CAPSTONE TEAM 4

Alvin Kuo & Sundar Murugesen

12/11/2023

I. INTRODUCTION, PROJECT STATEMENT & ARTICLE REVIEW

1. Introduction and Project Statement

The trend of collaboration in the music industry caught our eyes from 10%-20% around 3 decades ago almost tripling to approaching 30%+ of the top music charts nowadays. We randomly take the 2023/7/28 Billboard Top 100, there are 40 songs out of 100 songs are collaborations. The 40% landmark signals loudly the new trend in the music industry. It motivates us to create a “Collab Index/Score” dashboard concept to analyze how the artist is popular for collab success and how the audio features could be successful based on this collab in our project Part 1. We generate the recommendation for collab success in audio features details and more in Part 2 with LLM(Large Language Mode) leverage including LLM managing platform Langchain, GPT from OpenAI, and Mistral from Mistral AI (through HuggingFace)to demonstrate the application of the collaboration ideas. Eventually, we turn the data script into a shareable web app with Streamlit.

Overall we leverage the data and API services from global #1 music chart provider Billboard and worldwide #1 music streaming service provider Spotify. The application would be for the users to apply our collab index success with more customized feedback to brainstorm their music collab potential.

2. Article Review

1) Billboard 200: The Lessons of Musical Success in the U.S. (Gourévitch, Boris, 2023) - This article provides a solid suggestion to look up the BB200 database and the music history of the rise and fall nature of popular music. The best benefit is 1) We may land this dataset to start instead of a single, and 2) We may benefit from an album viewpoint to study our focus "music collaboration" to compare the songs to collab or not to collab from a better angle. 3) We may from a lot of brilliant visuals and results think through the question we ask and whether

the analysis, simulation, and prediction we may conduct would be valuable or not! A truly awesome read!

2) Collaborative Song Dataset (CoSoD): An annotated dataset of multi-artist collaborations in popular music(Duguay, M., Mancey, K., Devaney, J., 2023)- This article provides a very in-depth analysis of music collaboration from a very music-oriented song analysis for the success of music collaboration. 331 songs in 10 years of BB100 could be probably not that large dataset. However, the analysis goes very deeply into gender, types of music collaboration even how they collaborate in intro, verse, chorus, bridge, to outro. It totally opens another new territory and possibility for our music collaboration study.

3) Detecting Collaboration Profiles in Success-based Music Genre Networks(Oliveira, P., Silva, M., Seufitelli, D., Lacerda, A., & Moro, M.,2020) - This article based on 1958-2020 BB100 data to start from genre to detect and analyze the network. We love the finding that rap, hip-hop, and R&B were listed as top genres with more music collaboration while rock ranks much lower. It may provide us a hint that genre will play a key role in the success, simulation, and prediction of the music collaboration. The network analysis also offers us a brand new window to think about if the network framework could be part of our visuals and models.

II. DATASET, API, Tookit & LLM

1. Dataset

The 5 datasets to help model training are the BB200(Billboard 200) song Spotify audio features and the BB100(Billboard 100) Chart dataset.

1) acoustic.csv - BB200 data with acoustic features (from source)
2) album.csv - BB200 data between 1963-01-05 and 2019-01-19 (from source)
3) charts.csv - BB100 data 1958-08-04 and 2021-11-13 (from source)
4) album_231201.csv - BB200 data up to date 2023-12-02 (tailored-made by our project)
5) charts_231201.csv - BB200 data up to date 2023-12-02 (tailored-made by our project)

1). Acoustic.csv

Around 340K (339,850) rows containing acoustic data for tracks from Billboard 200 albums from 1963 to 1/19/2019. Each row contains 19 columns including 1) id - track ID on Spotify, 2) song - track name, 3) album - album name, 4) artist - artist name, 5) 13 values for Spotify EchoNest acoustic data a) acousticsness, b) danceability, c) duration_ms, d) energy, e) instrumentalness, f) key, g) liveness,h) loudness, i) speechiness, j) mode, k) speechiness, l) tempo, m) time signature, and n) valence, 18) album_id - album ID on Spotify, and 19) date - release date of the album. This table is titled “acoustic.csv”.

2). album.csv

574K around(573,947) rows containing all albums in the Billboard 200 from 1/5/1963 to 1/19/2019 - say 56 years. (However, it was the top 150 albums starting from 1963, then around the top 200 albums from 1967 to 2019- we have 2,705 data, which is around 52 years) Each row contains the 7 columns: 1) id - album_id in Spotify, 2) date - chart date, 3) artist - the artist name of this album, 4) album - album name, 5) rank- album’s place in the charts, 6) length - length of the album which means how many songs (in average it’s around 11 songs), 6) track_length - length of the track in milliseconds (60K milliseconds = 1 min, in average it’s around 3-4 minutes). This dataset is titled “album.csv”.

