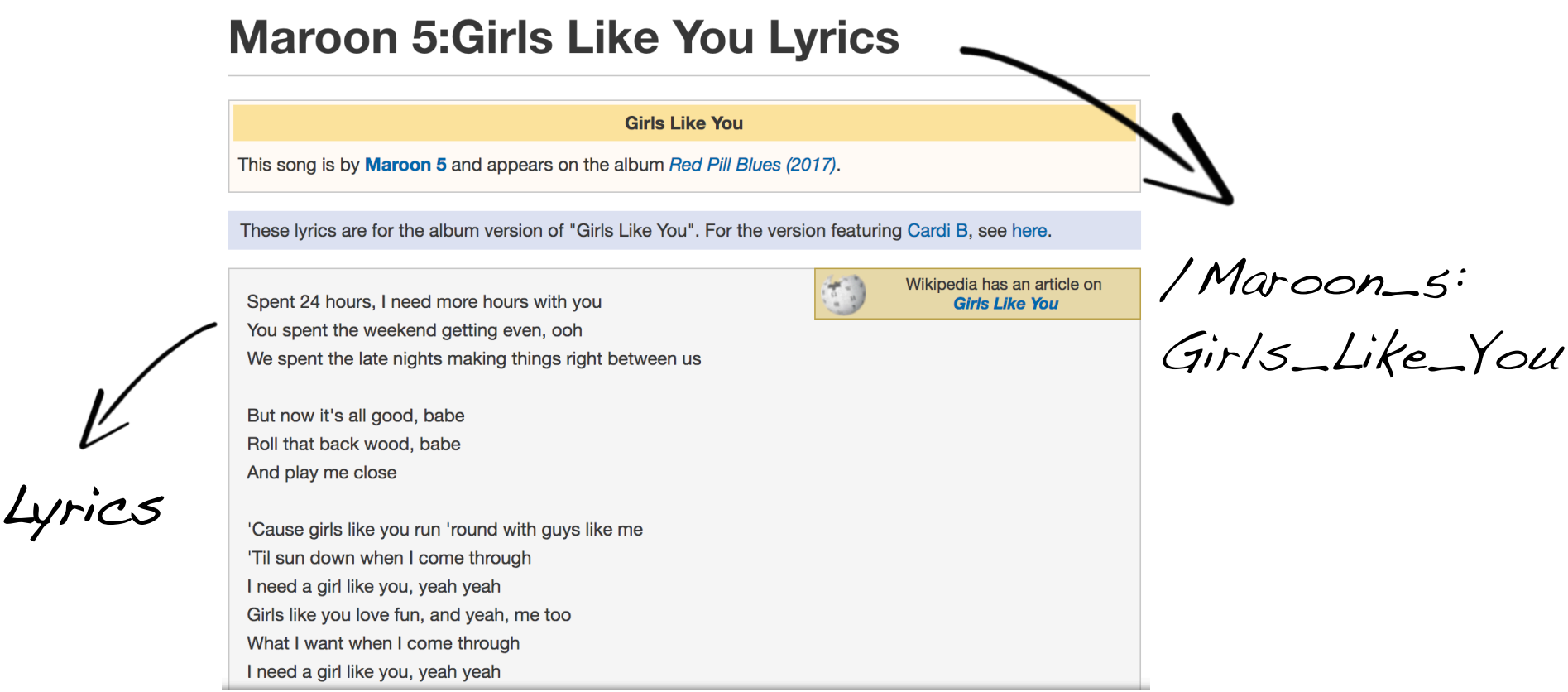


Scraping the Data

Billboard 100



Lyric Wiki



Maroon 5:Girls Like You Lyrics

Girls Like You

This song is by **Maroon 5** and appears on the album *Red Pill Blues (2017)*.

These lyrics are for the album version of "Girls Like You". For the version featuring Cardi B, see here.

Wikipedia has an article on ***Girls Like You***

Spent 24 hours, I need more hours with you
You spent the weekend getting even, ooh
We spent the late nights making things right between us

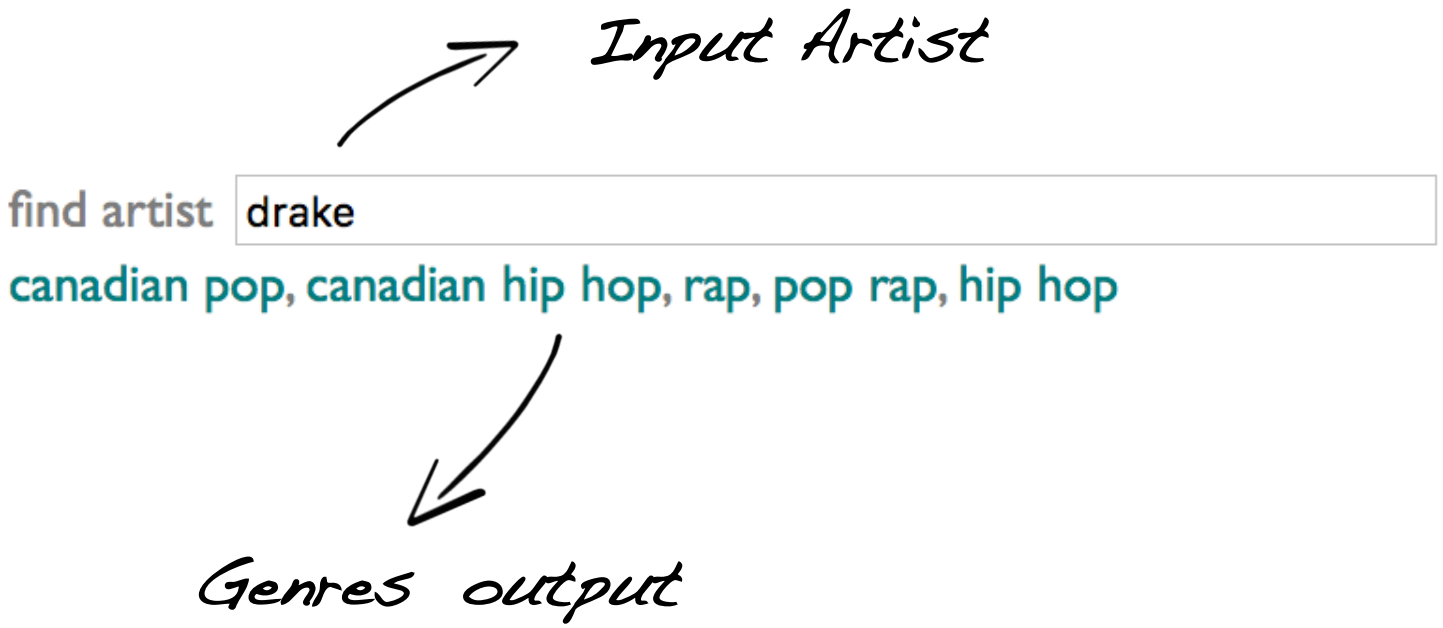
But now it's all good, babe
Roll that back wood, babe
And play me close

'Cause girls like you run 'round with guys like me
'Til sun down when I come through
I need a girl like you, yeah yeah
Girls like you love fun, and yeah, me too
What I want when I come through
I need a girl like you, yeah yeah

/Maroon_5:Girls_Like_You

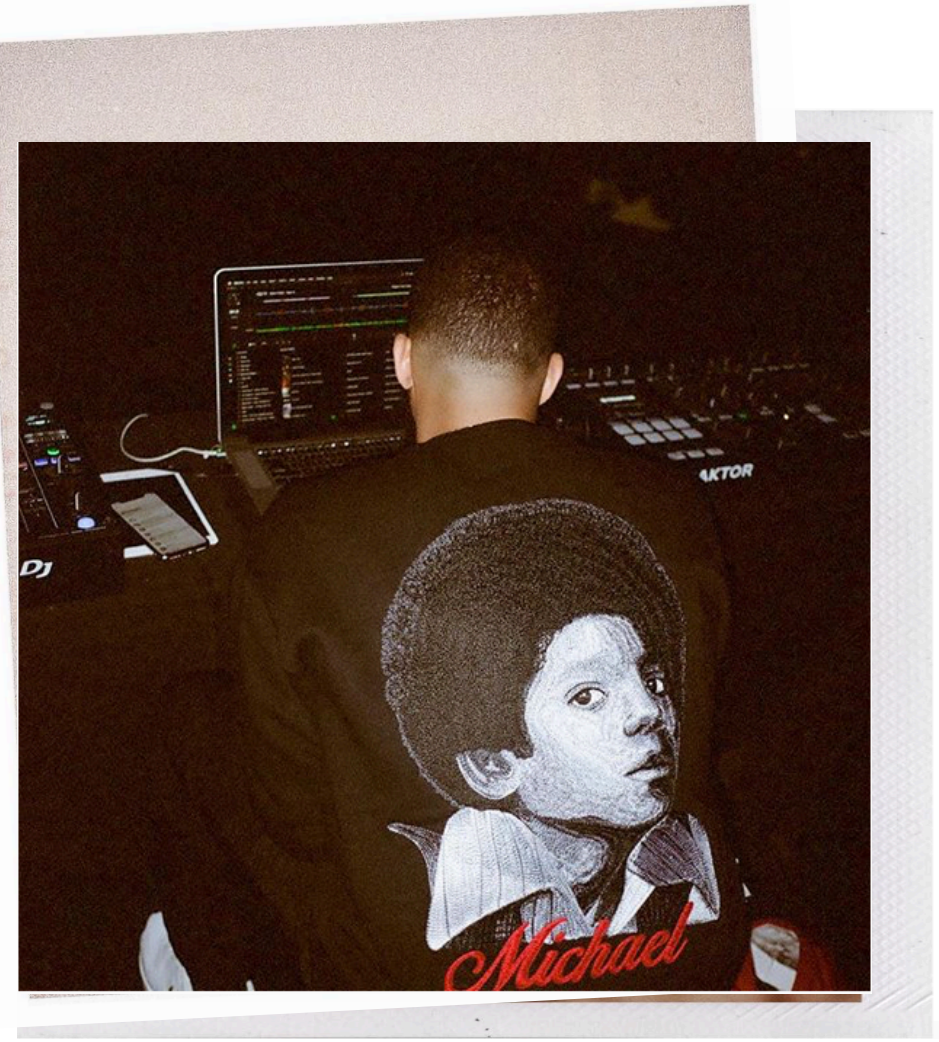
Lyrics

Every Noise


Input Artist
find artist
drake
canadian pop, canadian hip hop, rap, pop rap, hip hop
Genres output

EDA

What did I end up with?



