# Assignment 2 COSC 2671 (Social Media and Network Analytics) Report

**Topic : How Does Social Media Drive Music Consumption in the Streaming Era?**

**Group members details:**

| | Student Name | Student ID |
|---|---|---|
| 1. | N.Prashanth | S4016500 |
| 2. | Manntript Kaur | S4036115 |
| 3. | Zain Abbas | S4031481 |

# Table of Contents

**Introduction** ..... 3

**Method** ..... 3

**Data Collection** ..... 5

- **Spotify API (Spotify library):** ..... 5
- **Reddit API (PRAW):** ..... 8

**Data Preprocessing** ..... 11

**Analysis** ..... 13

- **Sentiment Analysis:** ..... 13
- **Topic Modeling:** ..... 16
- **Community Detection and Analysis:** ..... 25

**Conclusion** ..... 28

**Sources And References** ..... 29

## Introduction

With this project, we will compare how people feel about songs on Reddit to how popular they are on Spotify to see how social media changes how much music people listen to. The way people find and listen to music has changed because of streaming services like Spotify. It's essential to know how social media interactions affect the success of a song in the modern music business.

This project aims to use Spotify's API to find out how famous a track is, how lively it is, and how vibrant it sounds in popular mixes. Scraping posts and comments about these songs from the Reddit API (PRAW) gives it user-generated details about the music. We then use the Vader mood tool to examine how people feel about these Reddit posts. And also community detection to identify formation of different communities among all the different users and songs. We can learn about how people think about the posts this way.

The main objective is to discover any connections between how individuals feel on social media and the number of famous songs on Spotify. If we look at these trends, we want to know if good or bad relationships on social media can tell us about or change how well a song does. This study will help us understand how public arguments on sites like Reddit can change how people listen to music.

## Method

The primary sources of data for this project were Spotify and Reddit. Data extraction, sentimental analysis, and visualization methods were used to examine how social media changes music use.

1. Data Extraction:
    - Spotify API: The Spotify library was used to get information about how famous, energetic, and live songs were from Spotify playlists. This was the primary set of songs we used for study.
    - Reddit API (PRAW): The Praw tool took discussions and notes about the songs from Reddit. Based on how people interacted with the tracks, we figured out how people generally felt about them.

```
# Libraries for data extraction from spotify and Reddit loaded
import spotipy
import json
from spotipy.oauth2 import SpotifyClientCredentials
import praw
from praw.models import MoreComments
```

2. Sentiment Analysis: The Vader sentiment analyzer from the NLTK library was used to determine whether the mood in Reddit posts was positive, neutral, or negative. This research helped find patterns in how people felt about the songs.

```
# Sentiment Analysis and topic modelling libraries loaded
import nltk
nltk.download('stopwords')
from nltk.sentiment.vader import SentimentIntensityAnalyzer
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.decomposition import LatentDirichletAllocation
import pyLDAvis
import pyLDAvis.lda_model
from wordcloud import WordCloud
```

3. Community Detection : imports the `community_louvain` module from the `community` package, which is commonly used for community detection in networks using the Louvain method.

```
[52]: import pandas as pd
      import networkx as nx
      import matplotlib.pyplot as plt
      import community.community_louvain as community_louvain
      from collections import defaultdict
```

4. Platforms and Tools:

   - The study used Jupyter Notebook and tools like Matplotlib, seaborn, and Pandas to process and display the data.
   - We used the nltk tool to do tokenization, stemming, and stopword removal on Reddit comments to prepare them for sentiment analysis.

These methods helped the project by showing how Reddit opinion links to the success of a song on Spotify. This helped researchers figure out how social media affects music trends.

## Data Collection

Spotify and Reddit were the key places where data for this project came from. The reason for getting information from these places was to examine the connection between how popular a song is on Spotify and how people feel about it on Reddit. Given below is a list of the data sources, factors gathered, and APIs utilized:

### Spotify API (Spotify library) [2]:

- Purpose: To retrieve track metadata and audio features, which were used to assess the popularity and characteristics of songs.
- Data Collected:
    - Track ID: Unique identifier for each song.
    - Name: Title of the song.
    - Album: Album the song belongs to.
    - Artist: Name of the performing artist.
    - Release Date: When the song was officially released.
    - Length: The duration of the song is in milliseconds.
    - Popularity: Popularity score (0–100), reflecting how often a track is streamed.
    - Energy: A measure of intensity and activity in the song (0–1 scale).
    - Liveness: Detects the presence of an audience in the recording (0–1 scale).

These variables provide insights into the song's features and performance on the platform.

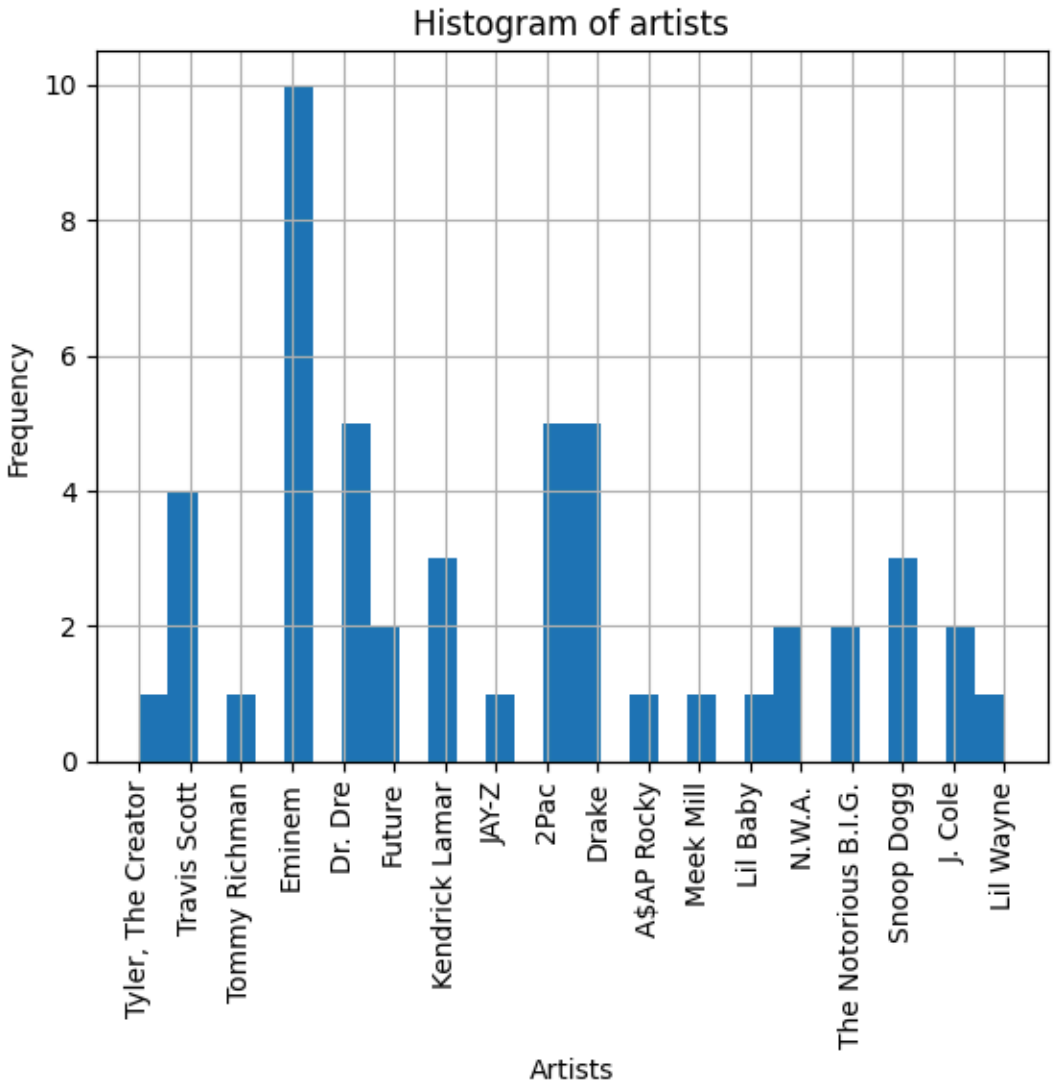
| | name | album | artist | release_date | length | popularity | energy | liveness |
|---|---|---|---|---|---|---|---|---|
| 0 | HIGHEST IN THE ROOM | HIGHEST IN THE ROOM | Travis Scott | 2019-10-04 | 175720 | 85 | 0.427 | 0.2100 |
| 1 | See You Again (feat. Kali Uchis) | Flower Boy | Tyler, The Creator | 2017-07-21 | 180386 | 89 | 0.559 | 0.1090 |
| 2 | Not Like Us | Not Like Us | Kendrick Lamar | 2024-05-04 | 274192 | 90 | 0.472 | 0.1410 |
| 3 | Without Me | The Eminem Show | Eminem | 2002-05-26 | 290320 | 86 | 0.669 | 0.2370 |
| 4 | Still D.R.E. | 2001 | Dr. Dre | 1999-11-16 | 270586 | 81 | 0.775 | 0.0543 |
| 5 | Like That | WE DON'T TRUST YOU | Future | 2024-03-22 | 267706 | 85 | 0.676 | 0.1190 |
| 6 | Ni**as In Paris | Watch The Throne | JAY-Z | 2011-08-08 | 219333 | 80 | 0.858 | 0.3490 |
| 7 | HUMBLE. | DAMN. | Kendrick Lamar | 2017-04-14 | 177000 | 84 | 0.621 | 0.0958 |
| 8 | The Real Slim Shady | The Marshall Mathers LP | Eminem | 2000-05-23 | 284200 | 85 | 0.661 | 0.0454 |
| 9 | Nuthin' But A "G" Thang | The Chronic | Dr. Dre | 1992-12-15 | 237573 | 76 | 0.821 | 0.1470 |