3). charts.csv

Around 330K (330,087) rows containing chart data for tracks from Billboard 100 singles from 8/4/1958 to 11/16/2021. It’s a total of 63 years. Each row contains 8 columns including 1) date - date of the chart on BB100, 2) rank - the place of the single on the chart, 3) song - song name, 4) artist - artist name, 5) last-week - the place of last week, 6) peak-rank - the latest peak rank for the specific song 7) weeks-on-board - the accumulated week on the chart date, there are total 2.6M data points. This table is titled “chart.csv”.

4). Album_231201.csv

To make it more efficient, we utilize the Spotify API to mitigate the data gap by 12/02/2023. The new CSV will be 624K around(624,746) rows containing all albums in the Billboard 200 from 1/5/1963 to 12/02/2023 - say 60 years. (However, it was the top 150 albums starting from 1963, then around the top 200 albums from 1967 to 2023 around 56 years) Each row contains the 7 columns: 1)date - chart date, 2) rank- album’s place in the charts 3) artist - the artist name of this album, 4) album - album name.

There are a total of 10K(10,169) around unique artists. There are a total of 35K(35,119) around unique albums. This dataset is titled “album_231202.csv”.

5). charts_231201.csv

To make it more efficient, we utilize the Billboard API to mitigate the data gap by 12/02/2023. The new CSV will be around 341K (340,887) rows containing chart data for tracks from Billboard 100 singles from 8/4/1958 to 12/02/2023. It’s a total of 64.5 years. Each row contains 8 columns including 1) date - date of the chart on BB100, 2) rank - the place of the single on the chart, 3) song - song name, and 4) artist - artist name. There are a total of 26K(25,704)songs. There are a total of 11K(10,720)artist combinations(including solo & collaboration). There are a total of 2.6M data points. This table is titled “chart_231201.csv”.

2. ToolKit

The reason why we chose LangChain is because LangChain wins over OpenAI API for free usage while OpenAI API requires paid services. LangChain wins over Hugging Face Hub due to LangChain could offer flexible management operations.

Figure 1: ToolKit Comparison: LangChain, Hugging Face Hub, and OpenAI API

ToolKit	LangChain	Hugging Face Hub	Open AI API
In-use of our project	Yes	No	No
Price	Free	Free	Pay as you go
Management	Flexible management process	Limted functions in management process	Flexible management process

3. API Services

The 4 APIs we utilized are Spotify API, Billboard API, and ChatGPT 3.5. And HuggingFace. Spotify API, Billboard API, and HuggingFace are free. However, the Spotify API requires your own client_ID and client_secret applied to run the API. Both Spotify API and Billboard API have specific frequency and request limits due to the nature of the free offering. We did need to pay for ChatGPT 3.5 with the API key.

4. LLM

The 2 LLMs we used are ChatGPT 3.5 through OpenAI and Mistral-7B-v0.1 through HuggingFace. The brief comparison is as follows:

Figure 2: LLM Comparison: ChatGPT 3.5 and Mistral-7B-v0.1

Feature	ChatGPT 3.5	Mistral-7B-v0.1(Hugging Face)
Type	LLM from API	Local LLM
Access	Remote	Local
Performance	Faster	Slower
Cost	Pay as you go	Free
Model Size	Large (~175B parameters	Small (~7B parameters)

III. METHODOLOGY

Part 1 is finalized by similarity score to compare the two artists' best-estimated audio features with the radar chart to visualize the various dimensions like danceability, loudness, valence, and so on.

Firstly, we started with the dashboard/index concept with data manipulation skills to generate the SCI(Simple Collab Index), ACI(Advanced Collab Index, and FCI(Full Collab Index). Then we dive into another dataset to visualize along with all audio features in AFI(Audio Features Index), AAFI(Artist Audio Features Index), and eventually to integrate them as the CCAFI(Current Collab & Audio Features Index). With the breakdown of two different distributions including PDAFI(Power-Law Distribution Audio Features Index) and NAFI(Normal-Distribution Audio Features Index), we eventually reach our ultimate output of the RAFI(Radar Audio Features Index).

Part II is focused on the application of Streamlit, which is well-known for its well-presented capability for results of data science/analytics. We demonstrate the fundamental capability of Streamlit to output our music collaboration data analytics result. The methodology is Streamit introduction to enhance the user experience like the value of the range in button to make it very intuitive to use.

Part III is centered on leveraging LangChain to deal with multiple LLMs to generate the recommendation. LangChain is the library and methodology developed primarily to simplify and streamline the process of building applications that leverage the capabilities of large language models like our choice of Mistral and OpenAI. Our picks of multiple LLMs rather than one will make the final cut of the faster one to highlight the capability of LangChain. Meanwhile, the different source of the LLMs is also part of the comparison since ChatGPT is directly from LLM, but the Mistral is indirectly from HuggingFace. LangChain can manage them and offer our comparison with multiple

LLM operations. With our data input, the LLM becomes smarter to output with more domain-knowledge-oriented content for users - a process from Generative AI to Narrow AI.

Part IV is the total integration of the combination of Part II and Part III - how to apply the results managed by LangChain with LLMs to be showcased in Streamlit with our Part I collab similarity analysis results. Streamlit could facilitate it for a better application presentation at the end. The methodology of this part is to confirm the vertical integration all the way from the data source all the way to LLM output seamlessly.

IV. ANALYSIS AND FINDINGS

Part I

Firstly, we started with the dashboard/index concept to monitor the

1. **SCI(Simple Collab Index)** - Simple Collaboration Index of Single Artist (or Multiple Artists) across the time. It's specifically useful to monitor an artist with the popularity across their periods. In this index, we equally assume the leading and the featuring roles have the same impact to show in one chart. In comparison to Ed Sheeran and Imagine Dragons, we can see Imagine Dragons has a longer pause for their popularity through this simple collab index which dropped to zero for almost two years. Imagine Dragons is a band, which naturally they do a very limited collaboration with other artists.

Figure 3: SCI(Simple Collab Index) - Ed Sheeran

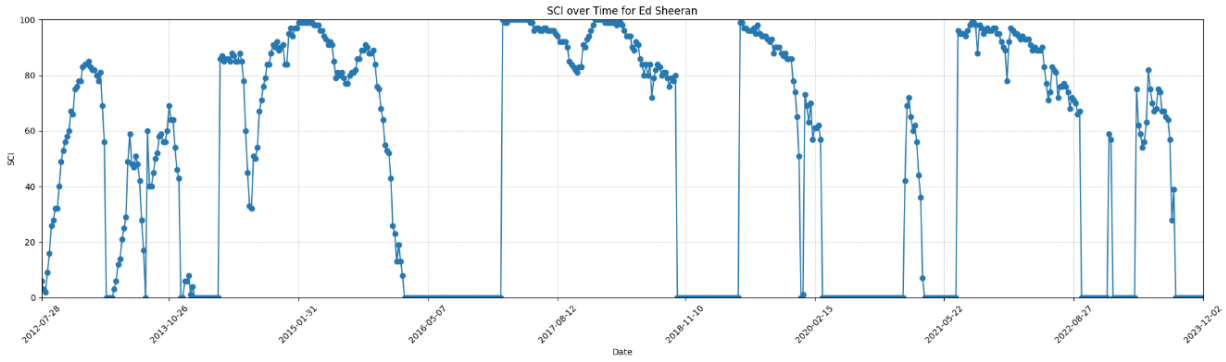
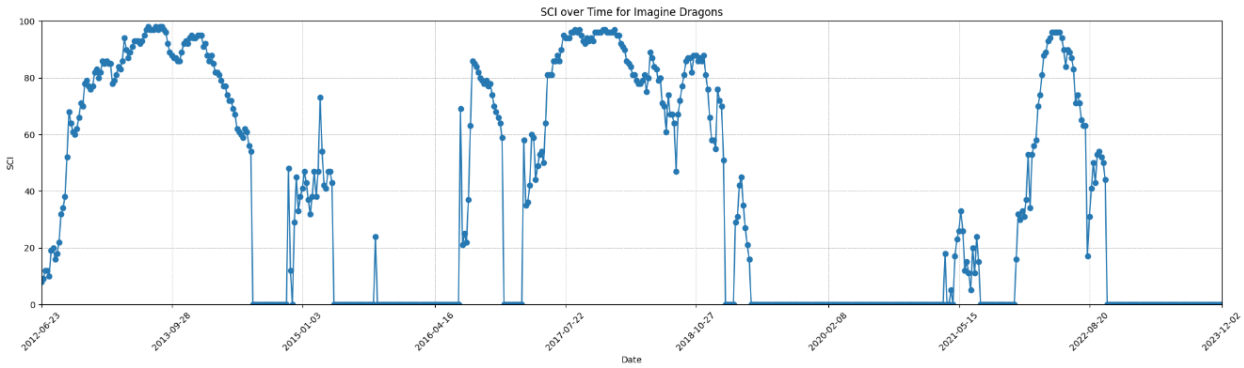


Figure 4: SCI(Simple Collab Index) - Imagine Dragons (From notebook)



2. **ACI(Advanced Collab Index)** - Advanced Collab Index which includes SCI and Lead Artist SCI across the period of the artist(s). It's designed to monitor an artist with the popularity across their periods with the breakdown of Lead Artist SCI. In this index, we equally assume the leading and the featuring roles have the same impact to show in one chart.

   From three different charts from artists of Ariana Grande, BTS, and Enimem. We could observe Ariana Grande started mostly from her solo works to gain popularity. However, in recent years after 2020, she started to kick off a collaboration work strategy to gain more and more popularity with Justin Bieber in "Stuck With You"(2020), with Weeknd in "Save Your Tears" (2020) and "Die For You"(2023). BTS started to work with Coldplay for "My Universe"(2021). Enimum is famous for working with multiple artists throughout his career so far and the most famous one is the collaboration with Rihana in "Love The Way You Lie" (2010).

Figure 5:ACI(Advanced Collab Index)- