Spotify Trends Analysis

Merve TASKAYA
Marianne STEPHANIDES
Bernhard WEGHAUPT
Tomasz WLODARSKI

Introduction

Problem/Motivation:

- To understand trends in order to make predictions for the music industry.
- Introduction to Spotify and its importance in the music market.
- Methods used to collect and analyze data from Spotify
- Predictions and recommendations for the future

Research Questions

- How can we effectively group and categorize music genres from the Spotify dataset into a smaller number of genre clusters, and what insights can we gain from these genre groupings in terms of popularity, artist diversity, and track characteristics?

- Can trends or sentiments in song lyrics using NLP techniques be identified, and how do these trends correlate with song popularity?

State-of-the-Art/Related Work

- **Title:** Genre Classification of Spotify Songs using Lyrics, Audio Previews, and Album Artwork
- **Authors:** Dammann, Tyler, and Haugh, Kevin. *Stanford University, Fall* (2017).
- **Content:** Attempt to classify Genres of Spotify Songs by predicting the Genre using only Song Lyrics, Audio Previews and Album Artwork.
- **Models:** Recurrent Neural Network, k-Nearest Neighbors, and Naive Bayes. Output of the three models combined.
- **Similarities:** Trying to accurately predict Spotify Genres, also considering Song Lyrics.
- **Differences:** Our study includes numerous and varied genres, while this paper only considered 4 distinct Genres (Christian, Metal, Country, Rap) and also used Audio Previews for prediction.

State-of-the-Art/Related Work

- **Title:** What makes a song trend? Cluster analysis of musical attributes for Spotify top trending songs
- **Authors:** Al-Beitawi, Zayd, Mohammad Salehan, and Sonya Zhang. *Journal of Marketing Development and Competitiveness* 14.3 (2020): 79-91.
- **Content:** Attempt to discover factors which make Songs Trend on Spotify.
- **Models:** Cluster Analysis, Feature Selection, Correlation Analysis, Comparison of Clusters.
- **Similarities:** Deploying Cluster Analysis on Spotify Songs to gain more insight.
- **Differences:** Only very few Songs considered (Top 100 of 2017 and 2018). Song Lyrics not considered at all. Genre not considered for Cluster Analysis.

State-of-the-Art/Related Work

- **Title:** Classification of musical genre: A machine learning approach
- **Authors:** Basili, Roberto, Alfredo Serafini, and Armando Stellato*.* *International Society for Music Information Retrieval*. 2004.
- **Content:** Classification of Songs into Genres.
- **Models:** Not specified, only referred to as Machine Learning techniques.
- **Similarities:** Extracting genre information from song data.
- **Differences:** MIDI transcriptions of the Songs used for genre classification. Musical Instrument Digital Interface - Contains melodic, rhythmic, and structural aspects of songs. Also, much smaller dataset (only about 300 song files).

State-of-the-Art/Related Work

- **Title:** Musical genre classification by ensembles of audio and lyrics features.
- **Authors:** Mayer, Rudolf, and Andreas Rauber. *Proceedings of international conference on music information retrieval*. 2011.
- **Content:** Musical genre classification by combining song lyrics with acoustic data.
- **Models:** Classifier Ensemble Techniques, Support Vector Machines
- **Similarities:** Genre Classifications, Song Lyrics considered as well.
- **Differences:** Considering acoustic data for genre classification. No sentiment analysis.

State-of-the-Art/Related Work

- **Title:** Machine Learning Approach for Genre Prediction on Spotify Top Ranking Songs.
- **Authors:** Luo, Kehan. University of North Carolina, (2018).
- **Content:** Classifying songs into Genres based on their audio features.
- **Models:** Principal Component Analysis, OneVsRest Classifier
- **Similarities:** Using Spotify Data for Genre Prediction.
- **Differences:** Much smaller Dataset, focusing on Daily Top-Ranking Songs on Spotify. Not considering song lyrics.

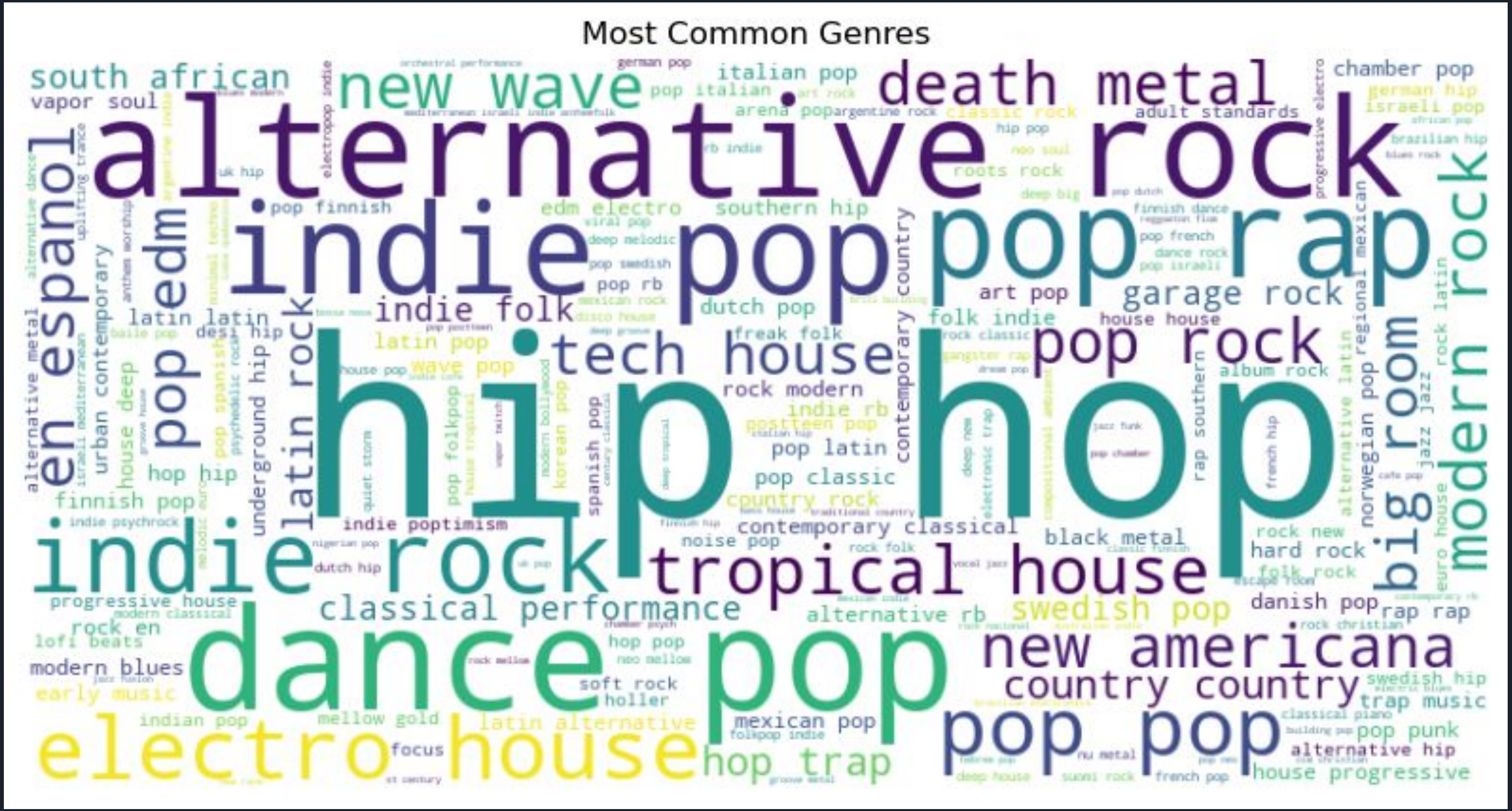
Our Data

- main dataset □ spotify_tracks
  - Audio features available:
    - Accousticness, dancebility
  - Genres (missing for 42% of dataset)
  - Lyrics (available in different languages)
- spotify_artists
  - Popularity
  - followers

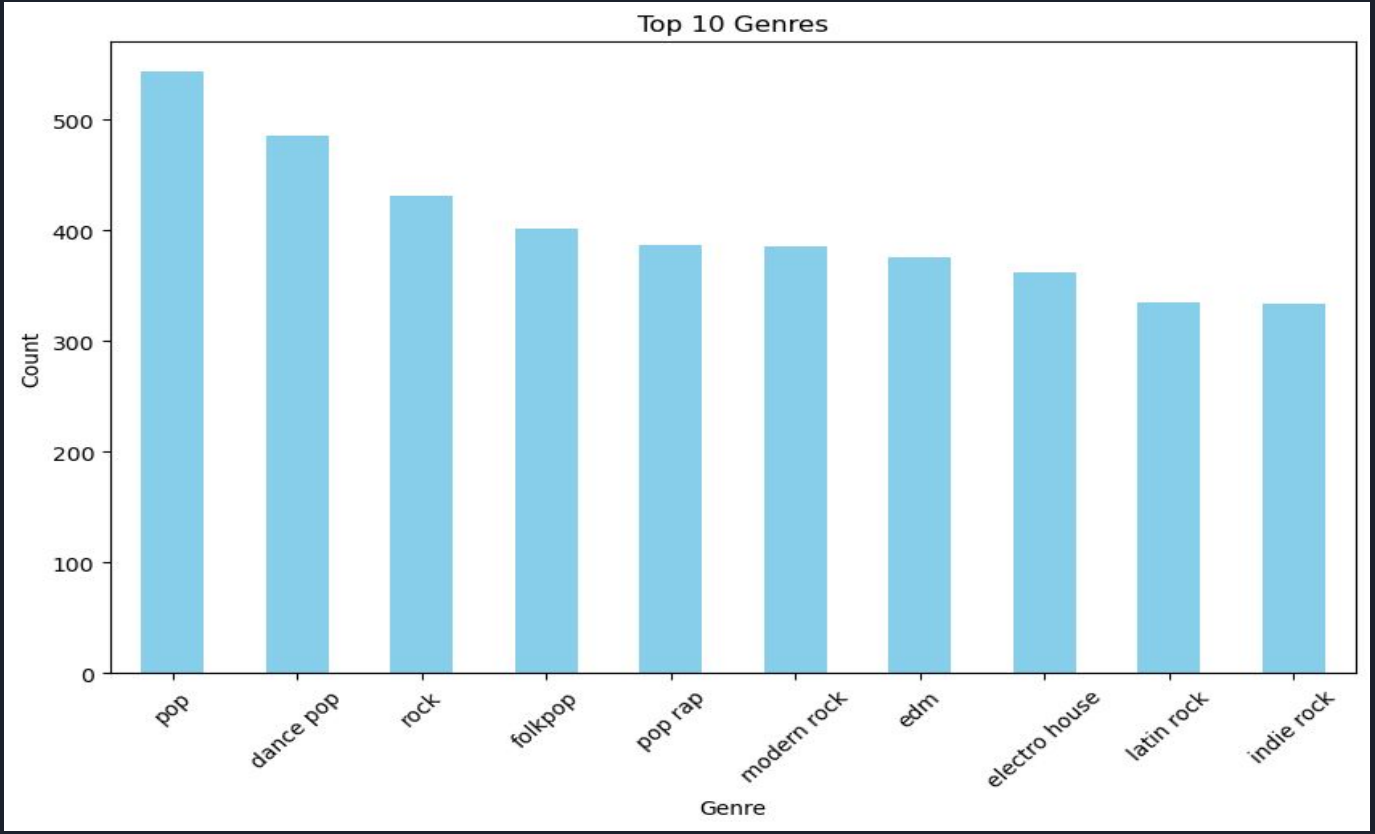
Preprocessing Steps - Genre

- Genres:
    - convertion into list
    - lowercase
    - remove characters
    - Tokenization
    - No „classic“ stopwords

Word Cloud - Genres



Top Genres


Top 10 Genres
Count
0
100
200
300
400
500
pop
dance pop
rock
folkpop
pop rap
modern rock
edm
electro house
latin rock
indie rock
Genre

Preprocessing Steps - Lyrics

1. Replace ‚\r\n‘ with ‘ ‘
2. Detect language
3. Translate lyrics into english (failed)
4. Lowercasing
5. Remove numbers and punctuation
6. Tokenization
7. Removing Stopwords (updated stopwords with music related ones)
8. Removing rows with non-english characters
9. Keeping only songs that are 90s-600s long and have 500-6000 words

Why did translation fail?

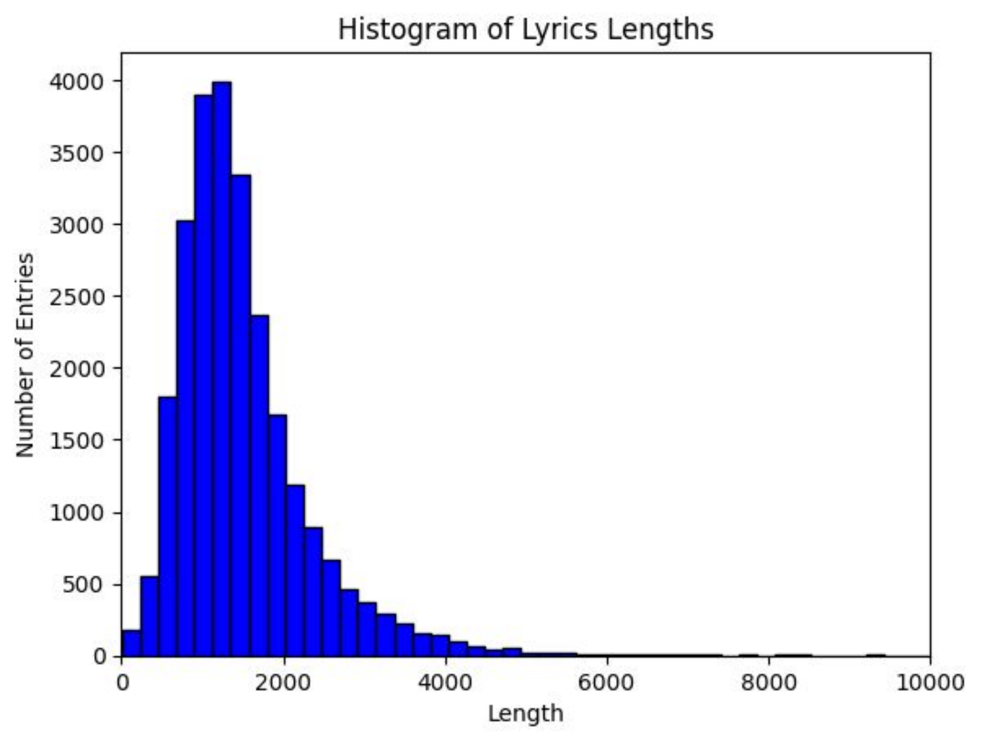
package googletrans:
- took forever
- api connection failed

DeepL API:
- Free version limited to 500.000 characters
- Pro version costs 4.99€ + + Usage-based pricing (20€ per 1.000.000 characters)
- Non-english lyrics consist of around 22.000.000+ characters

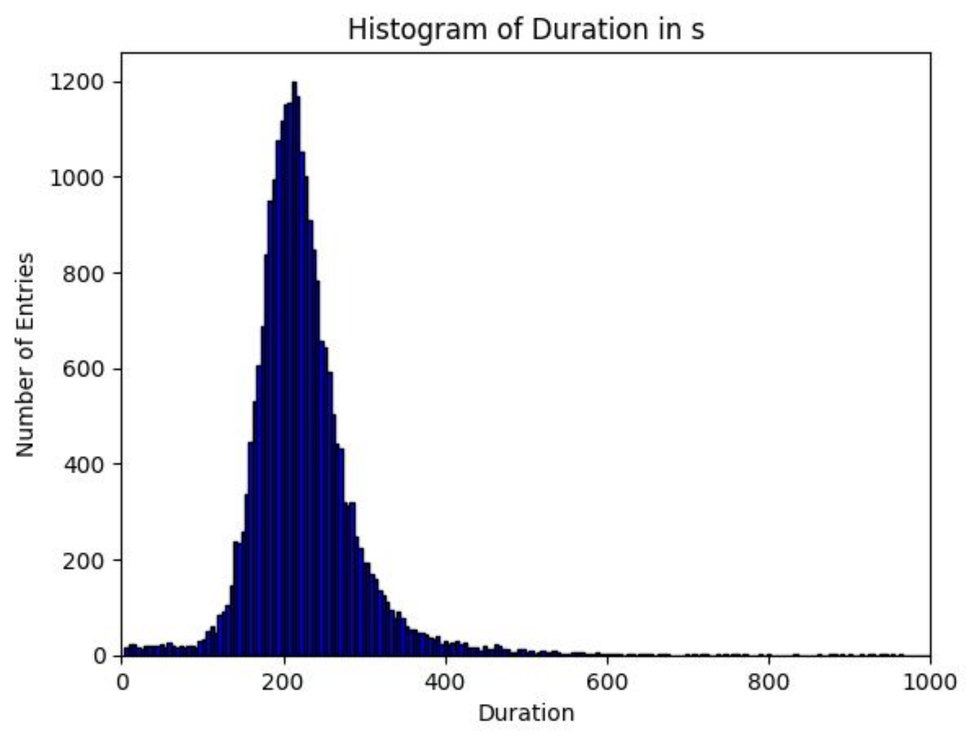
Statistics

Number of Documents: 23265
Total Number of Words: 8164428
Average Document Length: 350.93 words
Minimum Document Length: 80 words
Maximum Document Length: 1582 words
Vocabulary Size: 84558 unique words
Filtered Words after Stopword Removal: 4590429 words
Average Sentence Length: 350.93 words