The Spotify data consists of 50 songs, with an average track length of approximately 244,000 milliseconds (around 4 minutes). Popularity scores range from 69 to 90, averaging 78.8,

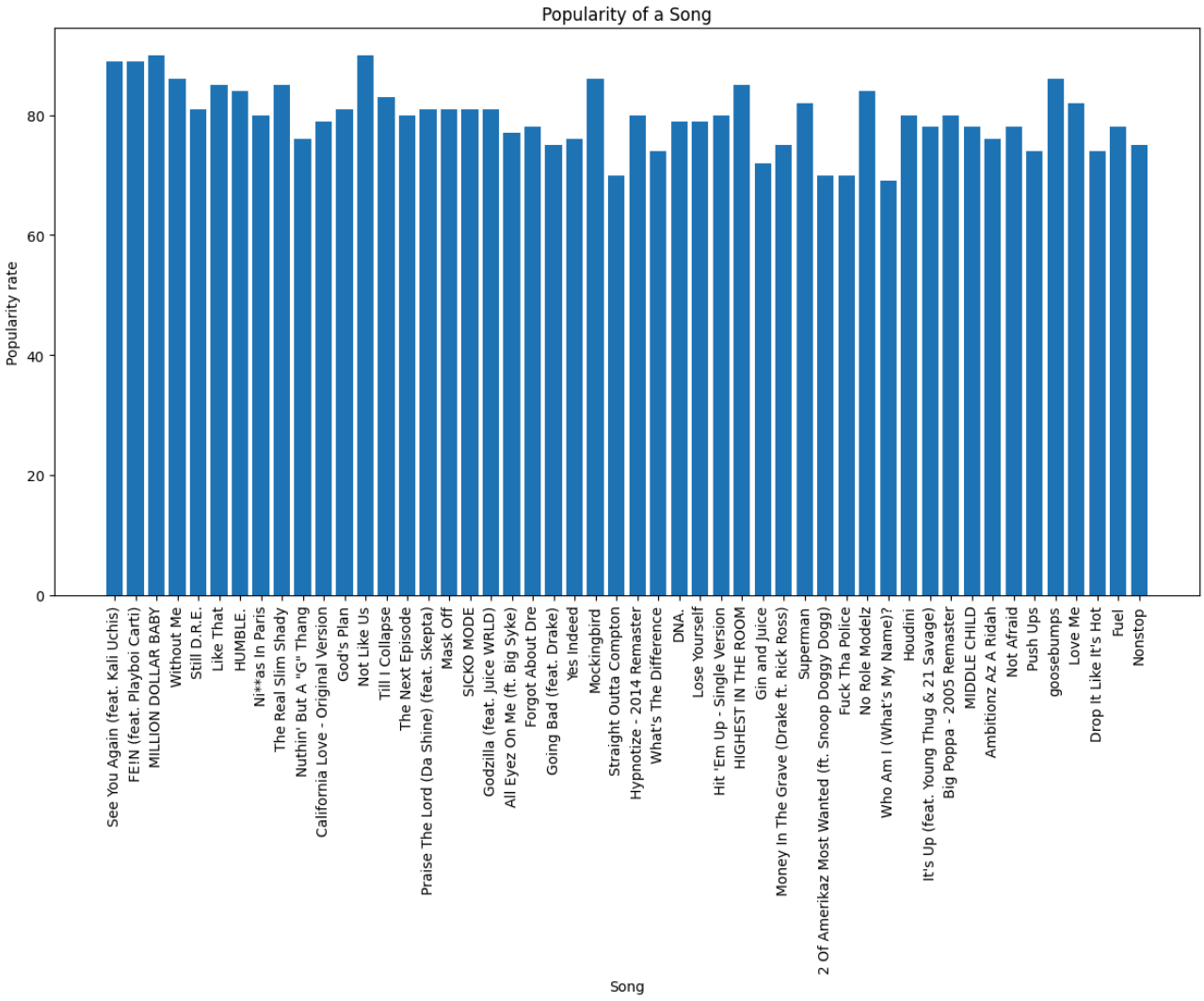
indicating that most tracks are trendy. Energy levels, measured on a scale of 0 to 1, have an average of 0.68, showing that most songs are moderately energetic, with values ranging from 0.35 to 0.95. On a 0 to 1 scale, the liveness scores average 0.19, indicating that most tracks are studio recordings with minimal live audience elements. The data reveals critical song characteristics such as popularity, energy, and recording context.

| | length | popularity | energy | liveness |
|---|---|---|---|---|
| **count** | 50.000000 | 50.000000 | 50.000000 | 50.000000 |
| **mean** | 243953.020000 | 78.800000 | 0.676580 | 0.184892 |
| **std** | 47180.742029 | 5.115004 | 0.157604 | 0.126527 |
| **min** | 142273.000000 | 69.000000 | 0.346000 | 0.045400 |
| **25%** | 211066.500000 | 75.250000 | 0.561500 | 0.096850 |
| **50%** | 244018.000000 | 79.500000 | 0.696500 | 0.145000 |
| **75%** | 277048.000000 | 81.000000 | 0.771750 | 0.261000 |
| **max** | 350320.000000 | 90.000000 | 0.954000 | 0.559000 |

The histogram displays the frequency of songs by various artists in the dataset. Eminem has the highest number of songs, with 10 tracks, followed by artists like Drake, 2Pac, and Kendrick Lamar, each contributing 4 to 6 songs. Artists like JAY-Z, Future, and A$AP Rocky have 2 to 3 songs in the dataset, while Lil Baby, N.W.A., The Notorious B.I.G., and Lil Wayne have fewer than 2 songs each. The chart provides a clear visualization of the distribution of songs by each artist, showing that certain artists, particularly Eminem, dominate the dataset.




**The bar chart displays the popularity scores of various songs, as measured on a scale from 0 to 100. Most songs in the dataset have high popularity scores, with many clustered between 80 and 90. The song "HIGHEST IN THE ROOM" appears to have one of the highest popularity scores, close to 90, while songs like "F**kin' Problems" and "The Watcher" show relatively lower scores but still maintain values above 70. The overall trend shows that the dataset includes songs with consistently high popularity, highlighting their significant streaming success. The variation in scores, while not dramatic, suggests differences in how frequently these songs are streamed or engaged with on the platform.**




## Reddit API (PRAW)[1]:

- **Purpose: To gather user discussions and comments on specific songs used for sentiment analysis.**
- **Data Collected:**
    - **Post Title: The title of Reddit posts mentioning songs.**
    - **Post Body: Content of the post.**
    - **Comments: User comments on posts, providing additional sentiment data.**
    - **Number of Comments: Total number of user comments for each post.**

- Comment Authors: Names of users who commented.

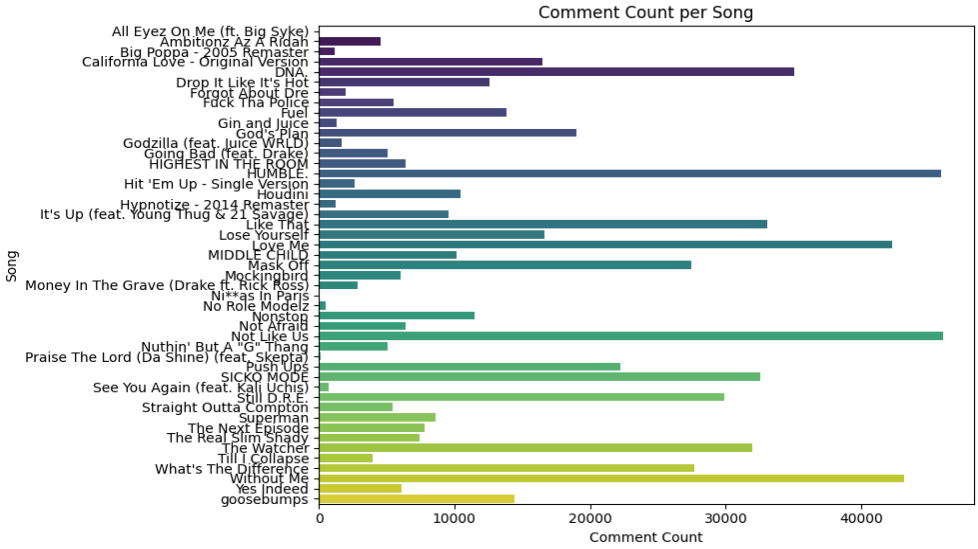
The "Comment Count per Song" horizontal bar chart is displayed in this image. It shows the quantity of comments left on different tracks, perhaps on a social media or music portal.

Song titles are displayed on the y-axis, and the comment count, which ranges from 0 to over 40,000, is displayed on the x-axis. A horizontal bar representing each song is used, and the length of the bar indicates how many comments it has received.

Several noteworthy findings:

- With more than 40,000 comments, the song "HUMBLE." has the most comments overall.
- Other songs that have received a lot of attention are "Love Me," "Not Like Us," and "DNA."
- "Ambitionz Az A Ridah" and "All Eyez On Me (ft. Big Syke)" are two songs that have fewer comments.
- Songs by musicians like Drake, Tupac, and Kendrick Lamar are mixed together on the list, showcasing a variety of genres.
- A rainbow effect is produced by the bars' progressive color transition from purple at the top to yellow at the bottom.

This visualization makes it easy to compare the engagement or popularity of songs based on the number of comments on various tracks.



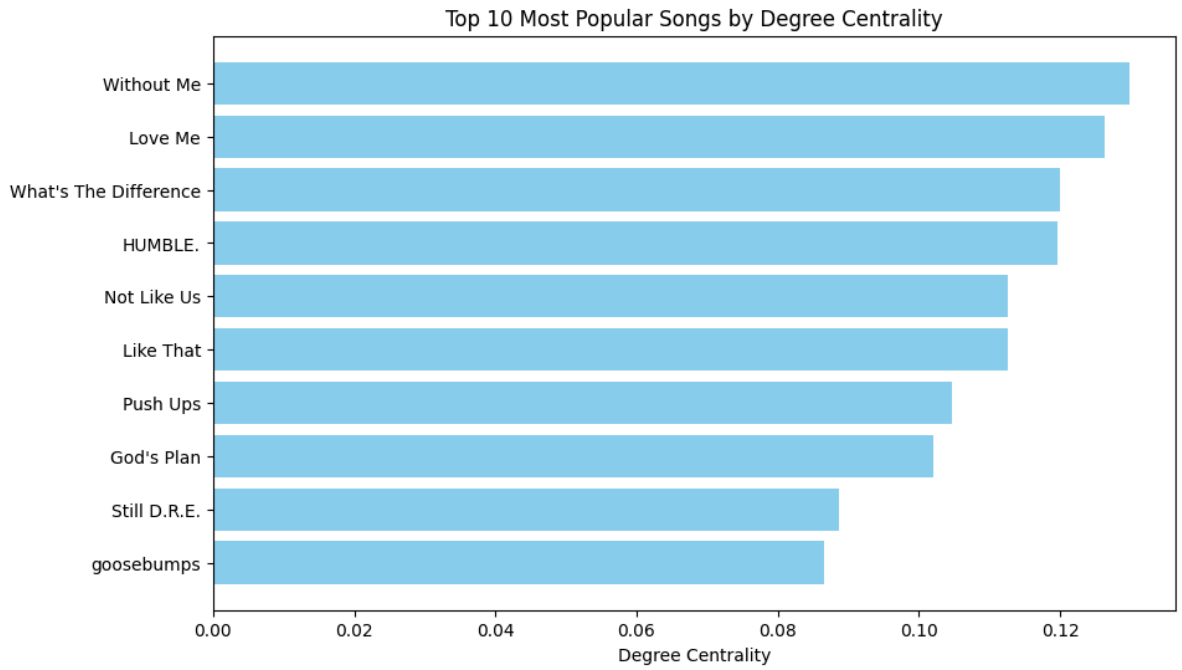
**"Top 10 Most Popular Songs by Degree Centrality" is the title of the horizontal bar chart displayed in this picture. Songs are ranked on the chart according to their degree of centrality, a metric that indicates how important or well-liked they are inside a network.**

**Important aspects of the chart include:**

1. **Song titles are displayed on the y-axis, while the degree of centrality, which ranges from 0 to around 0.12, is represented on the x-axis.**
2. **A light blue horizontal bar with a length that matches each song's degree centrality score is used to represent each song.**
3. **According to their degree of centrality, the songs are ordered descendingly.**

    **The top 10 songs listed are:**

    - **Without Me**
    - **Love Me**
    - **What's The Difference**
    - **HUMBLE.**
    - **Not Like Us**
    - **Like That**
    - **Push Ups**
    - **God's Plan**
    - **Still D.R.E.**
    - **goosebumps**
4. **"Without Me" has the highest degree centrality, slightly above 0.12.**
5. **The differences in degree centrality between songs are relatively small, especially among the top-ranked songs.**
6. **"goosebumps" has the lowest degree centrality among the top 10, at around 0.09.**




## Data Preprocessing

**The data preprocessing involved several vital steps to prepare both the Spotify and Reddit datasets for analysis. Initially, the Spotify data was checked for missing values using df.isnull().sum(), confirming that no missing values were present in variables like name, album, artist, release_date, length, popularity, energy, and liveness.**

```
#data pre-processing: checking for missing data
df.isnull().sum()
```

```
name           0
album          0
artist         0
release_date   0
length         0
popularity     0
energy         0
liveness       0
dtype: int64
```

**The processText function was applied to preprocess text from Reddit posts and comments. This function involved the following steps:**

- **Lowercasing: The function processText converts all text to lowercase using text. lower(). This ensures that words like "Happy" and "happy" are treated equally, avoiding case-based discrepancies.**
- **Tokenization: The TweetTokenizer from the nltk library breaks down the text into individual tokens (words). Tokenization splits sentences into manageable parts, allowing each word to be processed separately. This step is essential for feeding the text into sentiment analysis tools. Example: "I love this song!" becomes ["I," "love," "this," "song," "!"].**
- **Stripping Whitespace: The tokens are cleaned to remove any leading or trailing whitespace using the token after tokenization. strip(). This step ensures that each token is free of unnecessary spaces.**
- **Stemming: The PorterStemmer from nltk is used to perform stemming. Stemming reduces words to their base or root form. For example, "running" becomes "running," and "happiest" becomes "happy." This helps consolidate different variations of the same word. Example: Words like "liked," "likes," and "like" are all stemmed to "like."**
- **Stopword Removal: Common stopwords, such as "the," "is," and "in," are removed to reduce noise in the data. These words are not usually valuable for sentiment analysis because they don't carry significant meaning. The stopwords are combined with punctuation marks and specific irrelevant tokens like "rt" and "via" to create a comprehensive list of items to remove. Example: "I like this song a lot" becomes ["like," "song"] after stopword removal**
- **Regex Filtering for Numbers and URLs: Regular expressions are applied to filter out digits and URLs**
    - **regexDigit: This regular expression removes numbers (e.g., "1234") from the tokens. It's applied using re.compile("^\d+\s|\s\d+\s|\s\d+$") to ensure that standalone digits or numbers embedded in text are excluded.**
    - **regexHttp: This regular expression looks for words that begin with "http" or other similar URL patterns and gets rid of them. This ensures that any web addresses or links in the posts or comments are screened.**

```
def processText(text, tokenizer, stemmer, stopwords):
    """
    Perform tokenisation, normalisation (lower case and stemming) and stopword and keyword removal for reddit

    @param text: reddit submission or comment text
    @param tokenizer: tokeniser used.
    @param stemmer: stemmer used.
    @param stopwords: list of stopwords used

    @returns: a list of processed tokens
    """
    text = text.lower()
    lTokens = tokenizer.tokenize(text)
    lTokens = [token.strip() for token in lTokens]
    lStemmedTokens = set([stemmer.stem(tok) for tok in lTokens])
    regexDigit = re.compile("^\d+\s|\s\d+\s|\s\d+$")
    regexHttp = re.compile("^http")
    return [tok for tok in lStemmedTokens if tok not in stopwords and not tok.isdigit() and regexDigit.match(tok) ==
```

```
def preprocessText(textType):
    '''
    Main function for text preprocessing
    @textType: Text to be used to get tokens
    @lTokens: Token values returned in an array for each post
    '''
    tweetTokeniser = nltk.tokenize.TweetTokenizer()
    lPunct = list(string.punctuation)
    lStopwords = nltk.corpus.stopwords.words('english') + lPunct + ['rt', 'via', '...', '…', '"', "'", '`']
    tweetStemmer = nltk.stem.PorterStemmer()
    lTokens = processText(text=textType,
                          tokenizer=tweetTokeniser,
                          stemmer=tweetStemmer,
                          stopwords=lStopwords)
    return lTokens
```