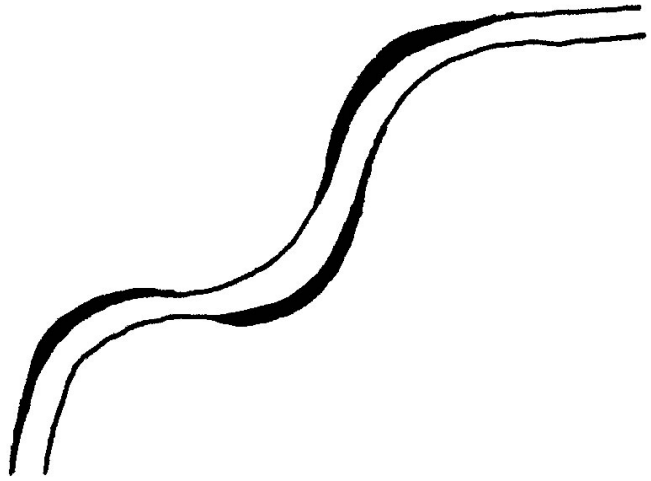
Cleaning the Data

Lyrics

- Remove non-English songs

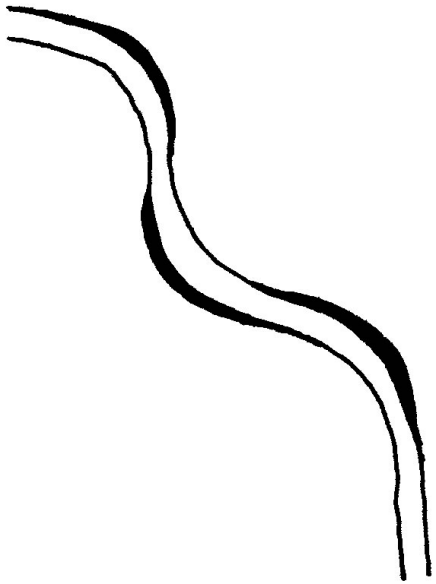
Genres

- Create broader groups



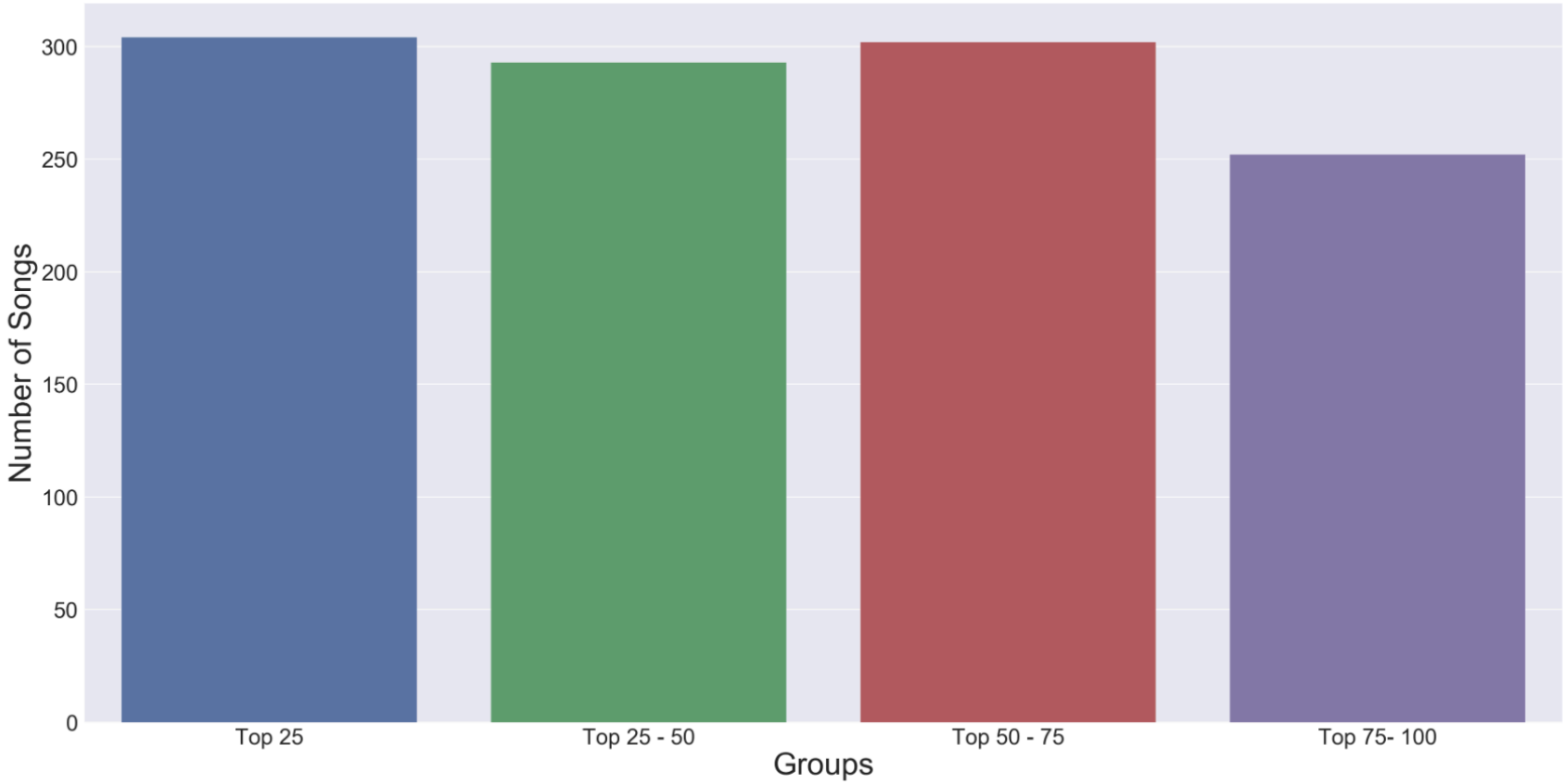
Data Debrief

- **Genres**: 22 types
- **Number of Songs**: 1,151
- **Time period:** August 2018, January 2016
- **Groups**: Top 25, Top 25 -50 , Top 50 – 75, Top 75 – 100