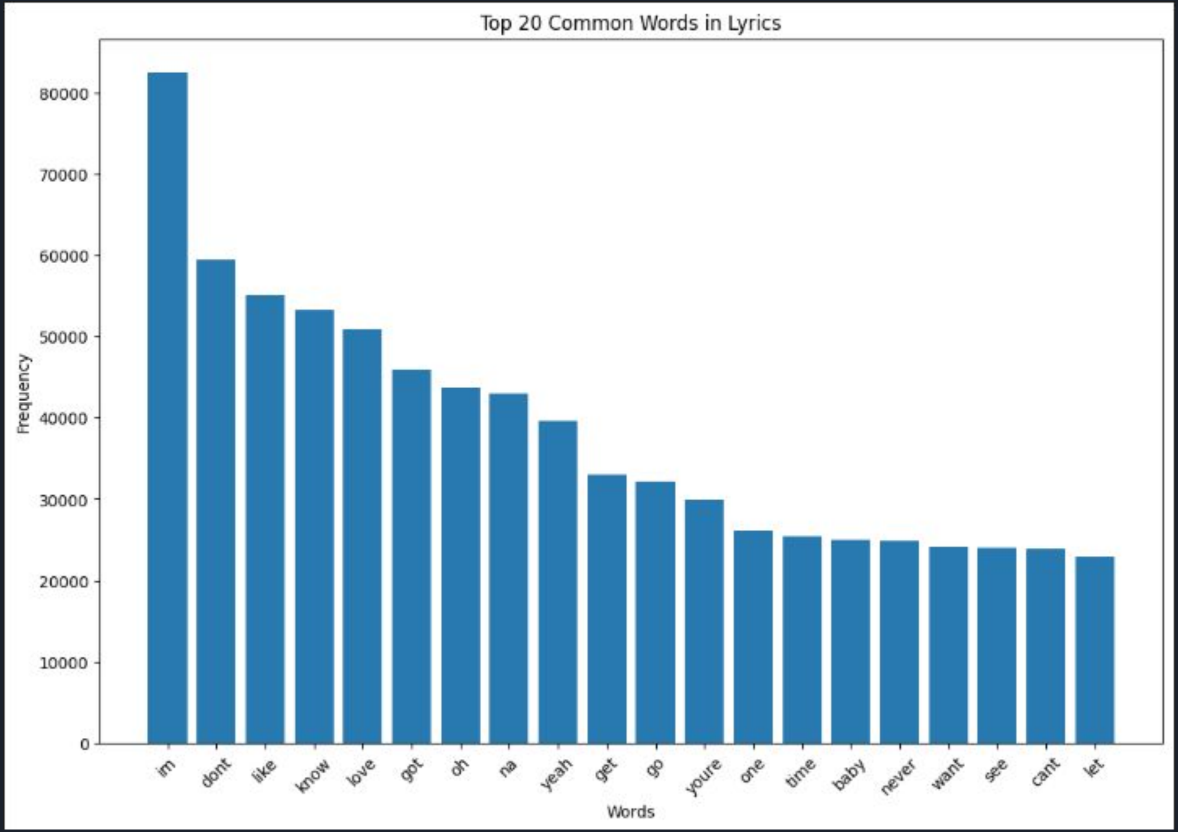
Lyrics length



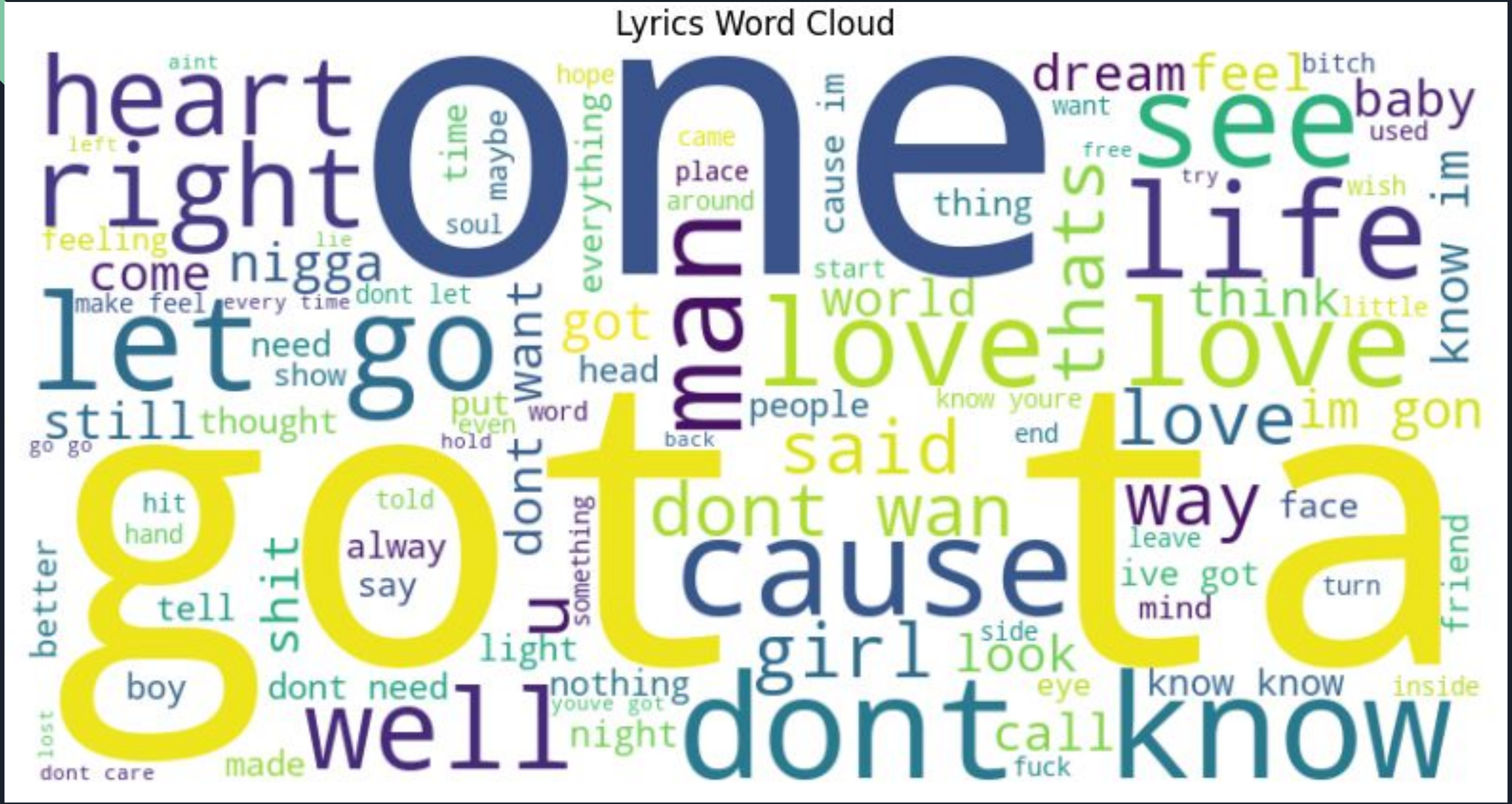
Song duration in s



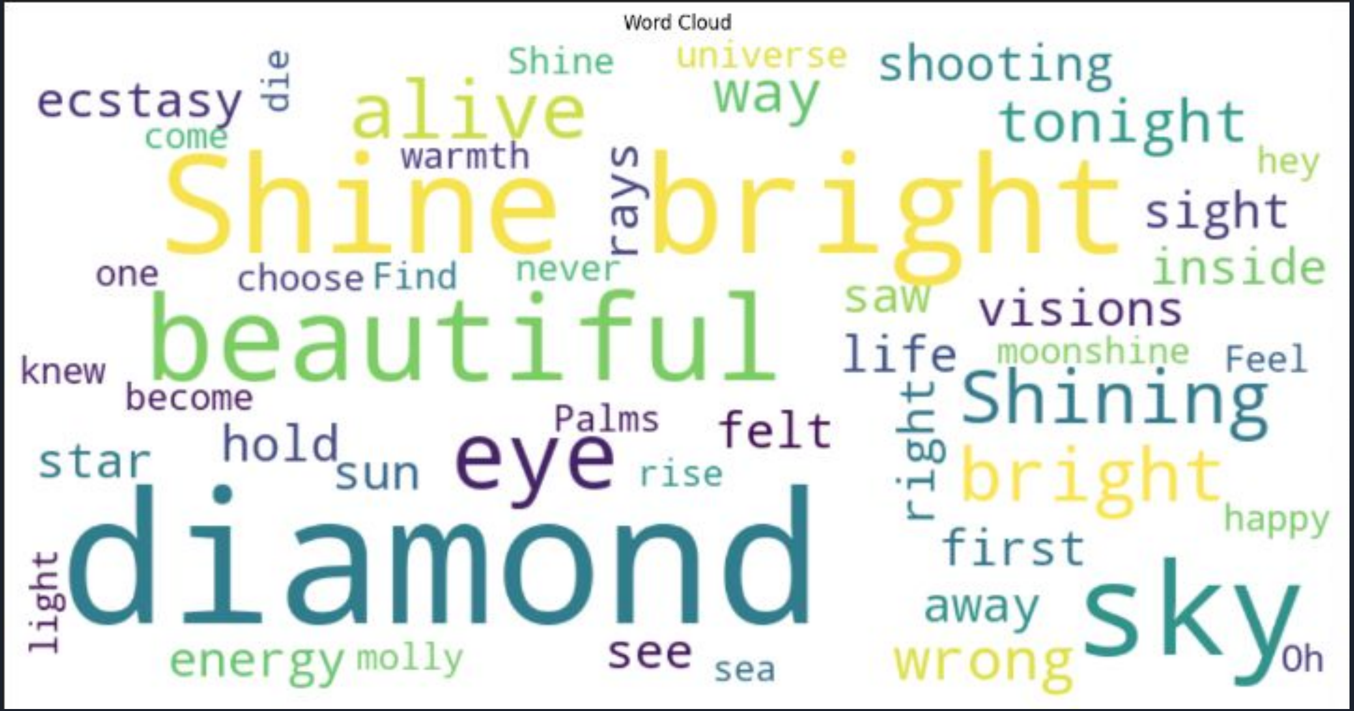
Top 20 words


Top 20 Common Words in Lyrics
Frequency
0
10000
20000
30000
40000
50000
60000
70000
80000
Words
im
dont
like
know
love
got
oh
na
yeah
get
go
youre
one
time
baby
never
want
see
cant
let

Word Cloud - all lyrics



Word Cloud - Diamonds by Rihanna



Feature Engineering

- Lyrics:
    - X One Hot Encoding
    - X Bag-of-Words
    - ✓ TF-IDF
    - ✓ Distributed Representations:
        - Word2Vec, GloVe, fastText
        - Doc2Vec
        - SBERT
    - X Hand-Crafted Features

Models

- Genre:
    - cleaned genre (2453 genres)
    - stemmed genre (2453 genres)
    - tokenized genre (1417 genres)

Model Evaluation

- Genres:
  - ✓ One Hot Encoding
  - ✓ Bag-of-Words
  - ✓ TF-IDF
  - X N-Grams
  - X Distributed Representations
  - X Hand-Crafted Features

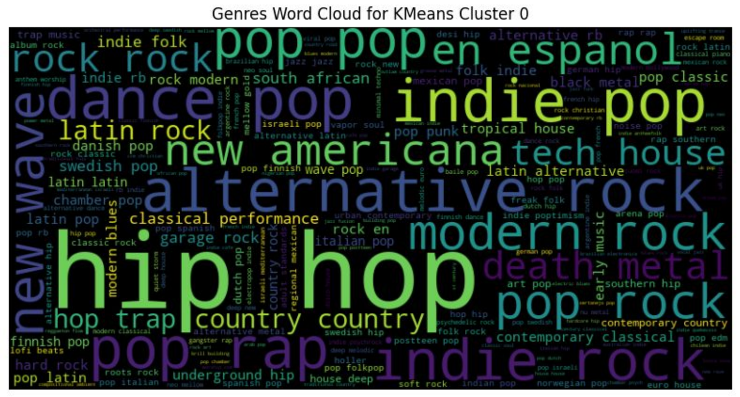
Research Question 1

- How can we effectively group and categorize music genres from the Spotify dataset into a smaller number of genre clusters, and what insights can we gain from these genre groupings in terms of popularity, artist diversity, and track characteristics?

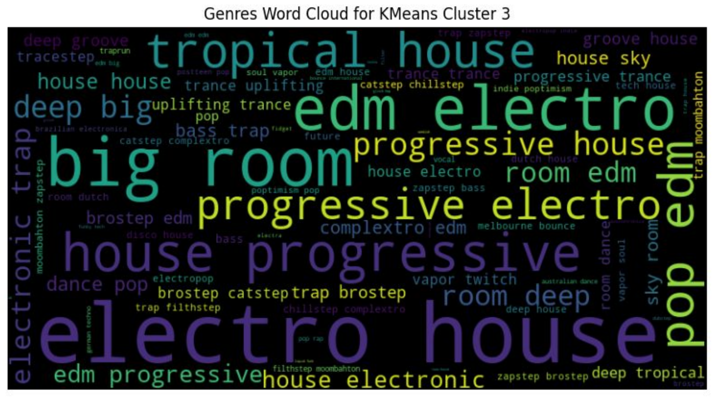
Clustering- Genres - OHE/TFIDF

- DBSCAN: using an optimized epsilon – (~750 OHE- 1300 TF-IDF CLusters) – tried to define a max of 10 clusters-> no epsilon found.
- Note: there is als hdbscan library- which seems nice ☺️
-
- Kmean Clustering: distortion/elbow method results in 5/8 Clusters – we can work with that-
- but still ´the number of clusters is not optimal and therefore results are not ideal

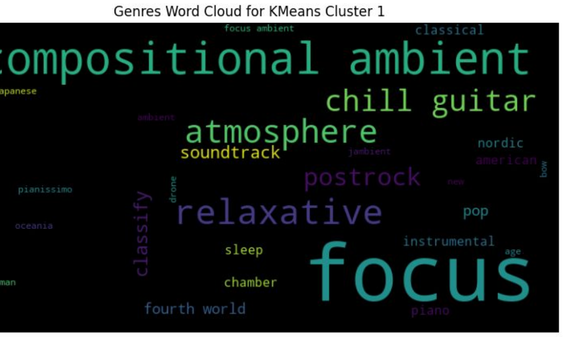
Clustering- kmeans (5-Clusters)



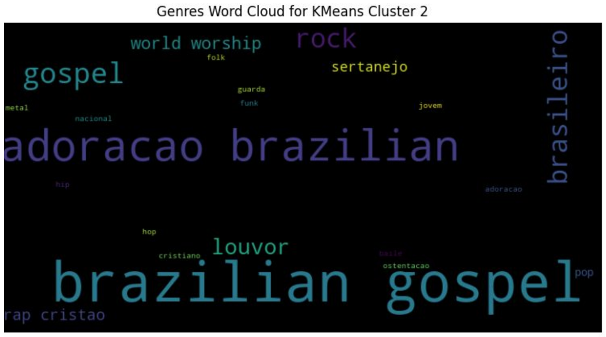
Genres Word Cloud for KMeans Cluster 0



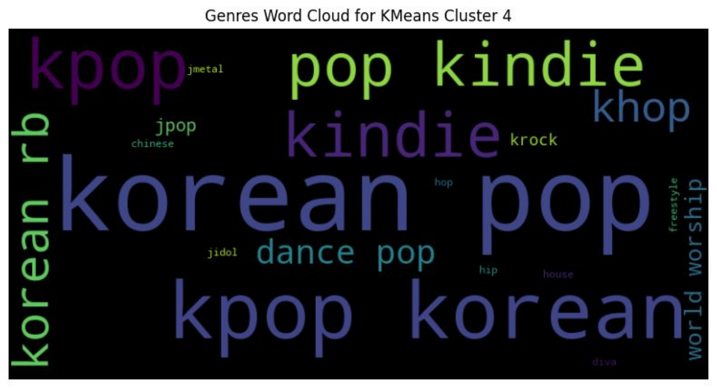
Genres Word Cloud for KMeans Cluster 3



Genres Word Cloud for KMeans Cluster 1



Genres Word Cloud for KMeans Cluster 2



Genres Word Cloud for KMeans Cluster 4

5 Clusters Descriptive

Cluster 0- Hip Hop/Rock:

- **Average Popularity: 40**
- **Average Followers: 128,009**
- **Unique Artists: 31,403**
- **Unique Genres: 2,451**

Cluster Brazilian- Gospel:

- Average Popularity: 46
- Average Followers: 206,862
- Unique Artists: 97
- Unique Genres: 15

●Cluster 1 – Background Music:

- Average Popularity: 51
- Average Followers: 6,926
- Unique Artists: 273
- Unique Genres: 25

Cluster 3 –Electro-House:

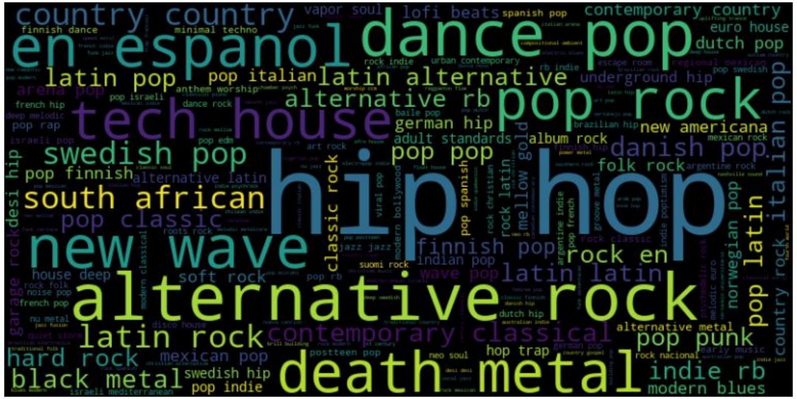
- Average Popularity: 58
- Average Followers: 408,663
- Unique Artists: 487
- Unique Genres: 117

Cluster 4- Korean/pop:

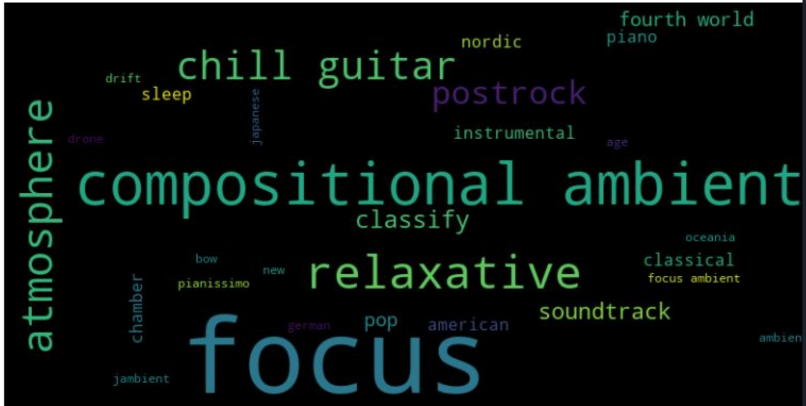
- Average Popularity: 48
- Average Followers: 270,439
- Unique Artists: 329
- Unique Genres: 14

8 Clusters

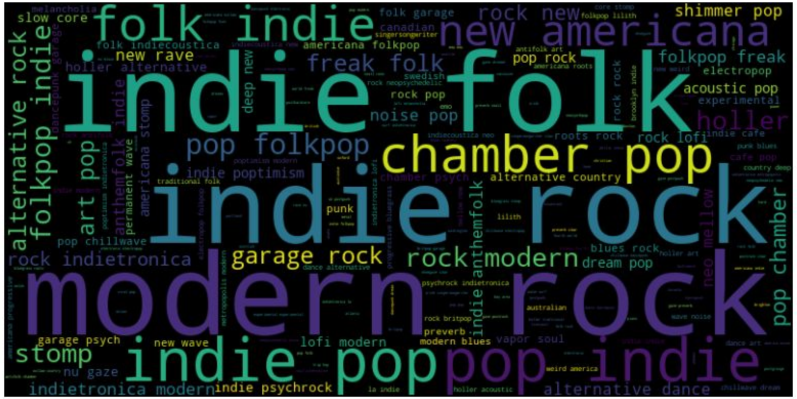
Genres Word Cloud for KMeans Cluster 0



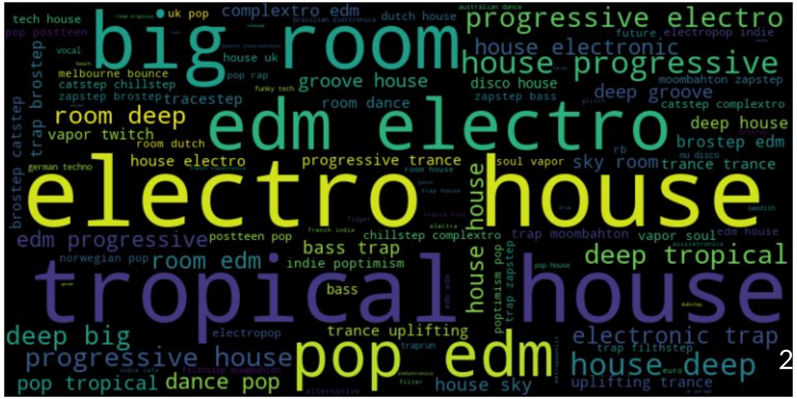
Genres Word Cloud for KMeans Cluster 2



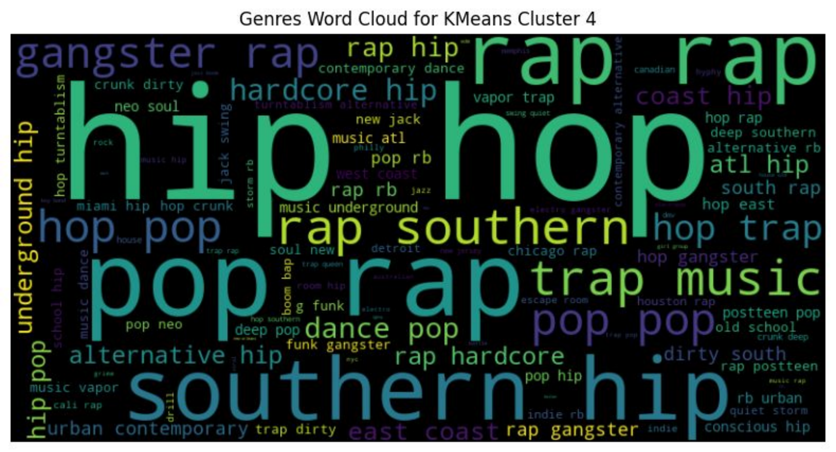
Genres Word Cloud for KMeans Cluster 1

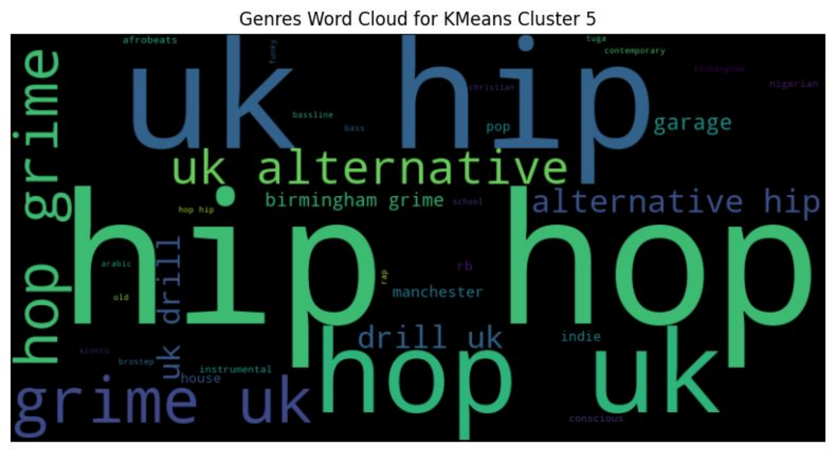


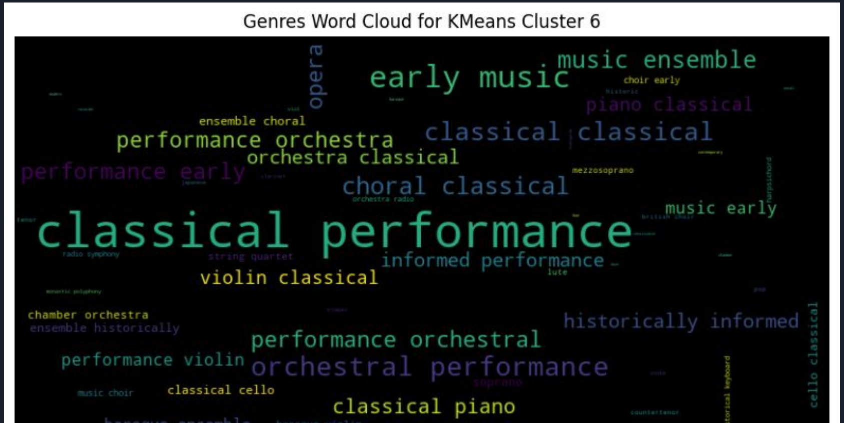
Genres Word Cloud for KMeans Cluster 3

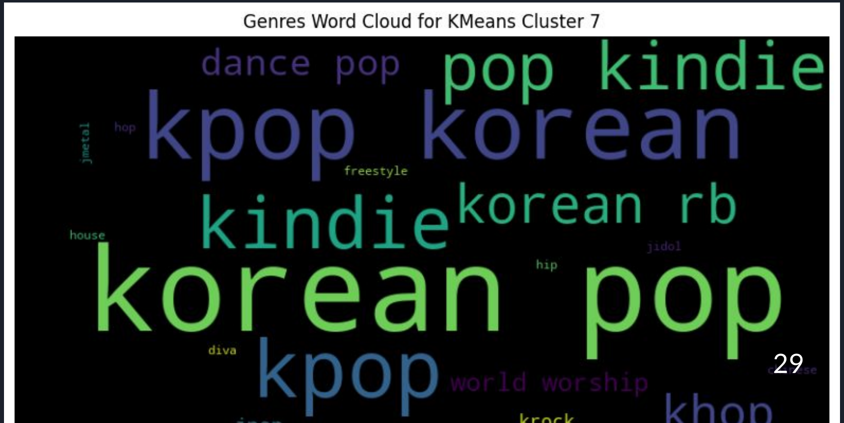


8 Clusters


Genres Word Cloud for KMeans Cluster 4


Genres Word Cloud for KMeans Cluster 5


Genres Word Cloud for KMeans Cluster 6


Genres Word Cloud for KMeans Cluster 7

8 Clusters Descriptive

Cluster 0 – Hip Hop/pop:

- **Average Popularity: 39**
- **Average Followers: 105,366**
- **Unique Artists: 29,661**
- **Unique Genres: 2,448**

Cluster 1 Indie/Rock:

- Average Popularity: 57
- Average Followers: 318,094
- Unique Artists: 691
- Unique Genres: 290

Cluster 2 - Background Music:

- Average Popularity: 51
- Average Followers: 6,926
- Unique Artists: 273
- Unique Genres: 25

Cluster 3 – Electro/tropical/House:

- Average Popularity: 58
- Average Followers: 335,089
- Unique Artists: 690
- Unique Genres: 147

8 Clusters Descriptive

Cluster 4 (hip hop) Rap:

- Average Popularity: 67
- Average Followers: 1,308,505
- Unique Artists: 497
- Unique Genres: 114

Cluster 5 - Alternative Hip Hop:

- Average Popularity: 53
- Average Followers: 59,565
- Unique Artists: 146
- Unique Genres: 28

Cluster 6 - Background Music:

- Average Popularity: 45
- Average Followers: 12,037
- Unique Artists: 302
- Unique Genres: 62

Cluster 7 – Korean Pop:

- Average Popularity: 48
- Average Followers: 270,439
- Unique Artists: 329
- Unique Genres: 14

Conclusion Clustering

- It did not work well - Many Genres might make sense
- Different Clustering Methods and combinations of other information would probably improve the results

Research Question 2

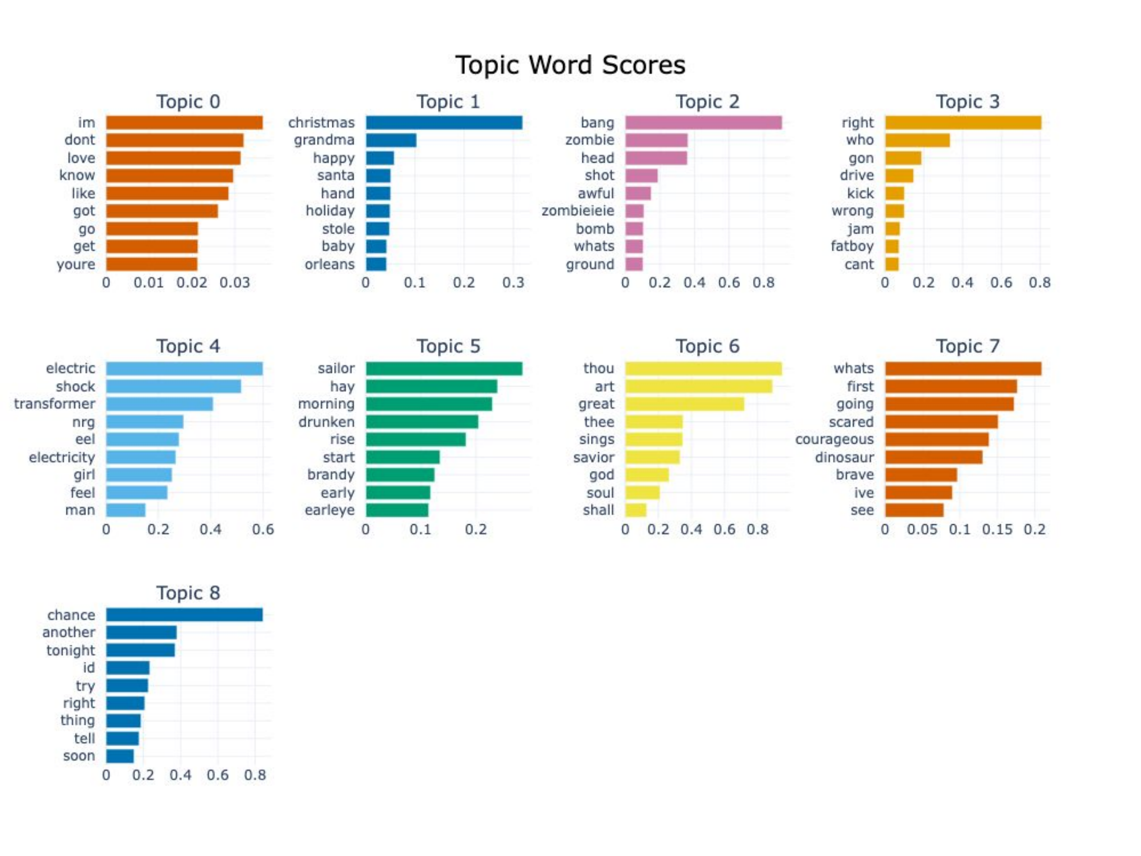
- Can trends or sentiments in song lyrics using NLP techniques be identified, and how do these trends correlate with song popularity?