After the text data was cleaned, it was used for more mood analysis. This made sure that the data was in a state that could be processed and analyzed.

## Analysis

The project used several analysis methods to look into how social media affects music usage, such as sentiment analysis, theme modeling, and graph analysis. Here is an in-depth look at each technique, along with pictures and descriptions of how they worked:

### Sentiment Analysis:

- Technique Used: The nltk library's Vader SentimentIntensityAnalyzer determined how people felt. VADER is made to look at how people think about things in social media posts, and it works well with text from Reddit and other sites. This tool gives four types of scores for text: positive, negative, neutral, and compound (total mood score).
- Process: Reddit posts and comments about each song were analyzed based on how people felt about them. The Reddit posts and comments were cleaned up by changing them to lowercase, getting rid of all the capitalization, and using the Porter Stemmer to break down words into their root forms. After that, the mood of every Reddit post and

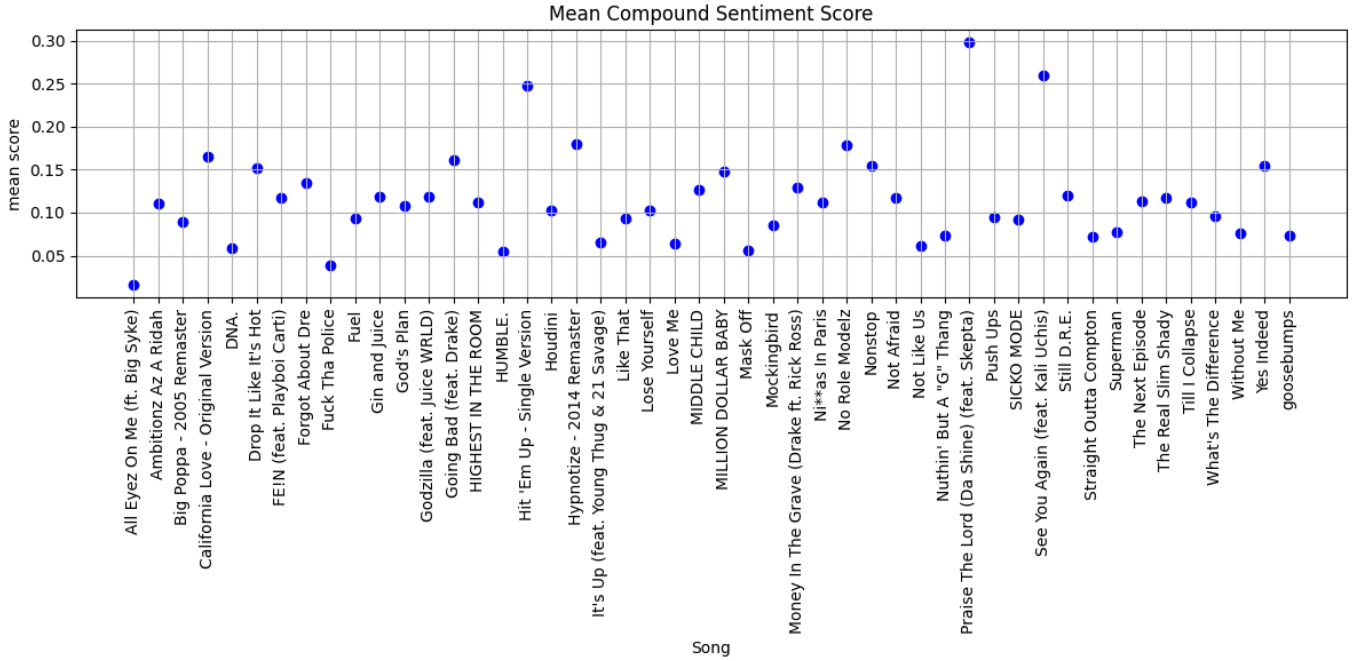
comment was evaluated, and a combined mood number was given to each one. This number goes from -1 (extremely negative sentiment) to +1 (extremely positive sentiment), with values around 0, meaning people generally don't feel either way.

- Analysis: The research showed that the feelings in different songs differed. There were positive and negative feelings in the Reddit conversations about songs like "All Eyez On Me" and "Ambitionz Az A Ridah," which got much attention. While some comments were very positive, which could mean that fans interacted with each other with much energy, others were negative, which could mean that people were having critical or controversial conversations. The average mood numbers gave each song's discussion a general emotional tone. People who talked about famous songs often had very different views, showing fan loyalty and critical responses.


```
{'All Eyez On Me (ft. Big Syke)': 0.3199,
 'Ambitionz Az A Ridah': 0.11075566397963514,
 'Big Poppa - 2005 Remaster': 0.08614840871021776,
 'California Love - Original Version': 0.16419074188882002,
 'DNA.': 0.06089386095521706,
 "Drop It Like It's Hot": 0.15286922552804905,
 'Forgot About Dre': 0.13557425488180883,
 'Fuck Tha Police': 0.04299502989536623,
 'Fuel': 0.09428911947820931,
 'Gin and Juice': 0.12568608169440243,
 "God's Plan": 0.11058714739069112,
 'Godzilla (feat. Juice WRLD)': 0.1309863025210084,
 'Going Bad (feat. Drake)': 0.16108433524569718,
 'HIGHEST IN THE ROOM': 0.1138921766072812,
 'HUMBLE.': 0.05452555415390161,
 "Hit 'Em Up - Single Version": 0.24943647148871032,
 'Houdini': 0.10185154377880183,
 'Hypnotize - 2014 Remaster': 0.1829882304526749,
 "It's Up (feat. Young Thug & 21 Savage)": 0.0671436060539592,
 'Like That': 0.0941757742167807,
 'Lose Yourself': 0.1047689749847468,
 'Love Me': 0.063004500691085,
 'MIDDLE CHILD': 0.12951548692182652,
 'Mask Off': 0.05682485285285285,
 'Mockingbird': 0.08527451481103167,
 'Money In The Grave (Drake ft. Rick Ross)': 0.13346079173838207,
 'Ni**as In Paris': 0.0853670731707317,
 'No Role Modelz': 0.17820631768953069,
 'Nonstop': 0.15312364644487933,
 'Not Afraid': 0.12105532518164099,
 'Not Like Us': 0.06156349773440224,
 'Nuthin\' But A "G" Thang': 0.07886214265889537,
 'Praise The Lord (Da Shine) (feat. Skepta)': 0.291496062992126,
 'Push Ups': 0.09723494734770198,
```


The chart below shows various songs' mean compound sentiment scores, indicating each track's average overall sentiment in Reddit discussions. The compound score is between -1 (most pessimistic) and 1 (most positive), which summarizes the overall emotional tone.