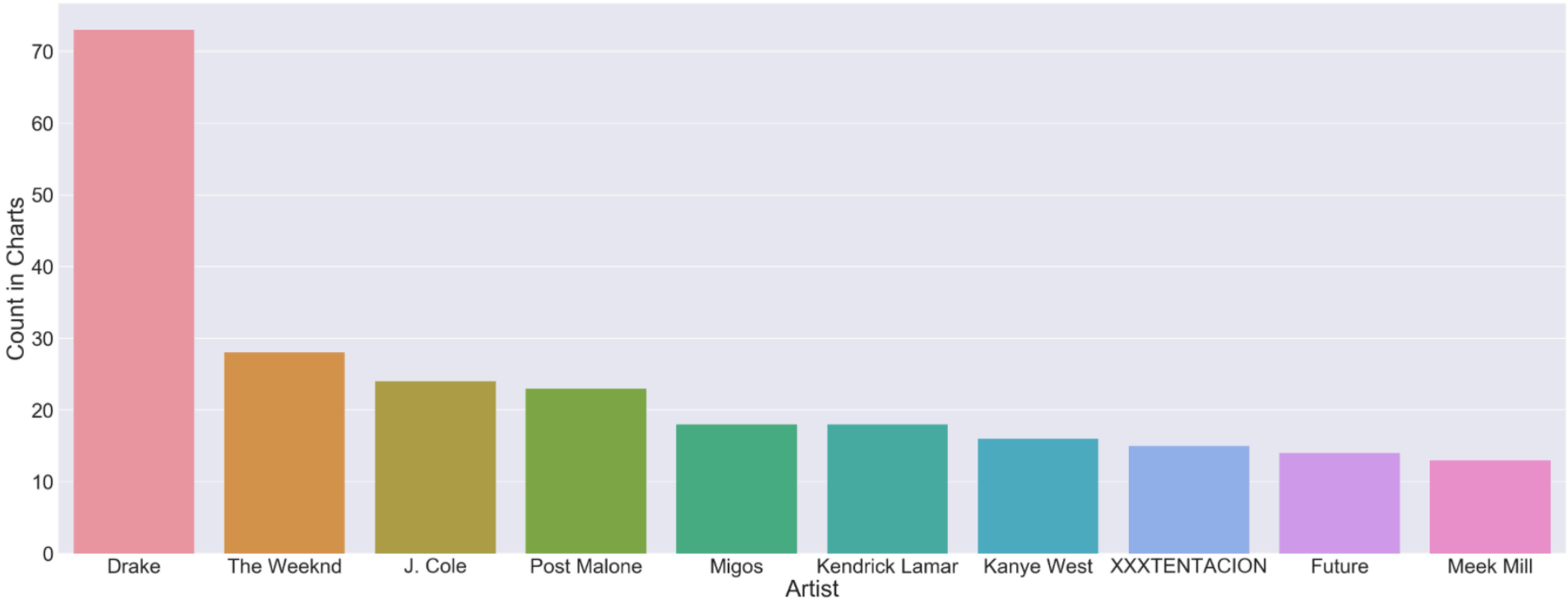


Exploring the Data

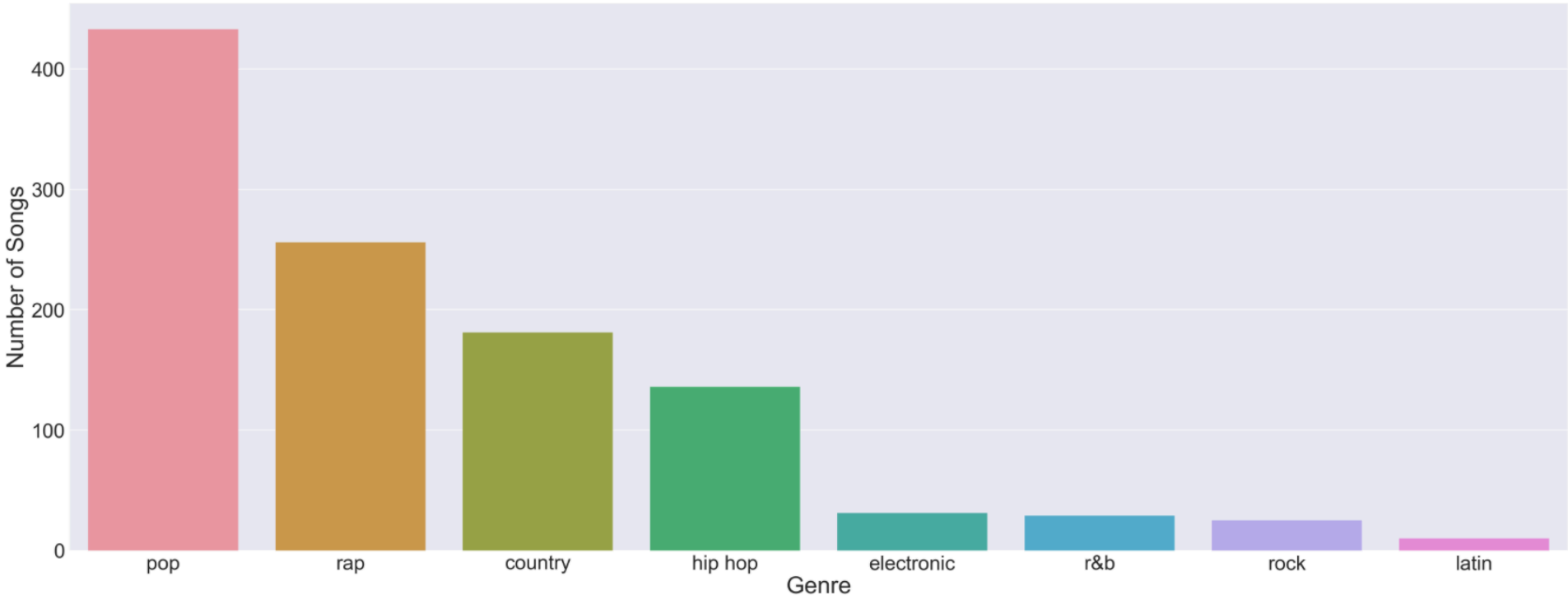
Distribution of Data by Groups



Top 10 Most Popular Artists

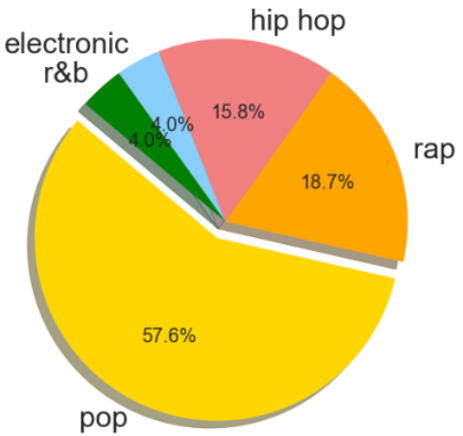


Top 8 Most Popular Genre

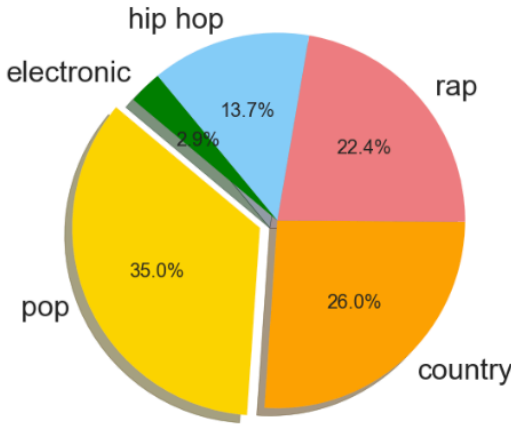


Genre Group Breakdown

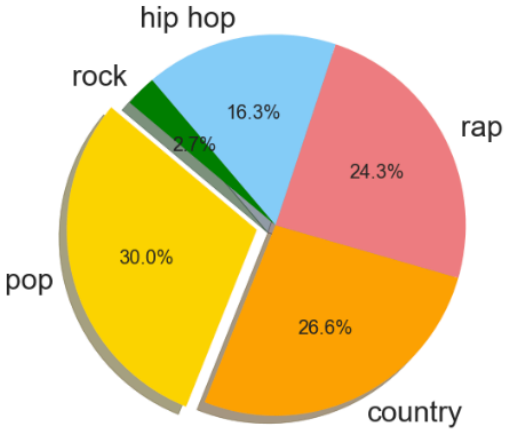
Top 25



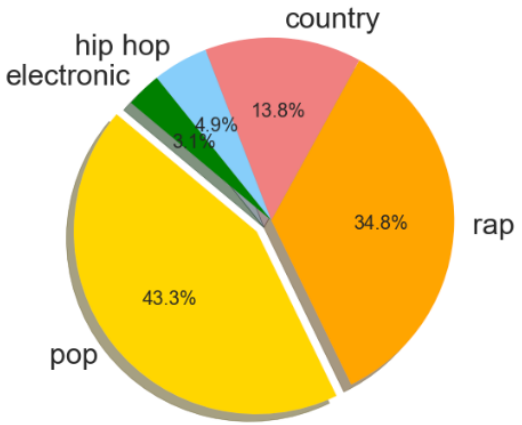
Top 25 - 50



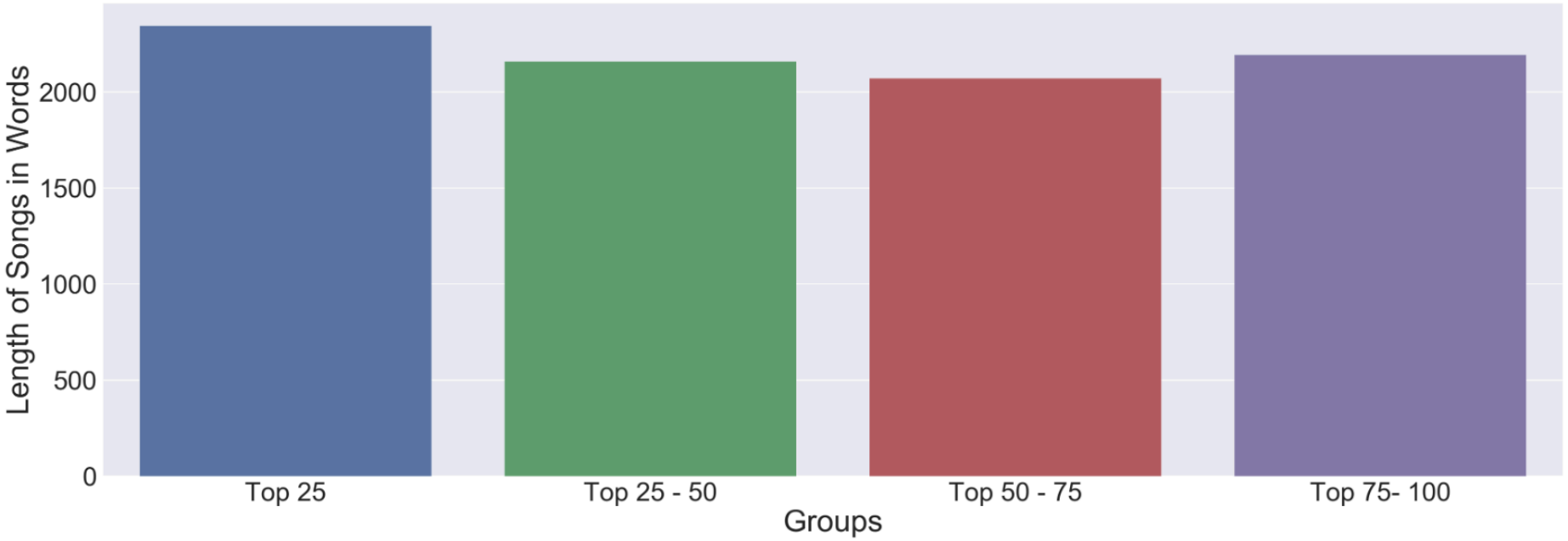
Top 50 - 75



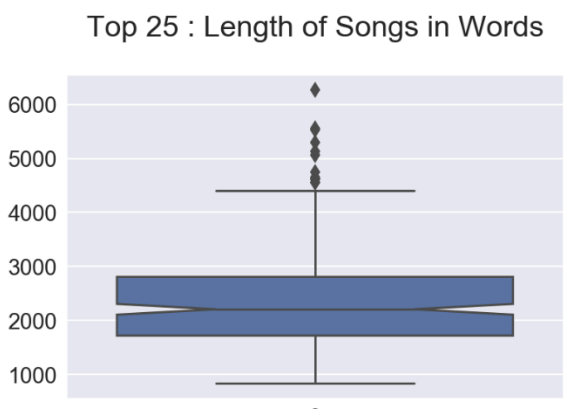
Top 75 - 100

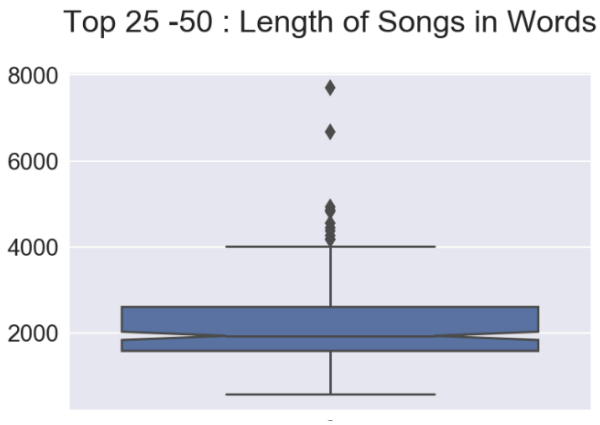


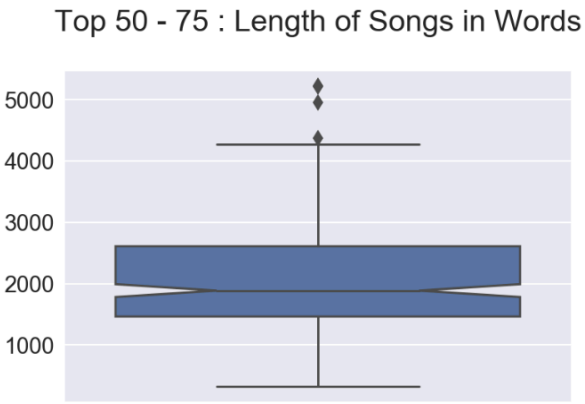
Length of Lyrics in Words

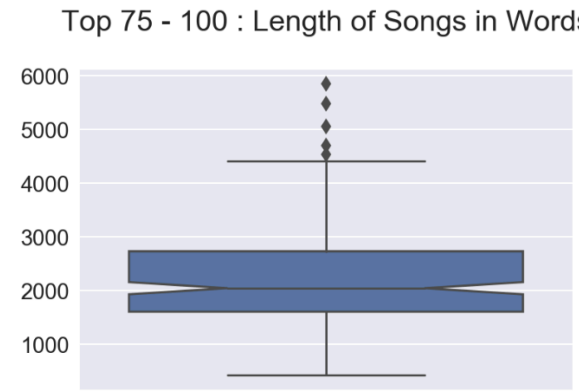


T-test - Length of Song in Words







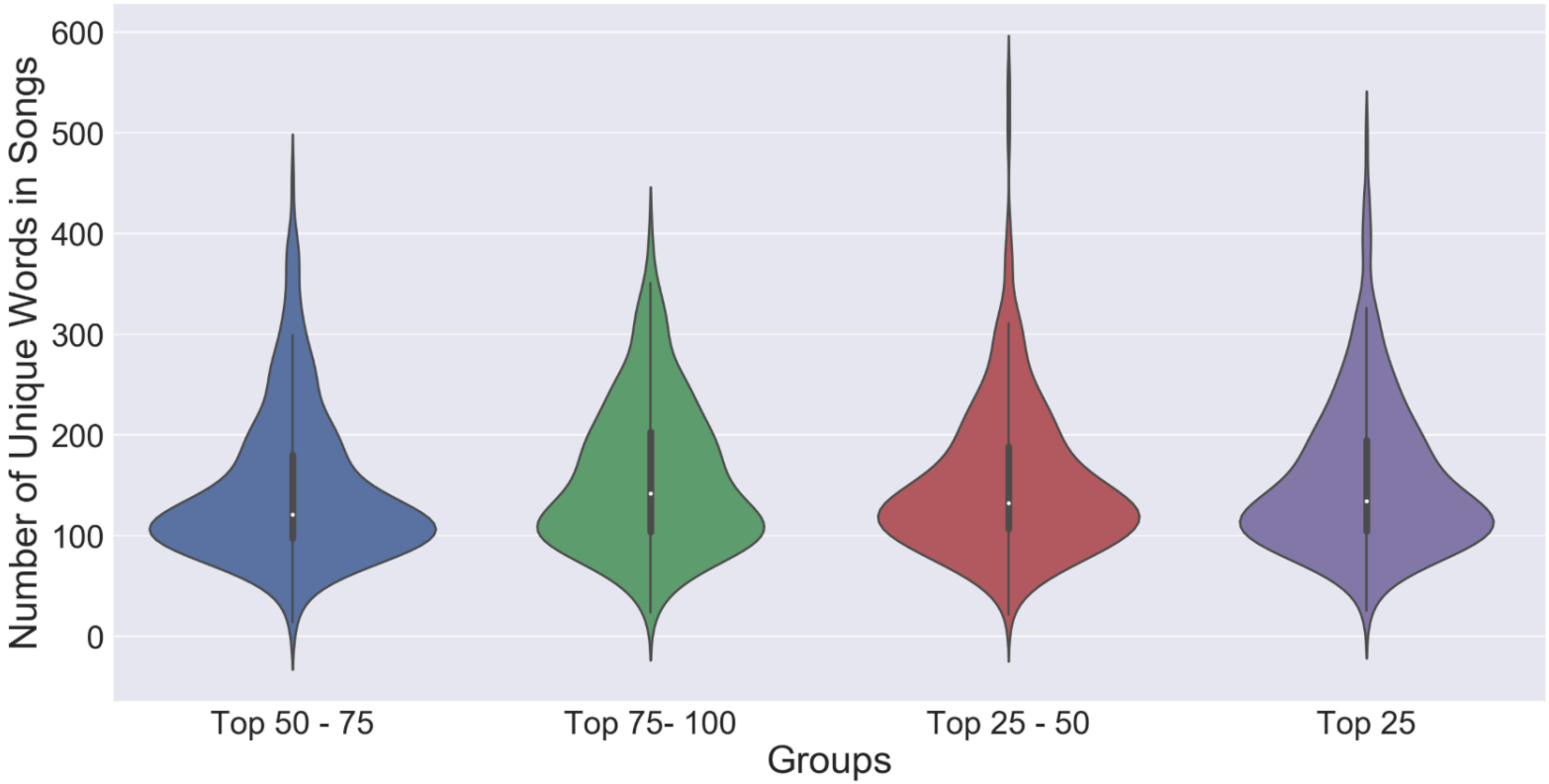


Null Hypothesis: There is no difference in the mean length of words in a song

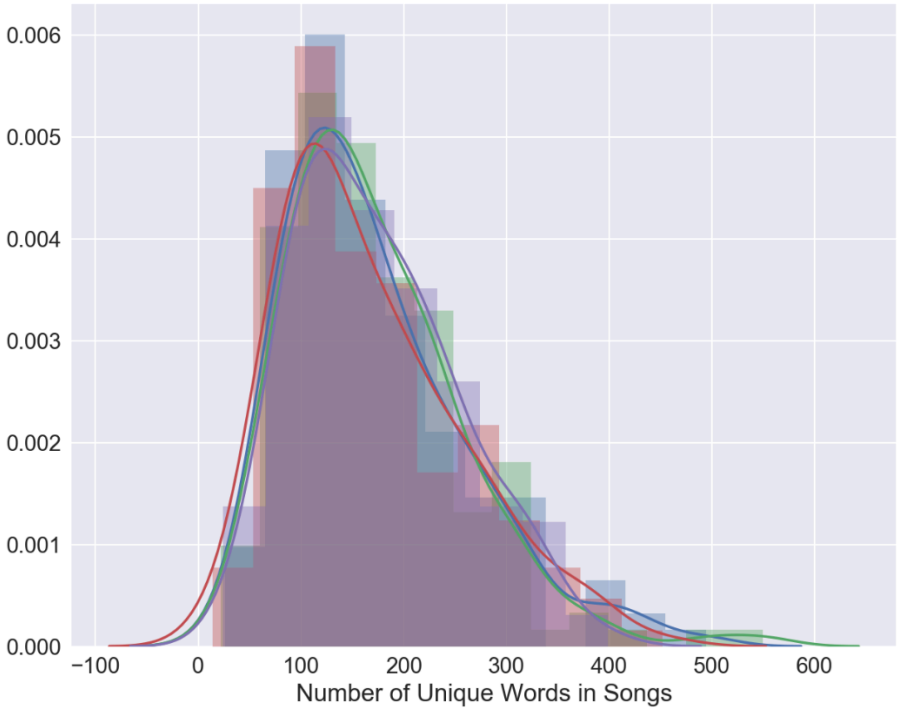
Alternative Hypothesis: There is difference in the mean length of words in a song