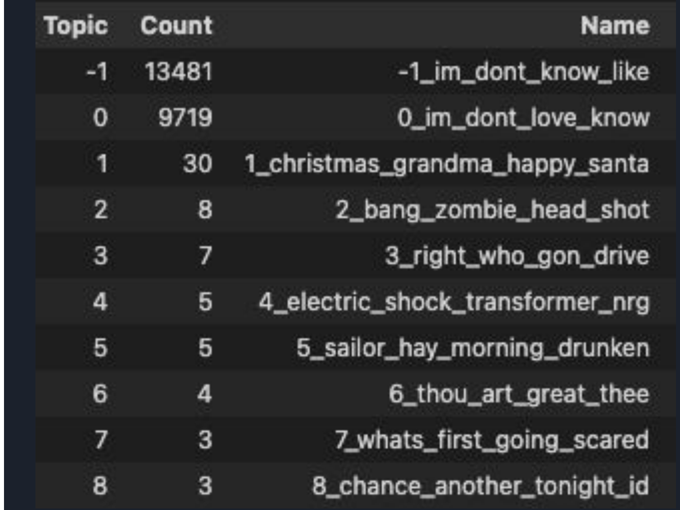
What did we do?

1. Topic Modelling with BERTopic and LDA (Latent Dirichlet Allocation)

2. Sentiment Analysis using TextBlob and Vader (Valence aware dictionary for sentiment reasoning)

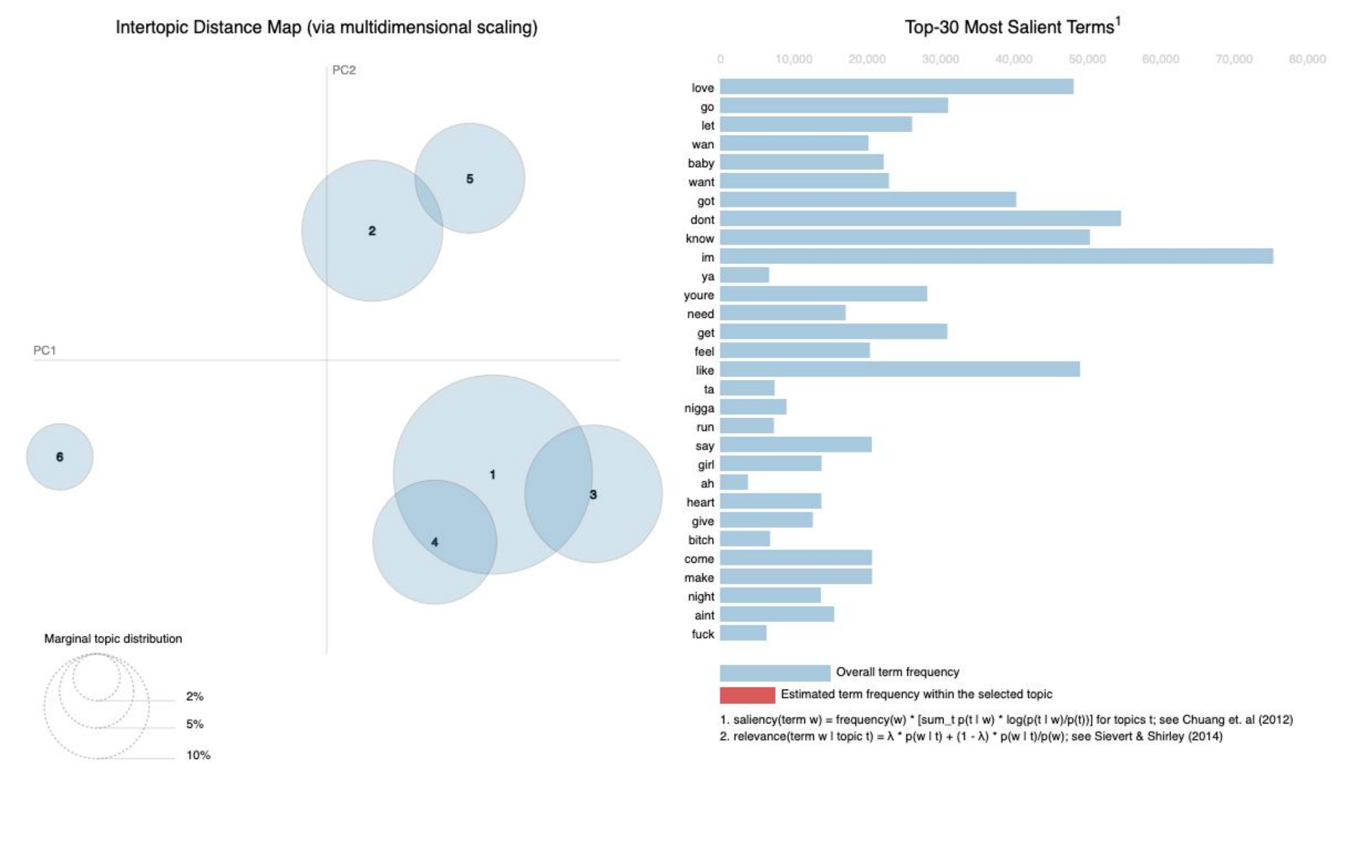
Topic Modelling - Bertopic





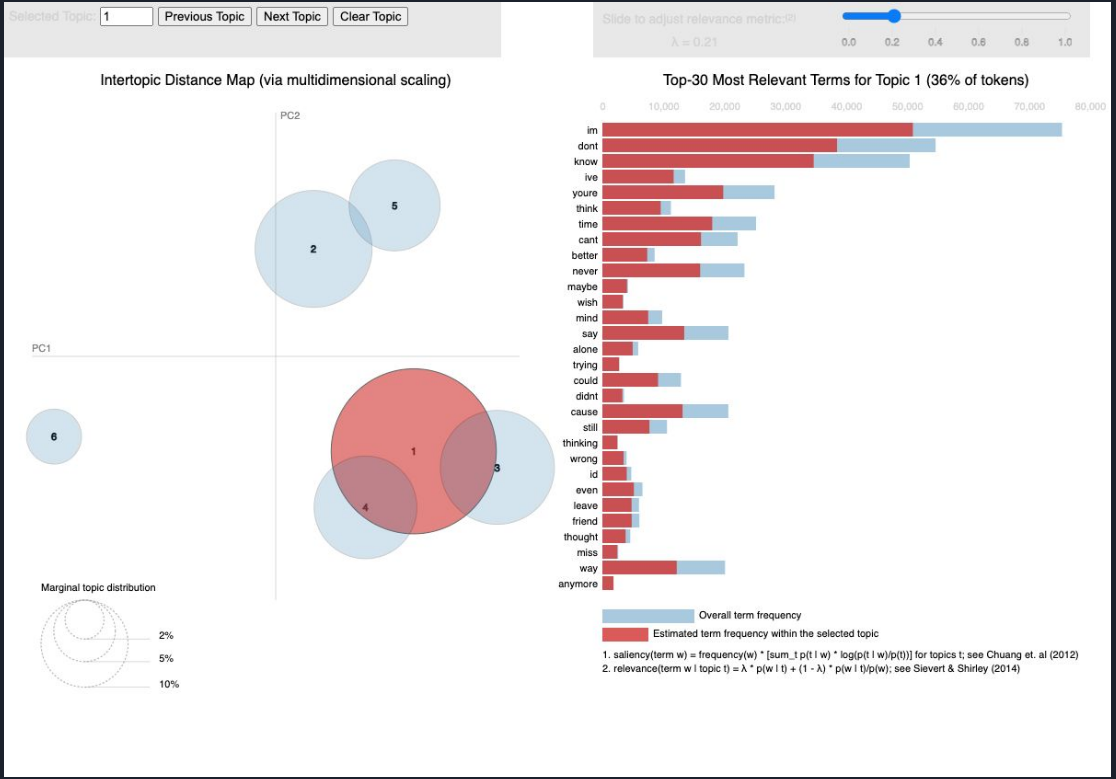
Topic	Count	Name
-1	13481	-1_im_dont_know_like
0	9719	0_im_dont_love_know
1	30	1_christmas_grandma_happy_santa
2	8	2_bang_zombie_head_shot
3	7	3_right_who_gon_drive
4	5	4_electric_shock_transformer_nrg
5	5	5_sailor_hay_morning_drunken
6	4	6_thou_art_great_thee
7	3	7_whats_first_going_scared
8	3	8_chance_another_tonight_id