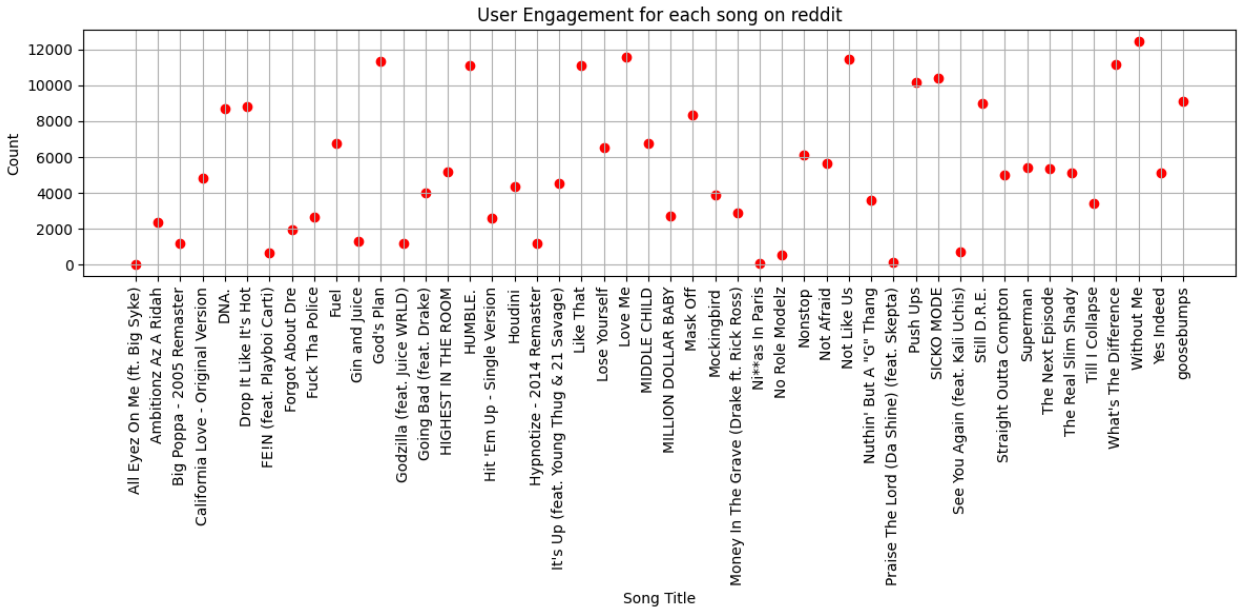
- Positive Sentiment: Songs such as "All Eyez on Me" and "It's Up" have the highest average sentiment scores, around 0.30, suggesting that the discussions around these songs were predominantly positive.
- Neutral/Lower Sentiment: Songs like "Goosebumps" and "What's The Difference" have relatively lower scores, close to 0.05, indicating that discussions about these songs were more neutral or mixed in sentiment.
- Variation in Sentiment: The spread of scores across the songs indicates that some tracks evoke stronger positive sentiments in online discussions, while others generate more neutral or less positive reactions. There are no significant negative sentiment scores, which suggests that most songs were generally well-received.




The graph displays user engagement for each song on Reddit, measured by the number of comments and posts for each song. The x-axis represents the song titles, and the y-axis represents the number of engagements (comments and posts).

- High Engagement: Some songs, such as "Lose Yourself" and "See You Again," show high engagement, with over 10,000 interactions. This indicates that these songs are popular topics of discussion on Reddit.
- Moderate Engagement: Songs like "Hypnotize" and "Mockingbird" have moderate engagement, with around 4,000–6,000 comments and posts. These songs still have a strong presence in Reddit discussions but are not as frequently discussed as the top songs.

- Low Engagement: Some tracks like "Goosebumps" and "Money In The Grave" show significantly lower engagement, with fewer than 2,000 interactions. This indicates that these songs generated less discussion or interest on Reddit.




## Topic Modeling:

Topic modeling is a method to uncover hidden topics or themes within a collection of documents (in this case, Reddit posts). In this project, the Latent Dirichlet Allocation (LDA) algorithm is applied to Reddit discussions to identify topics related to songs or artists. LDA is a generative statistical model that assumes each document (post) is a mixture of a small number of issues and each topic is a mixture of words.

Critical Components of the Code:

1. NumOfTopics: The number of topics to discover in the Reddit posts is set to 20. Each topic will consist of a group of words that frequently appear together across the posts.
2. WordNumToDisplay: This specifies how many words per topic will be displayed to understand the core idea behind each topic. In this case, it's set to 5 words per topic.
3. FeatureNum: The number of features/words used to describe the documents. It's capped at 2000, meaning the top 2000 frequent words are considered for the topic model.

Data Preparation: The function retrieveTopics reads the data from the songs_data.json file and processes each post title through the preprocessText function, which removes stopwords, tokenizes, and stems the text. These preprocessed tokens are then joined into a string to form the final dataset that will be used for topic modeling.

**Topic Modeling with LDA:**

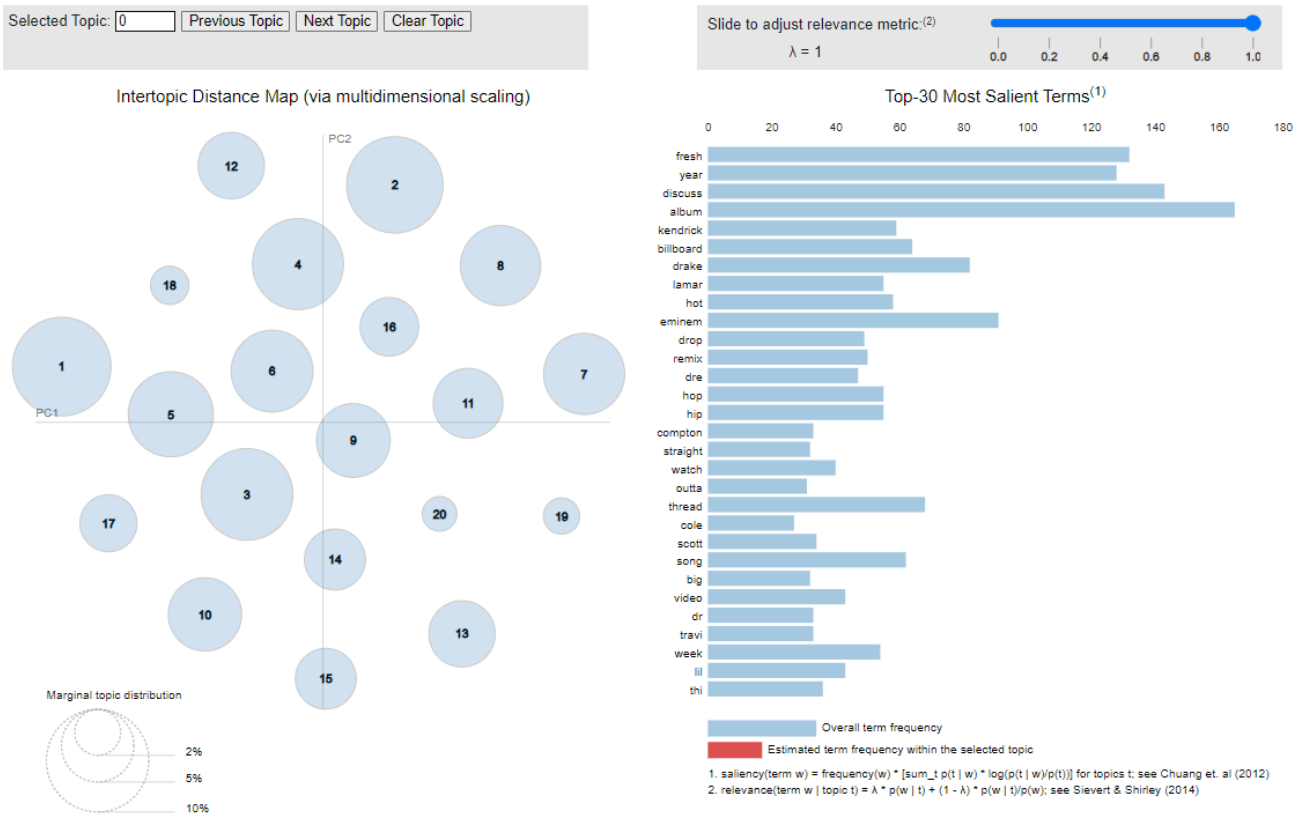
- **A CountVectorizer converts the text data into a term-frequency matrix (tf). This matrix is a representation of how frequently words appear in each post.**
- **The LDA model is created using the LatentDirichletAllocation class from sklearn. The model is trained on the term-frequency matrix (tf) with 20 topics (n_components=NumOfTopics). The model then learns a distribution of words that defines each topic.**