Conclusion:
Alternative Hypothesis accepted for songs in Top 25 versus rest

Number of Unique Words in Songs



T-test – Number of Unique Words in Songs



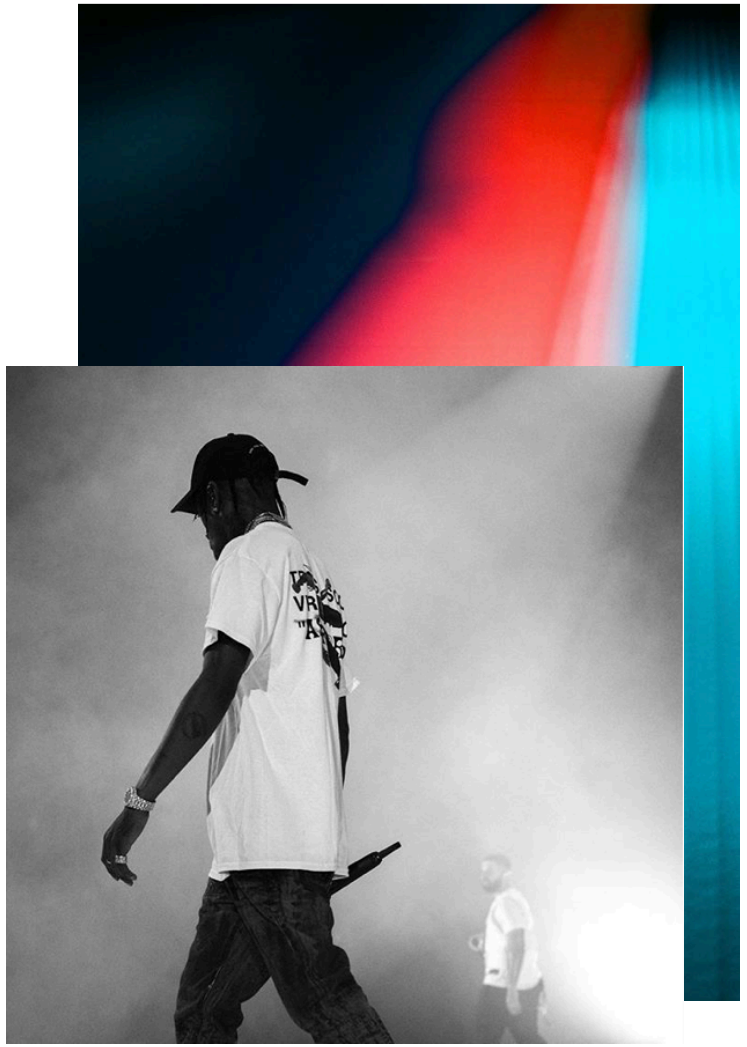
Null Hypothesis: There is no difference in the mean number of unique words in a song

Alternative Hypothesis: There is difference in the mean number of unique words in a song

Conclusion:
Null Hypothesis was accepted for all groups

Count Vectorizer

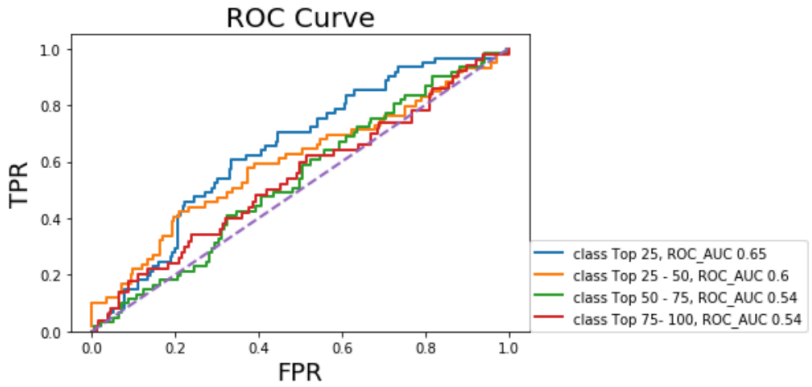
Based on vocabulary used, can you predict the group a song belongs to?



Support Vector Classifier

Count Vectorize Features

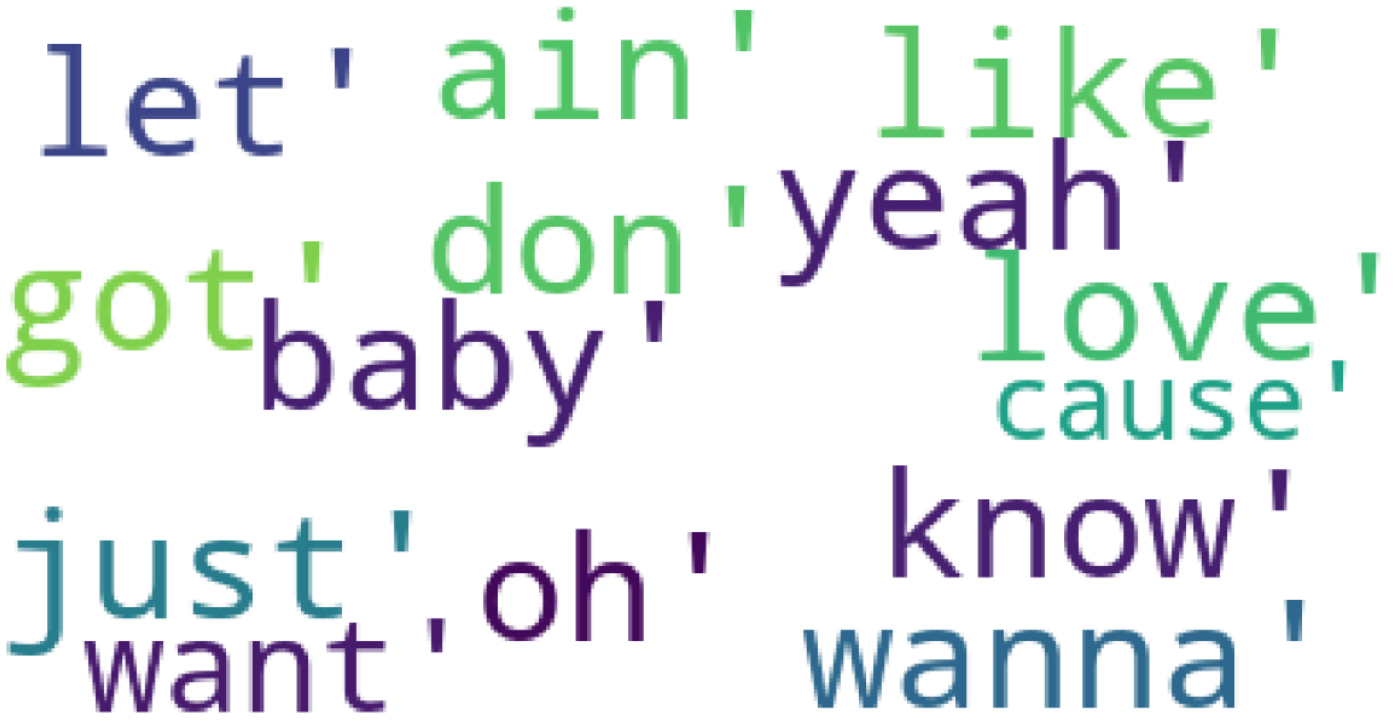
Max 3000 features
N grams (1,2)
Min df = 20