Topic Modelling - LDA



1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

T1: Reflexionen und Emotionen in persönlichen Beziehungen

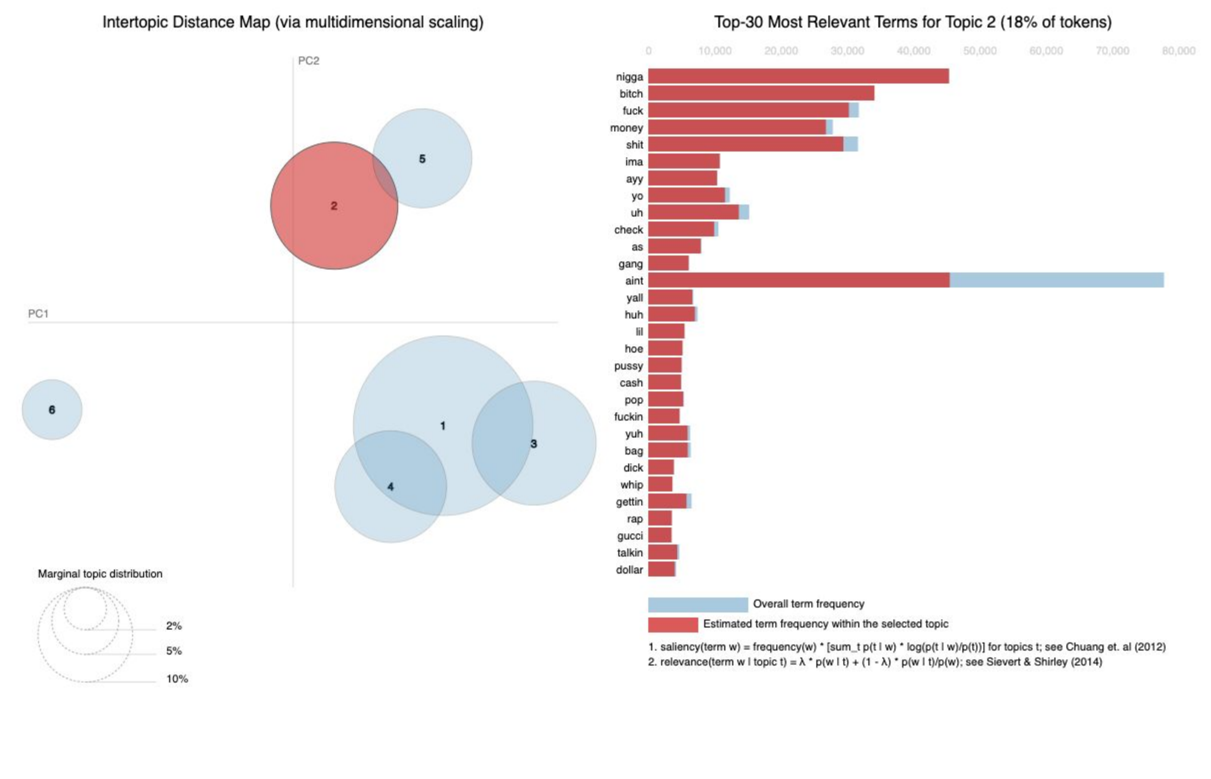


Selected Topic: 1

Slide to adjust relevance metric:(2) λ = 0.21

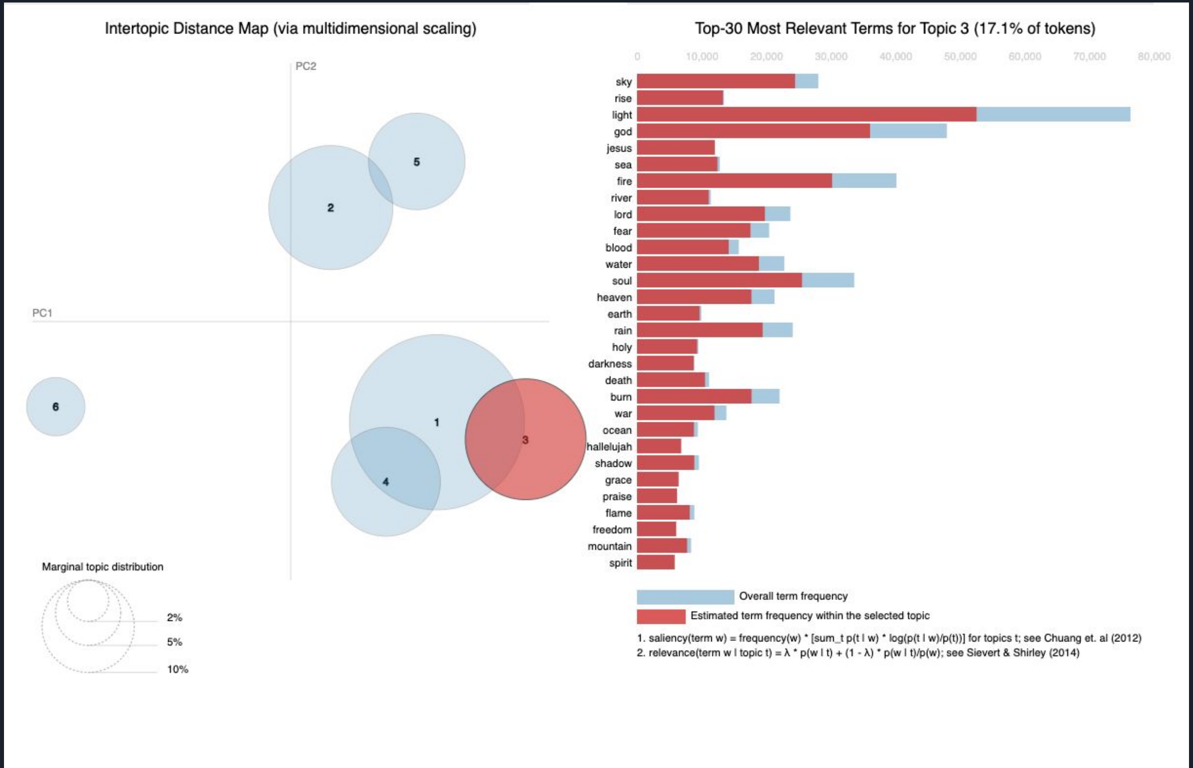
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

T2: Selbstdarstellung und Materialismus im zeitgenössischen Hip-Hop



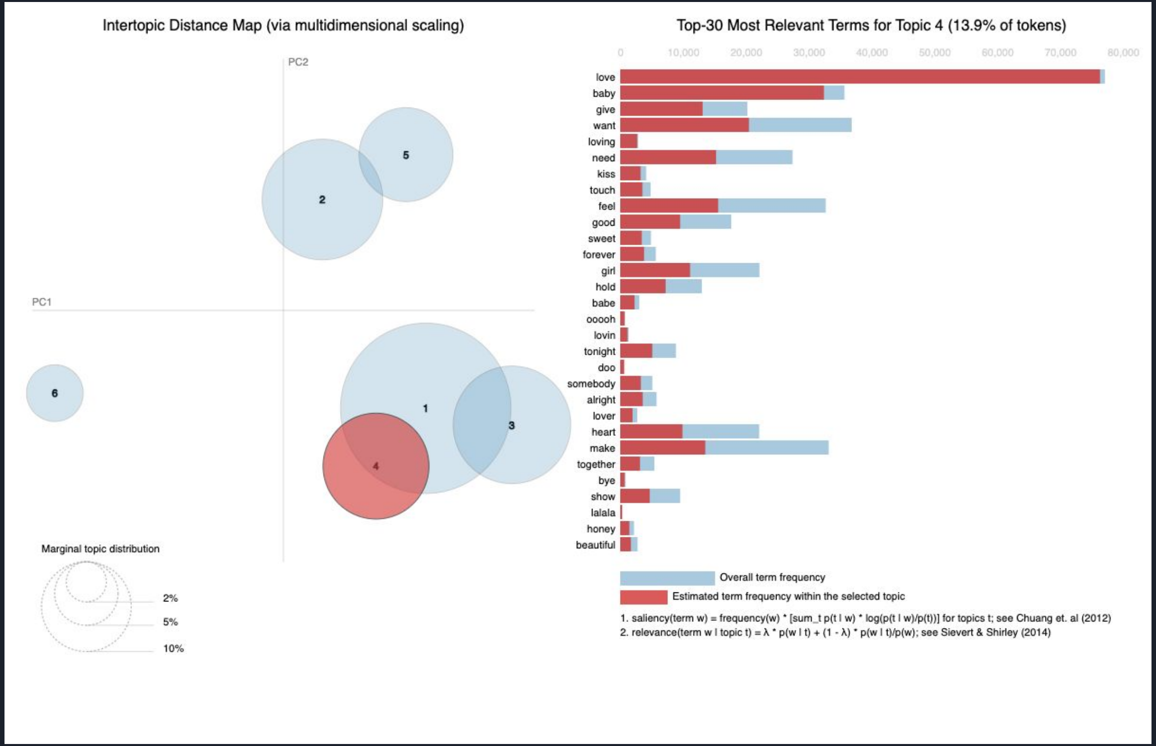
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

T3: Spirituelle Erhebung und die Suche nach Transzendenz



1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

T4: Romantik und Zuneigung

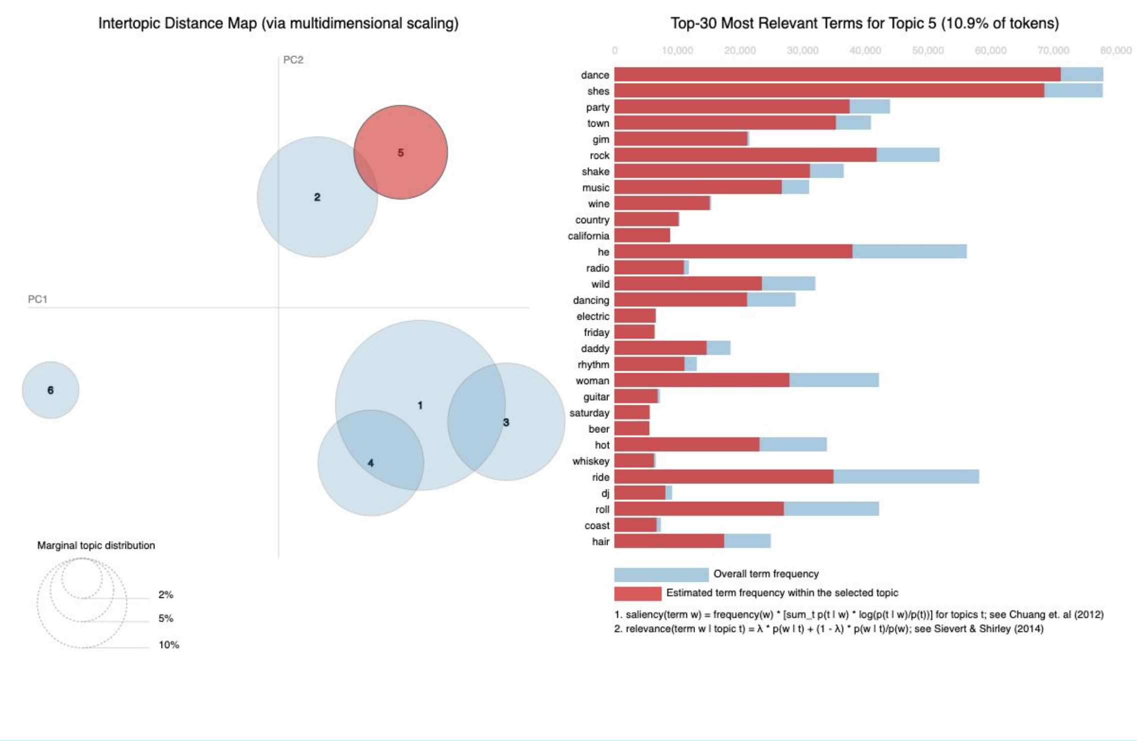


Intertopic Distance Map (via multidimensional scaling)

Top-30 Most Relevant Terms for Topic 4 (13.9% of tokens)

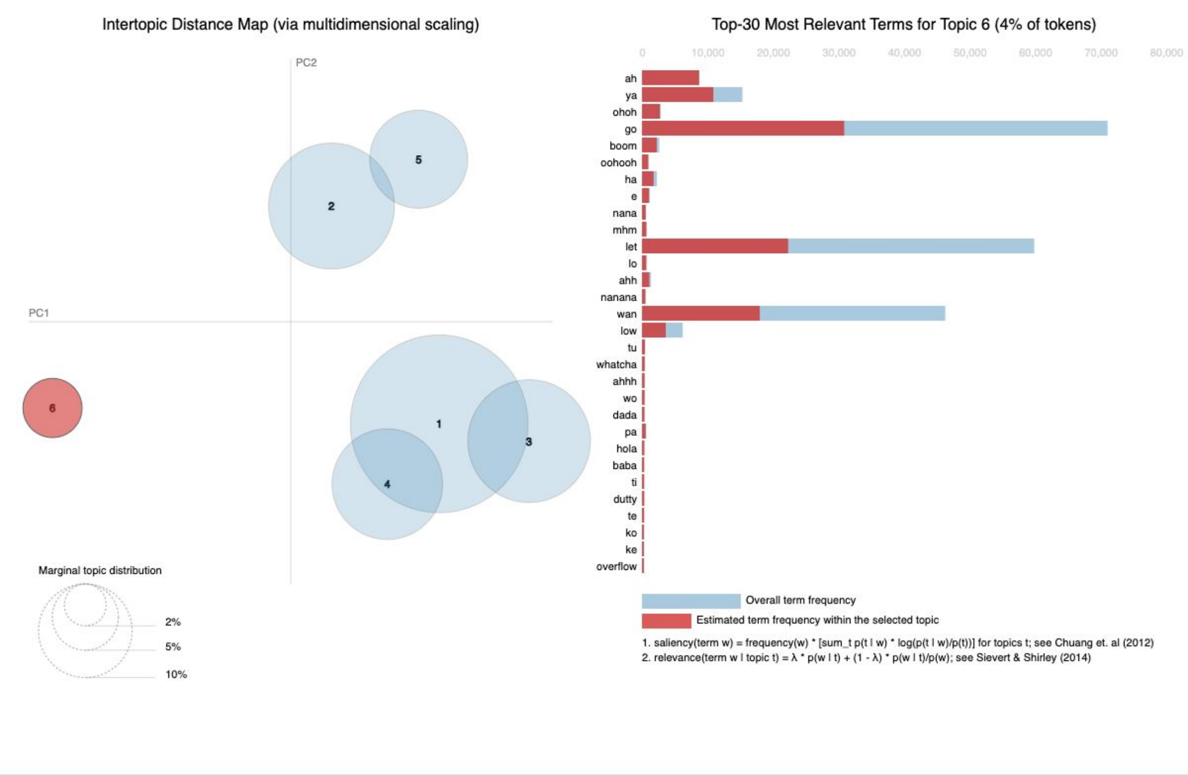
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