**Displaying the Topics:**

- **The function show_topics shows the most important words that describe each subject. Examples: Topic 0 has words like "thread," "daili," and "compton," which suggests that it might be about individual artist conversations, release threads, or playlists.**
- **Example Topics:**
    - **Topic 1: It has words like "album," "year," "week," and "fresh," which makes me think it might be about new albums coming out every week or so.**
    - **Topic 3: It has words like "hip," "hop," "throwback," and "denzel," which means it's mostly about old-school hip-hop acts and conversations.**
    - **Topic 17: It includes words like "drake," "lil," and "song," which make it sound like the subject is famous acts like Drake and talking about their songs.**

```
Topic 0:
thread daili discuss straight compton
Topic 1:
album year week fresh best
Topic 2:
cole fuck live tha polic
Topic 3:
hip hop throwback curri denzel
Topic 4:
fresh feat memori thread tyler
Topic 5:
discuss billboard hot 12 june
Topic 6:
az 13 ridah ambitionz pusha
Topic 7:
drop dre watch dr ft
Topic 8:
love know man featur method
Topic 9:
shadi album slim stori death
Topic 10:
kany west freestyl fresh drake
Topic 11:
kendrick lamar video fresh big
Topic 12:
nonstop mike perform wu tang
Topic 13:
rapper fuel life jay chanc
Topic 14:
tracklist lupe fiasco trust nba
Topic 15:
gener sunday parti rick ross
Topic 16:
eminem discuss year later thi
Topic 17:
drake lil song stream ye
Topic 18:
scott travi origin version leak
Topic 19:
juic gin wrld beat come
```

**Visualization with pyLDAvis: The topics are shown with pyLDAvis. It makes a dynamic panel for looking into the topics and how they relate. The topics are displayed in a two-dimensional space using a technique like t-SNE (t-distributed Stochastic Neighbor Embedding), where similar topics appear closer together. This visualization helps to understand how distinct or overlapping the topics are and provides insights into how frequently each topic appears across Reddit posts.**

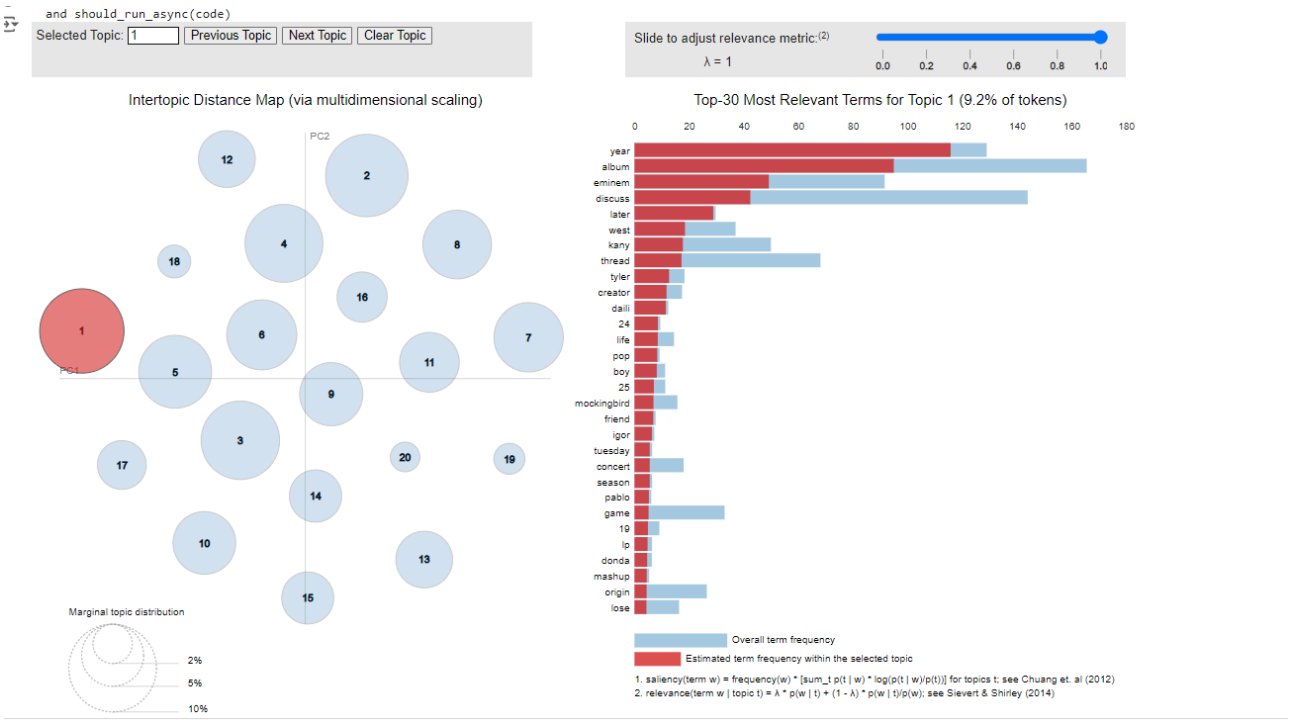



**For example, Topic 1 and Topic 2 are relatively close, indicating they might share some terms or themes in the discussion**

**Top 30 Most Relevant Terms per Topic**

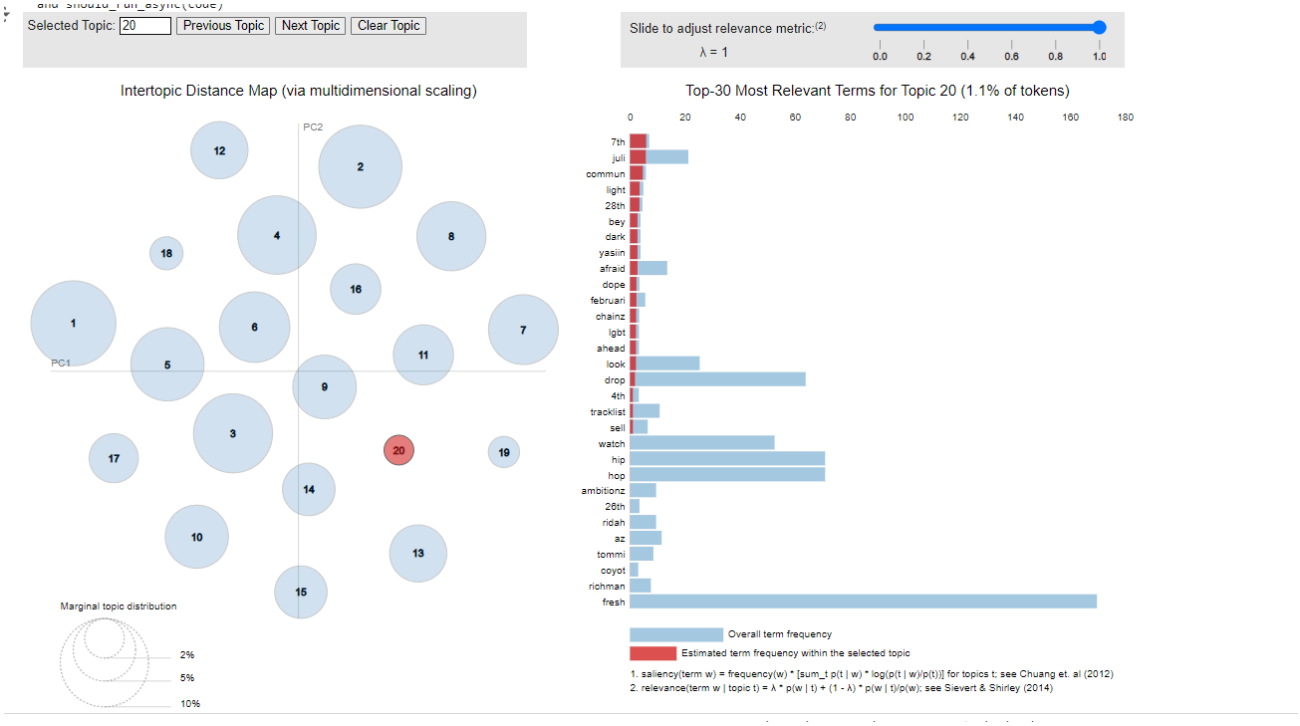
**Example 1 (Topic 1):**

- **Key terms: "year," "album," "discuss," "kanye," and "thread."**
- **Interpretation: This topic focuses on yearly album releases and general discussions about artists like Kanye West. The presence of terms like "thread" indicates that these are structured Reddit discussions, possibly weekly or yearly review threads about music.**