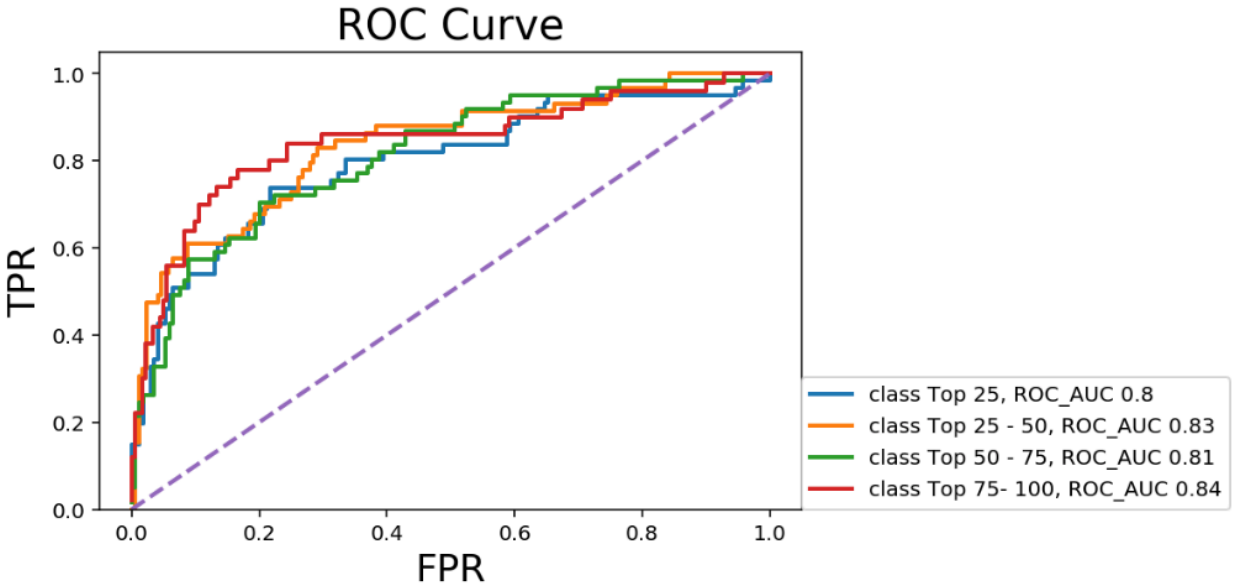
C = 6
Narrow margins, harsh on outliers

Score: 32%

Top Words Across all Groups



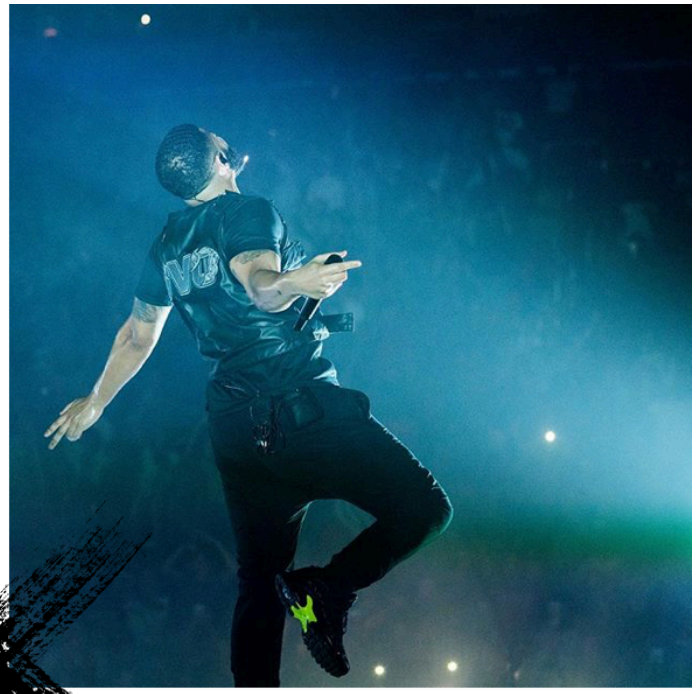
Count Vectorizer with Genres & Artist Popularity

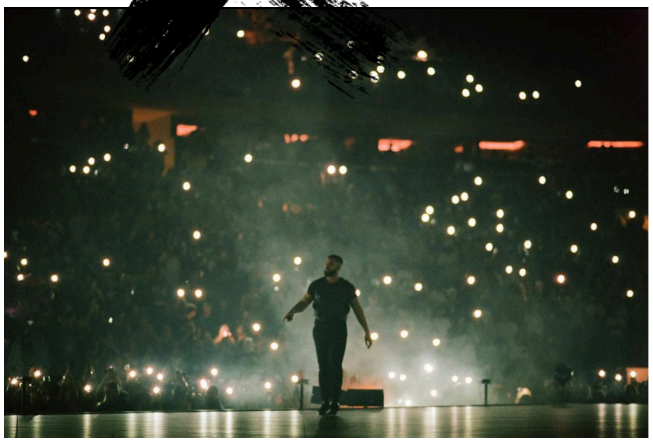


Score: 59%

- Best at classifying Top 75- 100
- Worst at classifying Top 25

Model: Logistic Regression





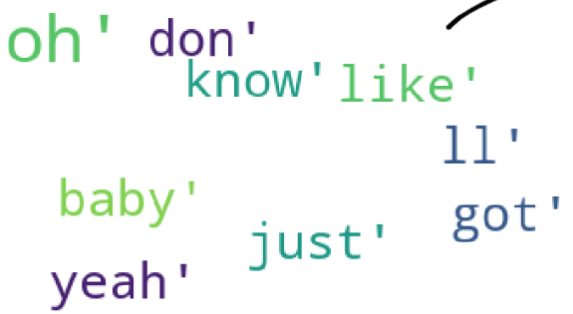
TF – IDF

Can you use uncommon words to better predict the group of a song?

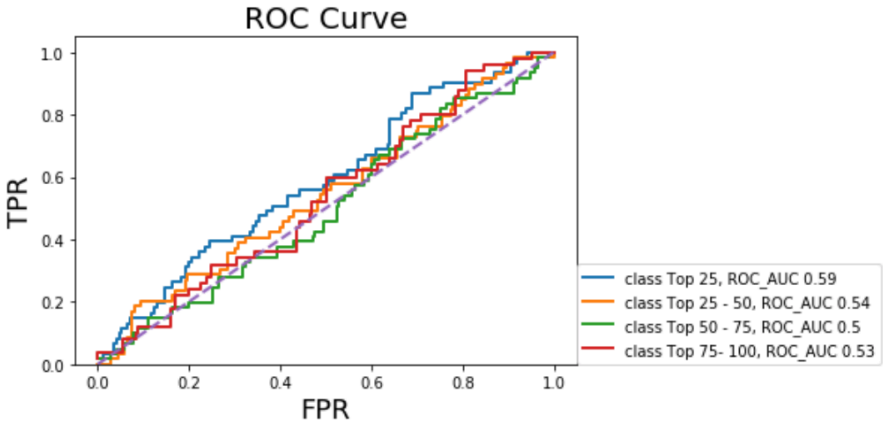
Logistic Regression

TF –IDF Features

Max 3000 features
N grams (1,2)

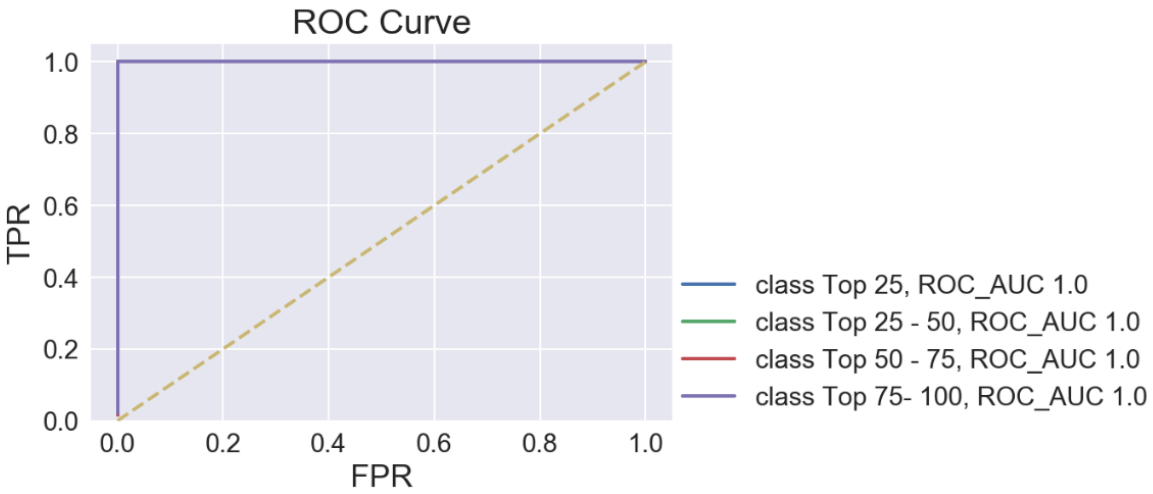


50% of Top 20 words the same across all groups



Score: 29%

TF – IDF, Genres & Artist Popularity



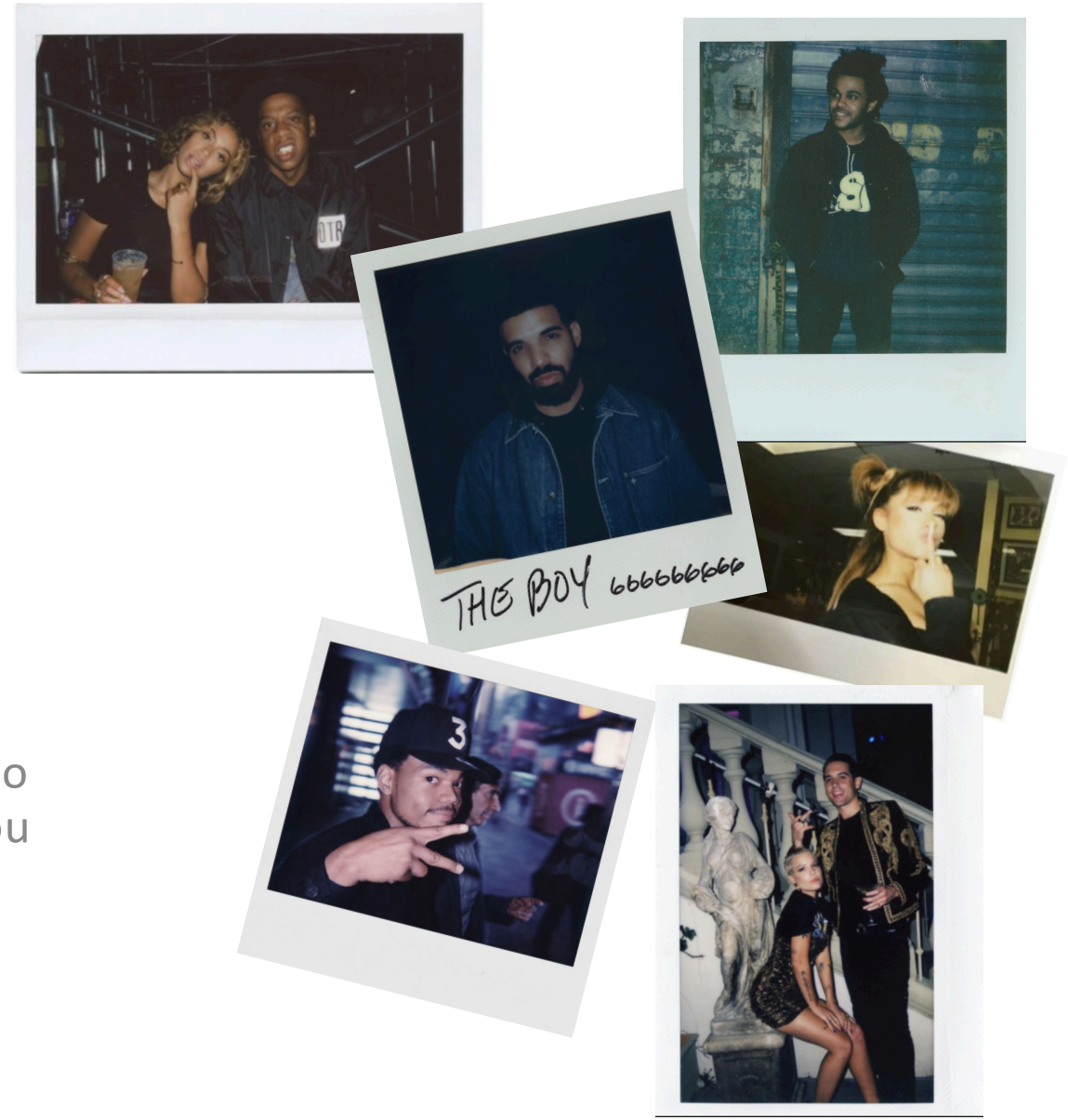
Score: 99.5%

Features: Max features =3000
N grams =(1,2)