T5: Feier und Ausgelassenheit der Nacht



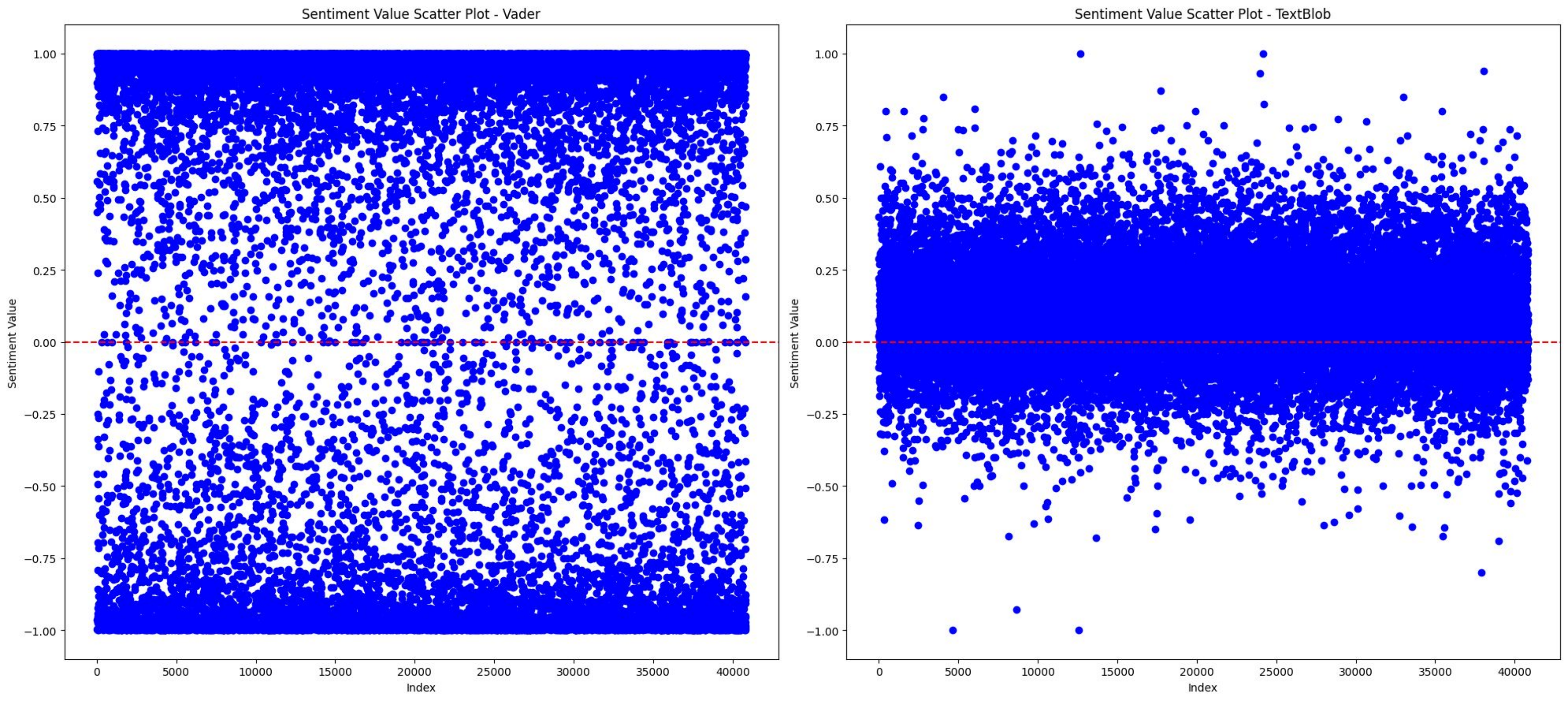
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

T6: Rhythmische Vibes und musikalische Energie

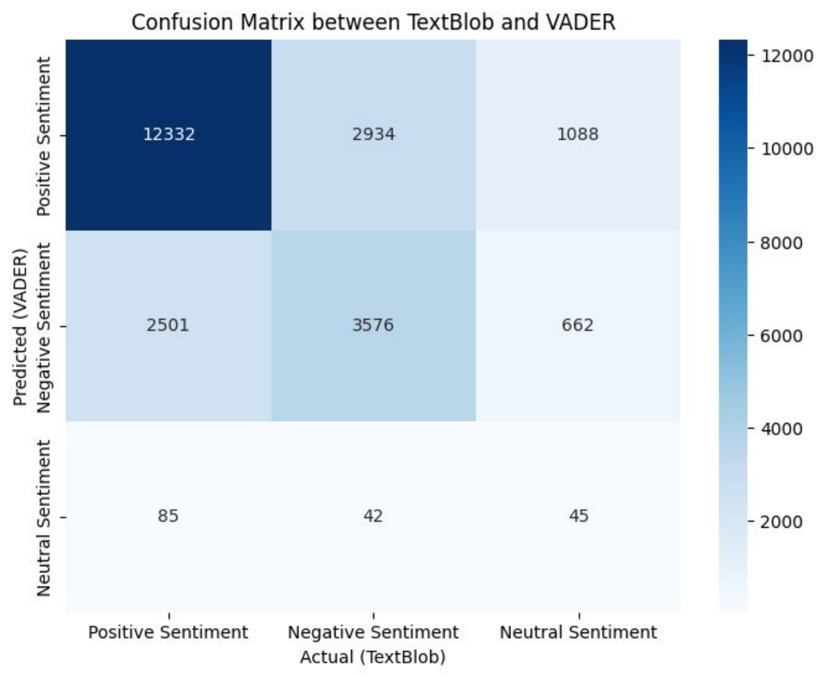


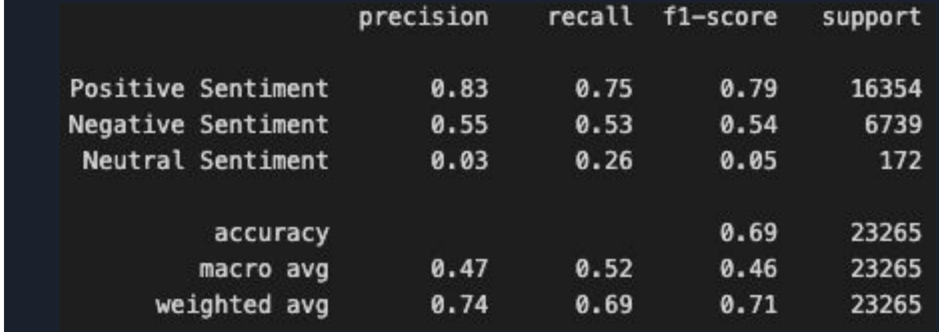
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Sentiment Analysis using Vader and TextBlob



Confusion Matrix Vader and TextBlob



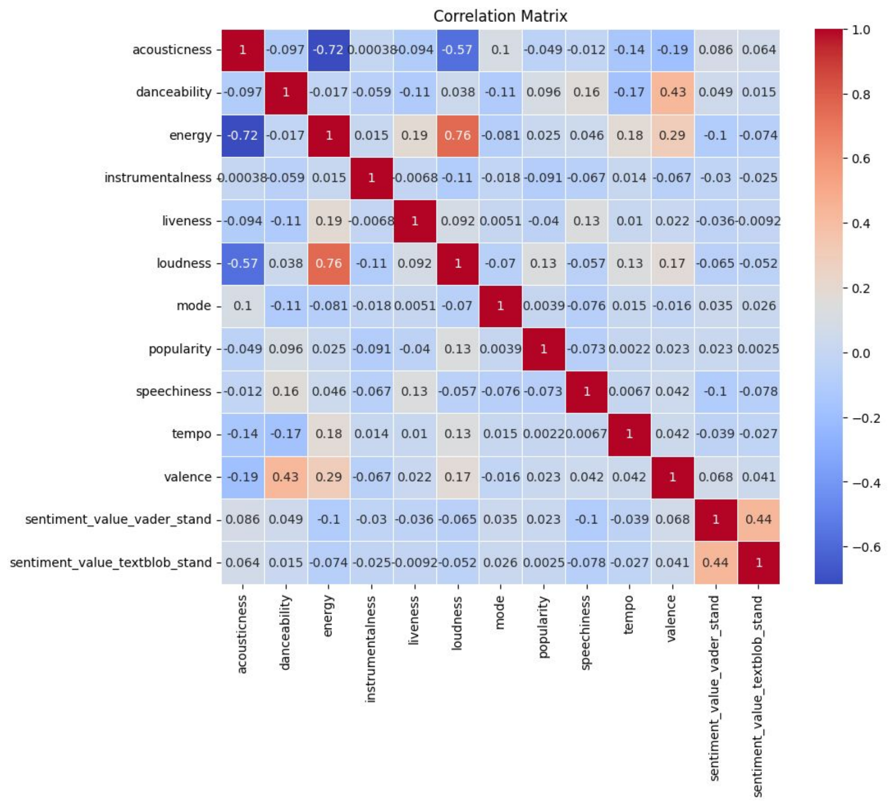


```
                    precision    recall  f1-score   support

Positive Sentiment       0.83      0.75      0.79     16354
Negative Sentiment       0.55      0.53      0.54      6739
 Neutral Sentiment       0.03      0.26      0.05       172

          accuracy                           0.69     23265
         macro avg       0.47      0.52      0.46     23265
      weighted avg       0.74      0.69      0.71     23265
```

No correlation between sentiment and music features



Correlation Matrix

	acousticness	danceability	energy	instrumentalness	liveness	loudness	mode	popularity	speechiness	tempo	valence	sentiment_value_vader_stand	sentiment_value_textblob_stand
acousticness	1	-0.097	-0.72	0.00038	-0.094	-0.57	0.1	-0.049	-0.012	-0.14	-0.19	0.086	0.064
danceability	-0.097	1	-0.017	-0.059	-0.11	0.038	-0.11	0.096	0.16	-0.17	0.43	0.049	0.015
energy	-0.72	-0.017	1	0.015	0.19	0.76	-0.081	0.025	0.046	0.18	0.29	-0.1	-0.074
instrumentalness	0.00038	-0.059	0.015	1	-0.0068	-0.11	-0.018	-0.091	-0.067	0.014	-0.067	-0.03	-0.025
liveness	-0.094	-0.11	0.19	-0.0068	1	0.092	0.0051	-0.04	0.13	0.01	0.022	-0.036	-0.0092
loudness	-0.57	0.038	0.76	-0.11	0.092	1	-0.07	0.13	-0.057	0.13	0.17	-0.065	-0.052
mode	0.1	-0.11	-0.081	-0.018	0.0051	-0.07	1	0.0039	-0.076	0.015	-0.016	0.035	0.026
popularity	-0.049	0.096	0.025	-0.091	-0.04	0.13	0.0039	1	-0.073	0.0022	0.023	0.023	0.0025
speechiness	-0.012	0.16	0.046	-0.067	0.13	-0.057	-0.076	-0.073	1	0.0067	0.042	-0.1	-0.078
tempo	-0.14	-0.17	0.18	0.014	0.01	0.13	0.015	0.0022	0.0067	1	0.042	-0.039	-0.027
valence	-0.19	0.43	0.29	-0.067	0.022	0.17	-0.016	0.023	0.042	0.042	1	0.068	0.041
sentiment_value_vader_stand	0.086	0.049	-0.1	-0.03	-0.036	-0.065	0.035	0.023	-0.1	-0.039	0.068	1	0.44
sentiment_value_textblob_stand	0.064	0.015	-0.074	-0.025	-0.0092	-0.052	0.026	0.0025	-0.078	-0.027	0.041	0.44	1

Conclusion and Future Work

- Song Lyrics are difficult to classify, even for humans.
- Managed to derive some interesting topics
- Try different embeddings for BERTopic
- Try different methods (LSA, PLSA, Ida2VEc)

- No correlation between Lyrics Sentiments and Music Features
- Try different methods for Sentiment Analysis

Thank You!