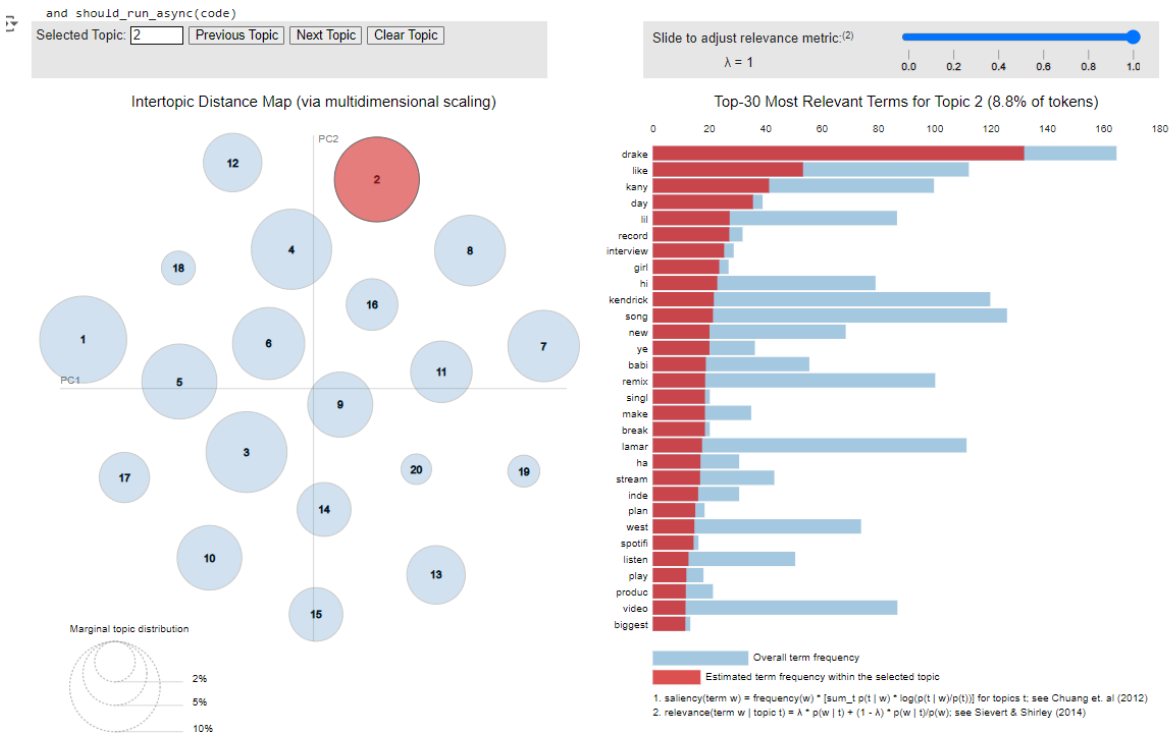

**Example 2 (Topic 20):**

- **Key terms: "hip," "hop," "watch," "ambitionz," "sell."**
- **Interpretation: This topic likely revolves around the hip-hop genre and specific discussions about artists and songs related to sales and the cultural impact of certain tracks (e.g., "Ambitionz Az a Ridah").**



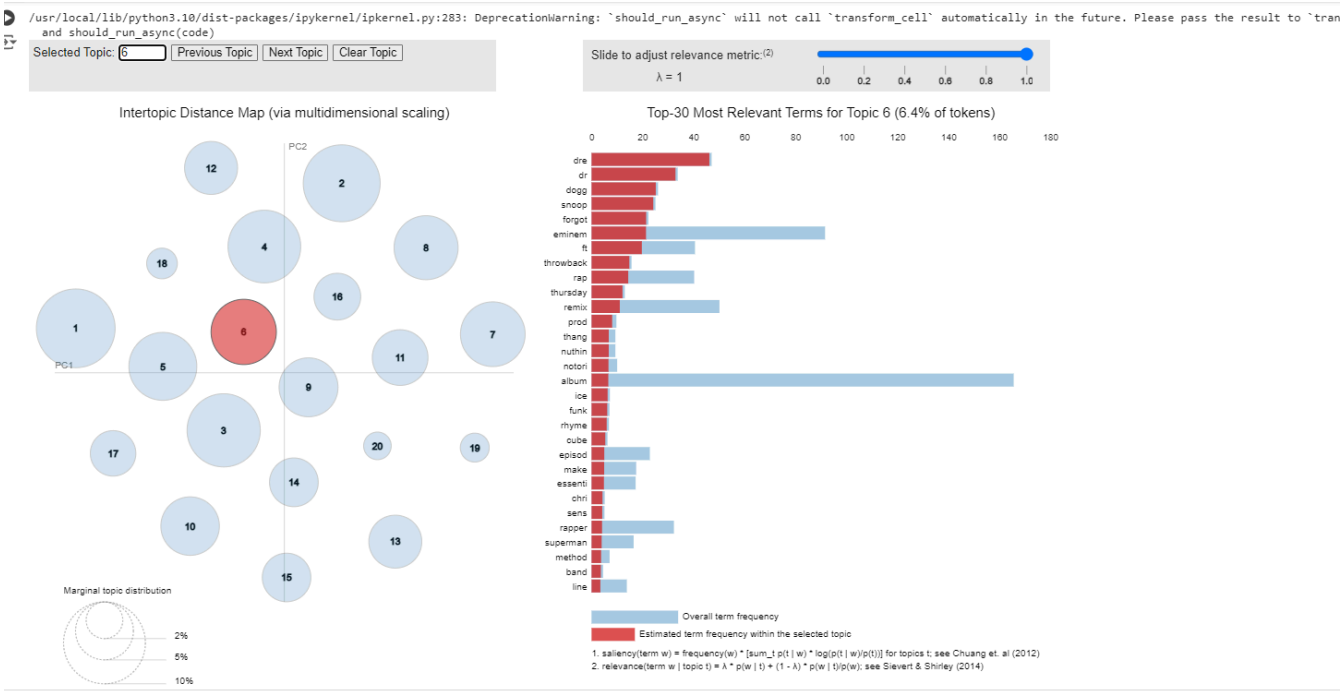

**Example 3 (Topic 2):**

- **Key terms: "drake," "kanye," "kendrick," "remix," and "album."**
- **Interpretation: This topic focuses on popular artists like Drake, Kanye West, and Kendrick Lamar, often mentioned in the context of albums, remixes, and new releases.**




**Example 4 (Topic 6):**

- **Key terms: "dre," "snoop," "dogg," "eminem," and "rap."**
- **Interpretation: This topic likely revolves around discussions related to Dr. Dre, Snoop Dogg, Eminem, and the broader rap genre. It includes words like "throwback," indicating that discussions may be centered around classic or older rap songs.**




The figure below contains word clouds representing the top terms for each of the 20 topics identified through topic modeling. Each word cloud visualizes the most frequent and relevant words within a specific topic, where larger words appear more often or are more strongly associated with that topic. Here's a breakdown of some key insights:

Key Observations from Selected Topics:

- Topic 1: Words like "memori," "thread," "offici," and "album" suggest that this topic is related to discussions about music memories, official releases, and album threads, potentially focusing on reviews or music nostalgia.
- Topic 2: The words "eminem," "perform," and "live" suggest that this subject may be about live performances and Eminem, especially his live shows or performances.
- Topic 6: Some of the words in this topic are "new," "listen," "thing," and "music," which makes me think of general conversations about new music releases or future tracks. They could also be conversations about music suggestions or events that are coming up.
- Topic 9: With words like "slim," "shadi," "album," and "rap," it's clear that this topic is mostly about talking about Eminem, especially his rap records and his alter ego "Slim Shady."
- Topic 11: The words "jul," "light," and "commun" suggest that this subject may be about music groups, possibly focussing on events or new releases that happen at certain times, like summer hits or music festivals.

- Topic 17: This topic seems to be mostly about talking about artists like Drake and Kanye West and how people feel about their songs or albums. Words like "drake," "like," and "kany" make this clear.
- Topic 19: With words like "dogg," "dre," and "snoop," this topic is probably about how rap and hip-hop artists like Snoop Dogg and Dr. Dre have changed the genre.