Absolute Coefficients	Feature
15.535310	Artist_Top25_Count
7.753641	Artist_Top25_50_Count
6.859544	Artist_Top50_75_Count
4.977335	Artist_Top75_100_Count
0.732758	won let
0.606586	hallelujah hallelujah
0.601081	heaven
0.598898	holy holy

Sentiment Analysis

With commutes radio’s are more likely to play happy songs than sad songs, are you able to classify songs by sentiment?



Happiest, Misclassified & Saddest Songs

Maroon 5 - Girls Like You

Leonard Cohen - Hallelujah

Jon Bellion - All Time Low

KNN Classifier

Score: 31%

Number of Neighbors = 5

Predicted	Top 25	Top 25 - 50	Top 50 - 75	Top 75 -100
Top 25	30	12	12	7
Top 25 – 50	21	17	16	5
Top 50 - 75	25	15	15	6
Top 75 - 100	18	13	10	9

Highest number of Misclassified Points

Sentiments by Groups

Top 25

Sentiment	Value
Positive	0.107
Negative	0.142
Neutral	0.752
Compound	**0.201**

Top 25 - 50

Sentiment	Value
Positive	0.104
Negative	0.151
Neutral	0.744
Compound	**0.310**

Top 50 - 75

Sentiment	Value
Positive	0.098
Negative	0.147
Neutral	0.755
Compound	**0.357**

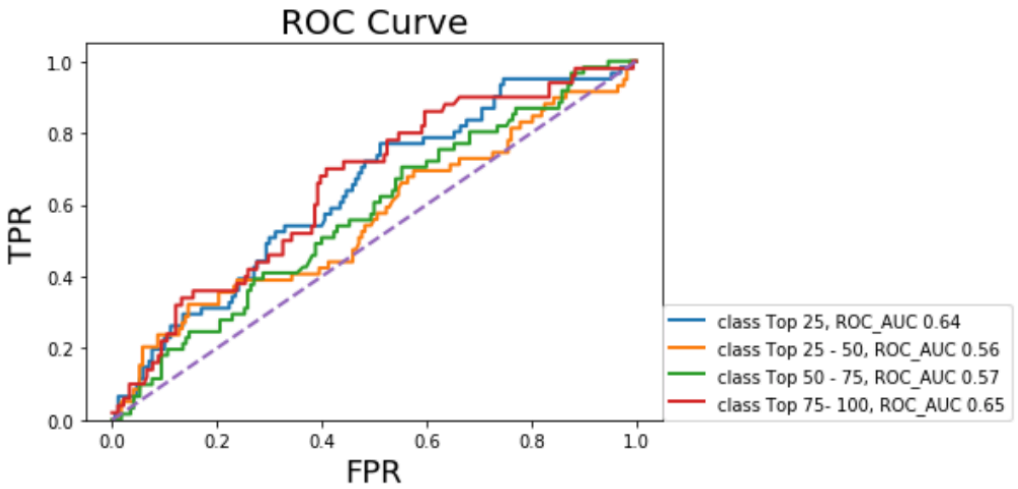
Top 75 - 100

Sentiment	Value
Positive	0.110
Negative	0.134
Neutral	0.756
Compound	**0.069**

Vader Compound & Genres

Score: 33 %

Logistic Regression (C = 5, solver= lbfgs)



Predicted	Top 25	Top 25 - 50	Top 50 - 75	Top 75 - 100
Top 25	43	9	3	6
Top 25 – 50	27	6	14	12
Top 50 - 75	32	11	12	6
Top 75 - 100	22	9	4	15

Repetition Analysis

Pop songs are more repetitive than rap songs, use the distribution of genres in groups to classify songs

Theory

Zopfli Compression

- Huffman Code
- Lempel Ziv (LZ77)

Repetition Calculation:

$$\frac{\text{length of the song} - \text{length of compression}}{\text{length of the song}}$$

Lempel Ziv with Most Repetitive Song

Feliz navidad
Feliz navidad
Feliz navidad
Prospero año y felicidad

Feliz navidad
Feliz navidad
Feliz navidad
Prospero año y felicidad

I wanna wish you a merry Christmas
I wanna wish you a merry Christmas
I wanna wish you a merry Christmas
From the bottom of my heart

We wanna wish you a merry Christmas
We wanna wish you a merry Christmas
We wanna wish you a merry Christmas

(Repeats 3 times)

Eliminating Repetition

Feliz navidad
Prospero ano y felicidad

I wanna wish you a merry Christmas
From the bottom of my heart

We

Huffman Coding – Most Repetitive Song

Feliz navidad
Prospero ano y felicidad

I wanna wish you a merry Christmas
From the bottom of my heart

We

Compressing

91.43%

Letter	Frequency	Code
A	9	11
E	7	00
Z	1	10000
M	2	10001

Repetition Breakdown

Groups

Groups	Repetition Percentage
Top 25	**71.4 %**
Top 25 – 50	69.6 %
Top 50 - 75	69.7 %
Top 75 - 100	69.1 %

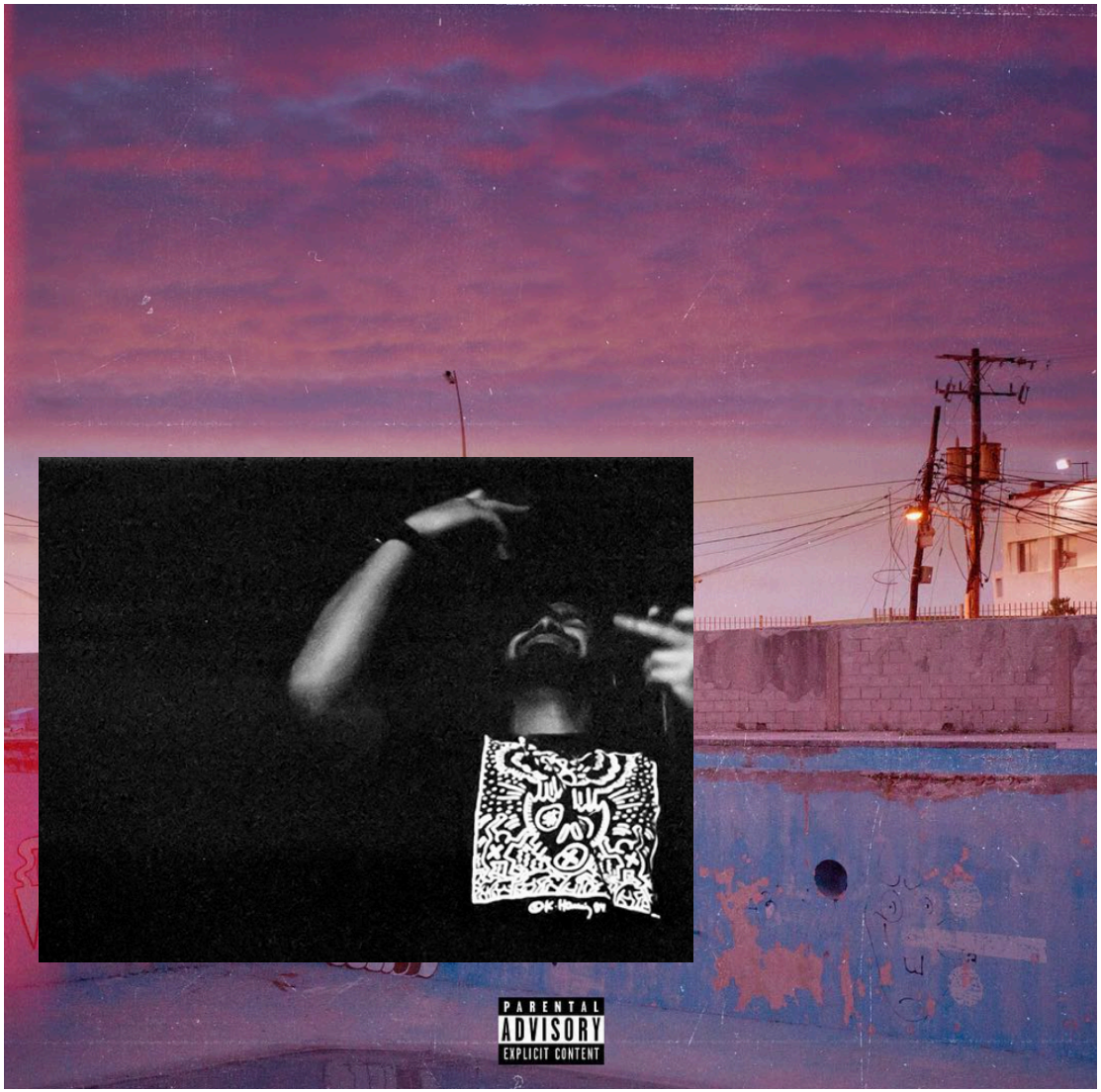
Repetition range:
43% - 91%

Genre

Genre	Repetition Percentage
Pop	**71.9 %**
Country	70.2%
Rap	68.4%
Hip Hop	**65.4%**
Electronic	74.3%
R&B	70.9%

Part of Speech Tagging

Use sentence configuration to classify songs



How does it work?