Descriptives OHE

Cluster 0 - Hip Hop:

- Average Popularity: 38.22
- Average Followers: 18,985
- Unique Artists: 60
- Unique Genres: 1

Cluster 1- german pop:

- Average Popularity: 48.92
- Average Followers: 94,705
- Unique Artists: 53
- Unique Genres: 1

Cluster 2:

- Average Popularity: 40.71
- Average Followers: 133,550
- Unique Artists: 32,373
- Unique Genres: 2,453

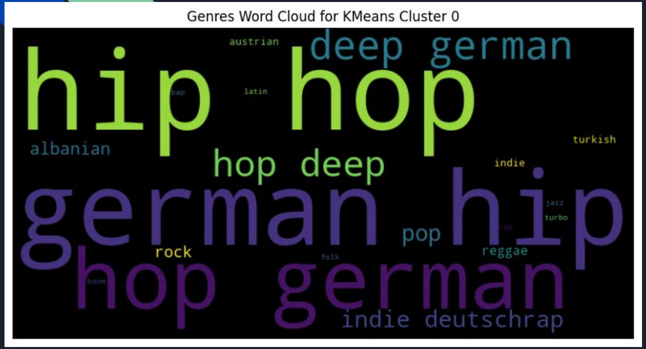
Cluster 3- psych afro:

- Average Popularity: 15.43
- Average Followers: 1,427
- Unique Artists: 28
- Unique Genres: 1

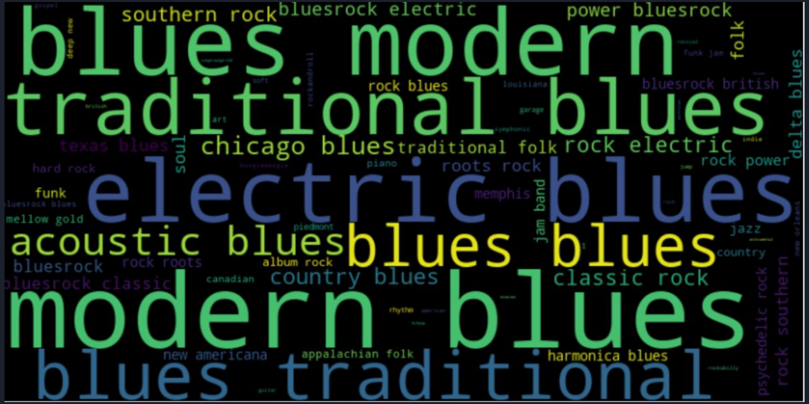
Cluster 4- orchestra:

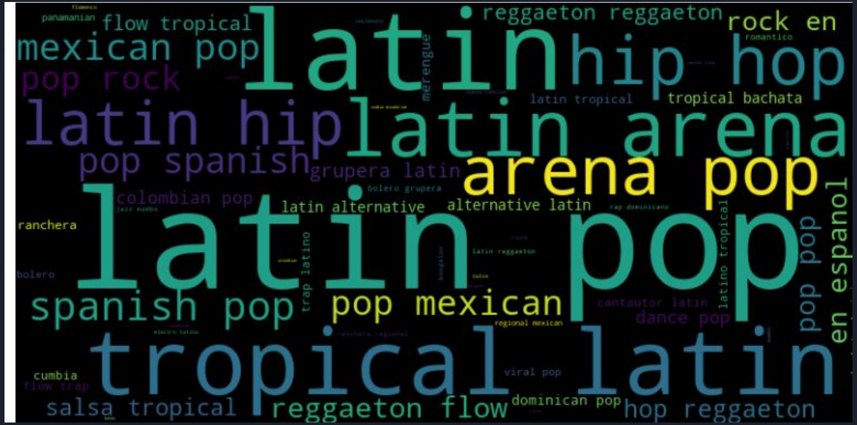
- Average Popularity: 35.65
- Average Followers: 2,295
- Unique Artists: 75
- Unique Genres: 1

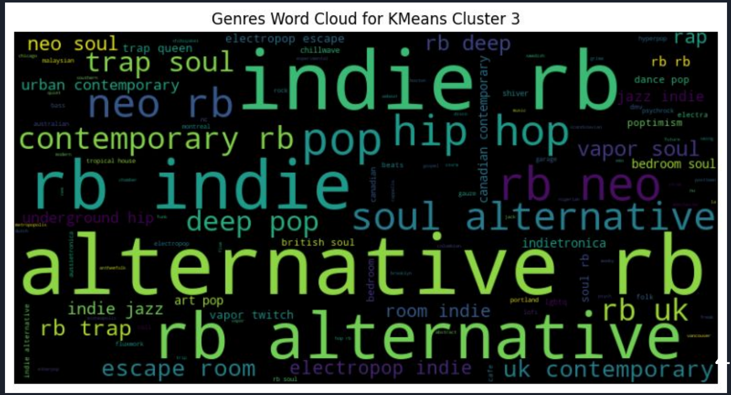
Genres –Clustering- kmeans– 7 Clusters



Genres Word Cloud for KMeans Cluster 0

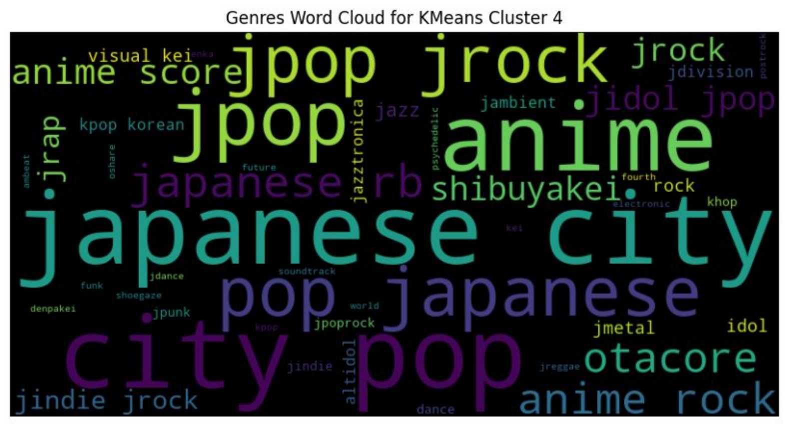


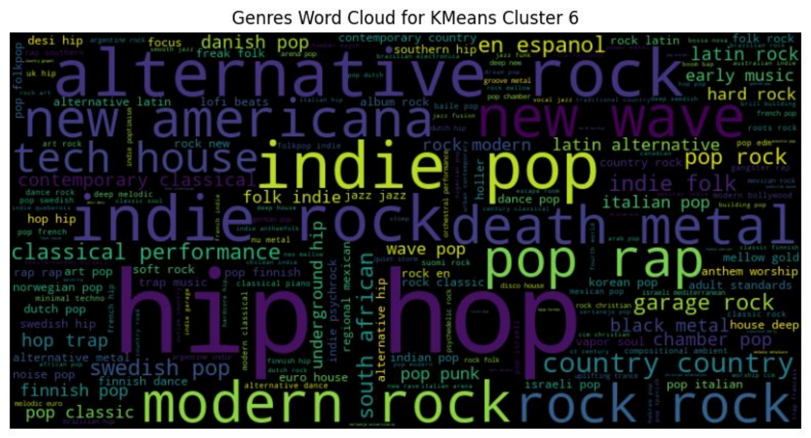


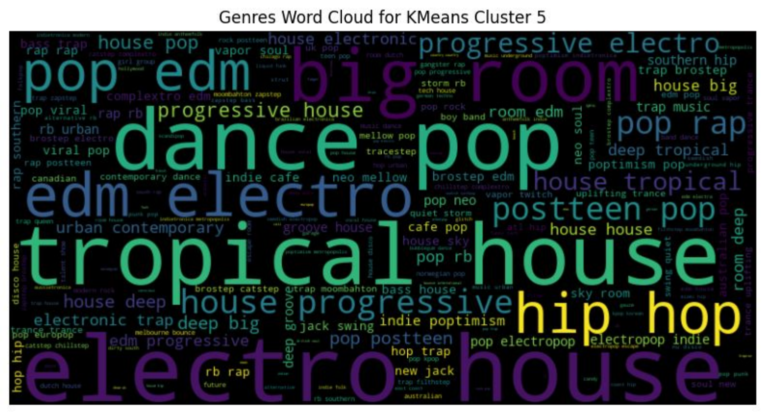


Genres Word Cloud for KMeans Cluster 3

Genres –Clustering- kmeans– 7 Clusters

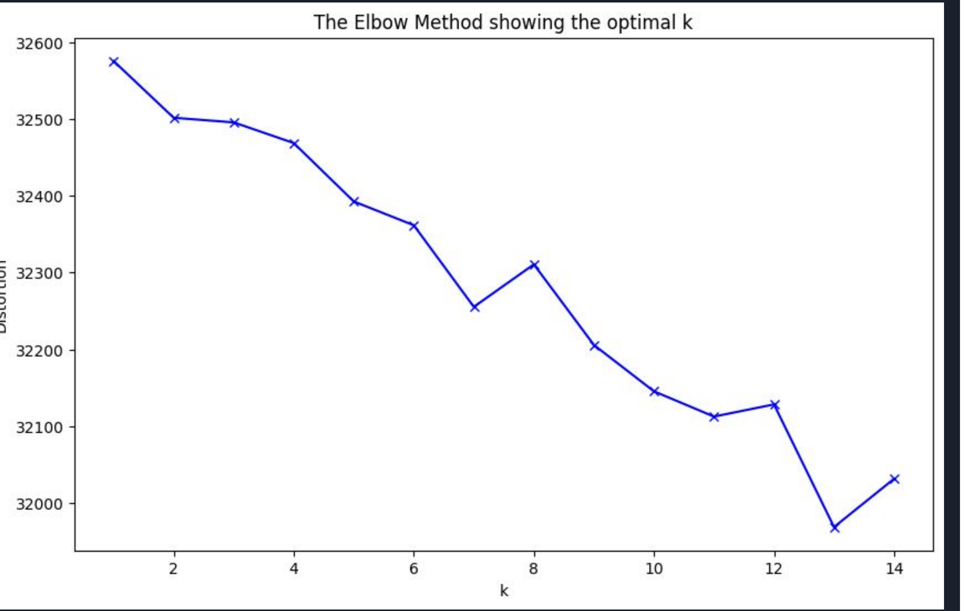

Genres Word Cloud for KMeans Cluster 4


Genres Word Cloud for KMeans Cluster 6

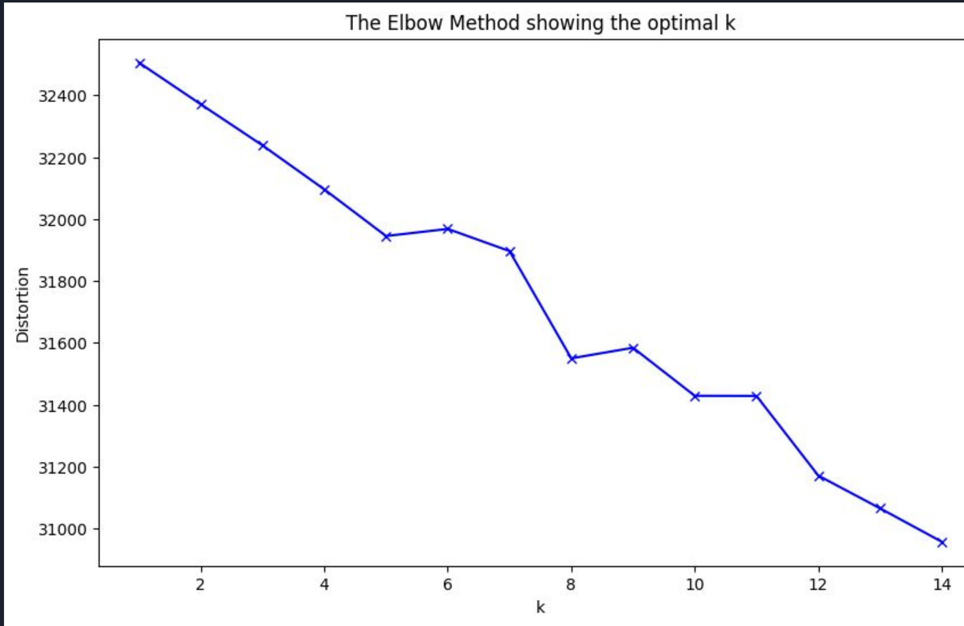

Genres Word Cloud for KMeans Cluster 5

Elbow Method

OHE



TF-IDF


The Elbow Method showing the optimal k
Distortion
k