Word clouds show a quick summary of the most common words related to a certain subject. The frequency of larger words in the text shows how important they are to the topic. This visulazation makes it easy to quickly find the main points of discussion for each subject. Like this:

- Topic 1 seems to focus on nostalgic or reflective discussions.
- Topic 9 clearly focuses on Eminem and his alter ego.
- Topic 19 highlights prominent figures like Snoop Dogg and Dr. Dre in rap conversations.

Topic modeling identifies groups of words that tend to appear together in discussions, providing insights into the major themes discussed on Reddit about specific songs or artists. The LDA model found 20 different themes. Some were about specific artists (like Drake, Eminem, or Travis Scott), while others were about things like album releases, song threads, or conversations about a certain genre.

### Community Detection and Analysis[3]:

In network analysis, the Louvain method is a commonly used technique for community detection, which finds clusters of strongly connected nodes within a larger network. An outline of the Louvain community detection method is provided below:

The Louvain method is an algorithm designed to optimize modularity, which measures the strength of division of a network into communities. It works through these key steps:

Initialization:

- Each node in the network starts as its own community.

Two-phase iterative process:

a) Local optimization:

- For each node, the algorithm considers moving it to the community of one of its neighbors.
- It calculates the change in modularity for each potential move.
- The node is moved to the community that results in the highest increase in modularity (if positive).
- This process is repeated for all nodes until no further improvement is possible.

b) Network aggregation:

- Communities found in the first phase are aggregated into "super-nodes".
- A new network is created where these super-nodes are connected by weighted edges.

Iteration:

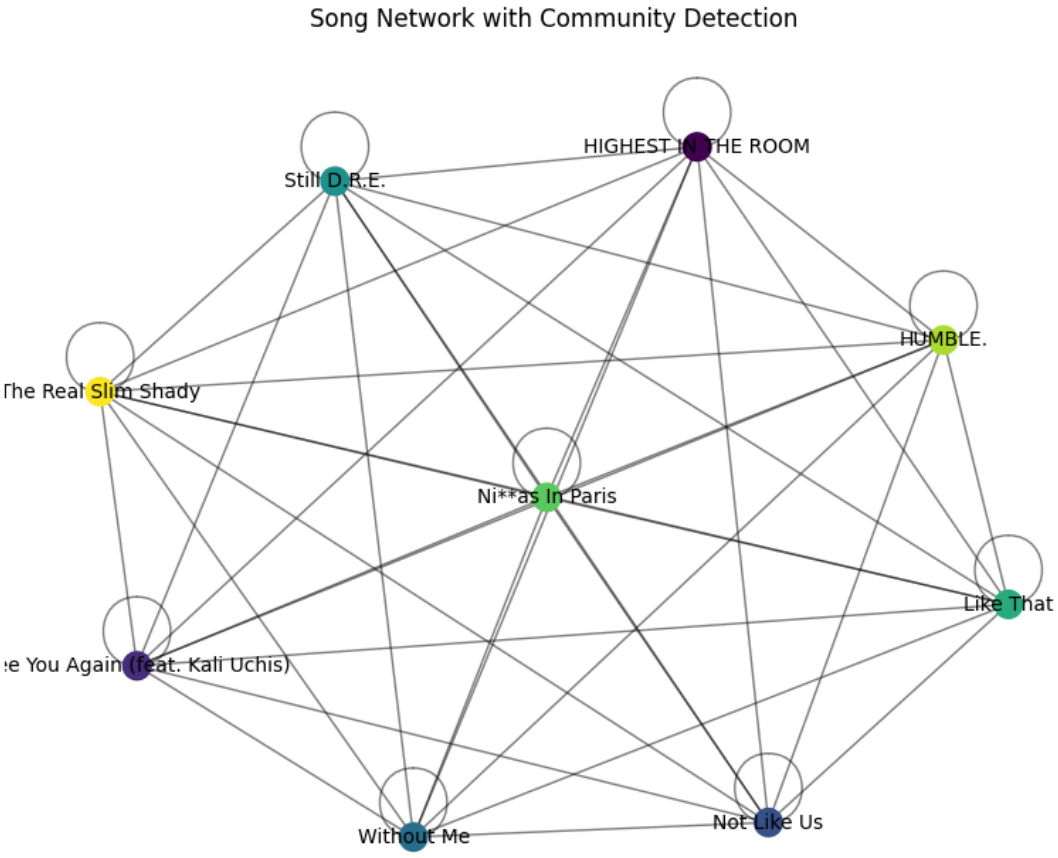
- Steps 2a and 2b are repeated on the aggregated network until no further improvement in modularity is possible.

Key features of the Louvain method:

- Hierarchical: It can reveal community structures at different scales.

- **Fast and efficient: Suitable for large networks.**
- **Resolution limit: Like many modularity-based methods, it may fail to detect small communities in large networks.**

**The Louvain technique, which is useful for evaluating complex systems in a variety of domains, including social networks, biological networks, and, in this case, networks of musical interactions, is well-known for its quickness and efficacy in identifying communities in huge networks.**




**This image shows a network graph titled "Song Network with Community Detection". The graph represents relationships between different songs, with each song represented by a node.**

**The community detection aspect is illustrated through different colored nodes, suggesting groups of songs that are more closely related to each other.**

**Detailed analysis of the graph:**

**Color-coded communities:**

- **The graph shows at least 5 distinct communities, each represented by a different color (purple, green, yellow, light blue, and dark blue).**

**Community sizes:**

- **Most communities appear to have 1-2 songs each, indicating a high level of distinctiveness between songs.**

**Largest community:**

- **The green community seems to be the largest, containing 3 nodes: "Still D.R.E.", "Ni**as in Paris", and "Like That".**

**Isolated communities:**

- **Some songs form their own single-node communities, such as "The Real Slim Shady" (yellow) and "HUMBLE." (light green).**

**Central nodes:**

- **"HIGHEST IN THE ROOM" (purple) appears to be a central node, connecting to many other nodes across different communities.**

**Cross-community connections:**

- **Despite the community structure, there are many connections between nodes of different colors, suggesting complex relationships between songs across different groupings.**

**Potential factors for community formation:**

- **The communities might represent similarities in genre, artist, era, or musical style, though this cannot be definitively determined from the graph alone.**

**Network density:**

- **The network appears to be fairly dense, with many connections between nodes, indicating complex relationships between songs.**

**Implications:**

- **This community structure could be useful for recommendation systems, playlist creation, or understanding musical trends and influences across different songs and potentially artists.**

## Conclusion

**In conclusion, this study highlights the significant role social media, particularly Reddit, plays in driving music consumption in the streaming era. Our analysis reveals that songs generating high positive engagement on Reddit often correlate with increased popularity on Spotify, illustrating how social sentiments directly influence streaming success.**

**By utilizing sentiment analysis, we were able to discern distinct trends in music conversations, which demonstrated that people regularly interact with particular musicians, musical genres, and significant music-related occasions like album launches and live performances. Prominent individuals such as Eminem, Drake, Snoop Dogg, and Kanye West surface as primary subjects, highlighting their cultural relevance and the manner in which they influence user dialogues.**

**These findings are further supported by the community detection study, which reveals unique clusters representing conversations around specific artists and genres. This arrangement highlights the interdependence of diverse musical inspirations and the group dynamic of fandoms, illuminating the intricate relationships within the hip-hop and rap music scenes. These networks' central nodes identify songs and performers with significant cultural influence, and the general design encourages the formation of discussion topics centered on hot new releases and genre-defining hits.**

**In the end, our research highlights how crucial social media platforms are in influencing how the general public views and interacts with music. Engaging in these online discussions on a regular basis would probably help artists become more visible and successful on streaming services. Social media, then, is a crucial part of the modern music industry and has a big impact on how listeners find, debate, and enjoy music.**

## Sources And References:

[ 1]. *The reddit instance#* (no date) *The Reddit Instance - PRAW 7.7.2.dev0 documentation*. Available at: https://praw.readthedocs.io/en/latest/code_overview/reddit_instance.html (Accessed: 20 October 2024).

[2].*Web api* (no date) *Web API | Spotify for Developers*. Available at: https://developer.spotify.com/documentation/web-api (Accessed: 20 October 2024).

[3].*Community API¶* (no date) *community API - Community detection for NetworkX 2 documentation*. Available at: https://python-louvain.readthedocs.io/en/latest/api.html (Accessed: 20 October 2024).