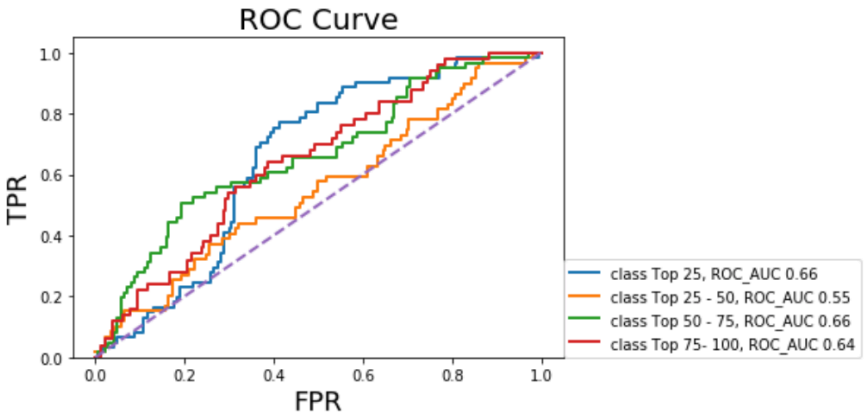
"I don't mind if you wanna go anywhere"

```
('I', 'PRP'),
('don', 'VBP'),
("'", "''"),
('t', 'JJ'),
('mind', 'NN'),
('if', 'IN'),
('you', 'PRP'),
('wanna', 'VBP'),
('go', 'VB'),
('anywhere', 'RB')
```

Top Tags Breakdown by Groups

Part of Speech Tags	Top 25	Top 25 -50	Top 50 - 75	Top 75 - 100
Nouns	**203.52**	182.65	175.96	195.88
Personal Pronouns	**48.49**	43.98	41.92	39.00
Preposition	**40.32**	33.55	37.50	38.19
Adverb	**29.77**	26.34	23.89	24.33
Determiner	34.41	33.82	32.72	**35.98**
Quotes	**34.19**	31.22	29.62	28.20
Comma	**38.89**	31.43	30.48	35.51

Repetition, Genres & Part of Speech Tags



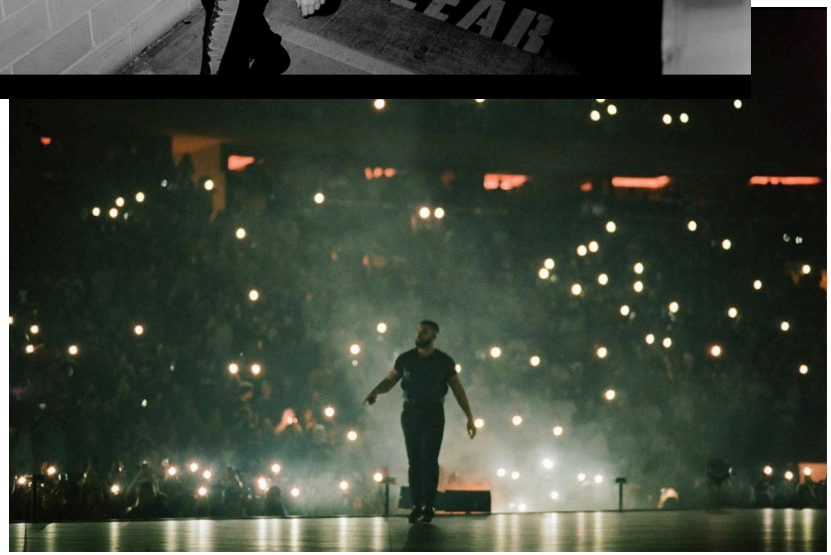
Predicted	Top 25	Top 25 - 50	Top 50 - 75	Top 75 -100
Top 25	32	8	5	16
Top 25 – 50	23	21	6	9
Top 50 - 75	17	14	18	12
Top 75 - 100	20	15	2	13

SVC Model
C = 1.06
Kernel = rbf

Score : 40%

All Features Together

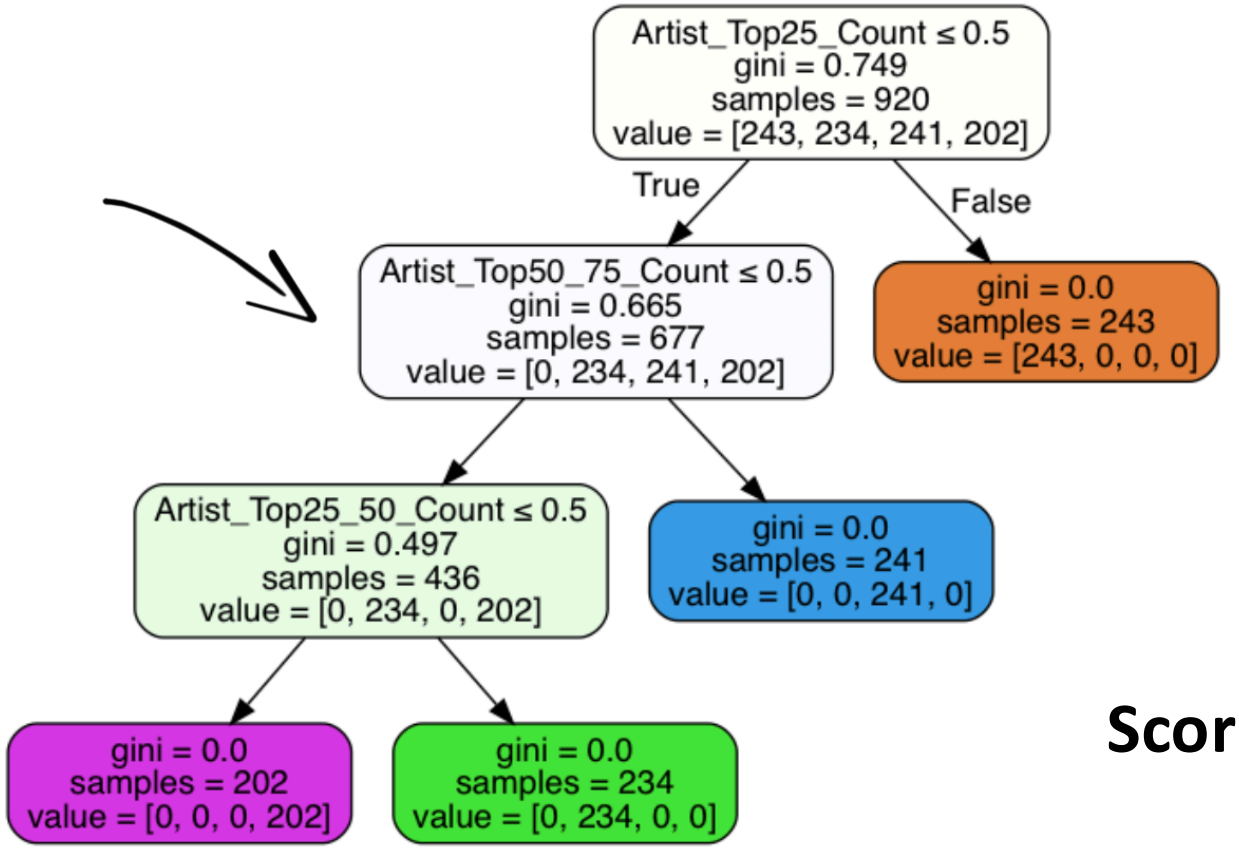
Using all NLP created features, how accurately can you predict the group a song belongs too?



Random Forest Decision Tree Classifier

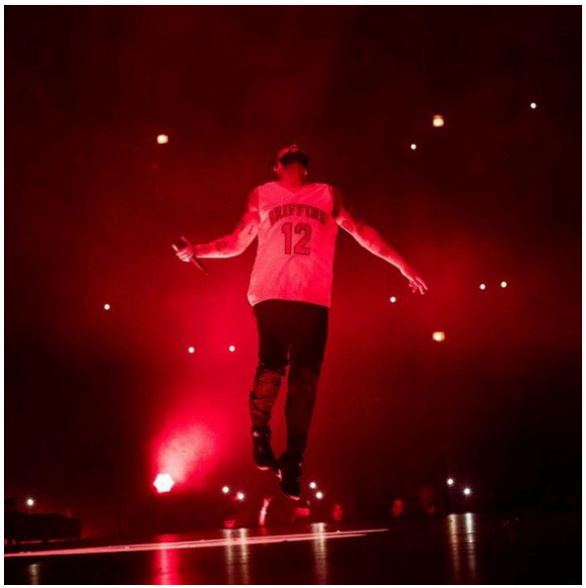
Inputs

- Artist Count per group
- Part of Speech Tags
- Repetition percentage
- Sentiment
- Genres



Score : 100 %

Results



Lyrics alone can reach a maximum of 40% accuracy in predicting the group of a song

60% of the remaining accuracy is due to the popularity of a song, and other features

Next Steps

- Look into dance-ability, energy & other features about the songs
- Create a pipeline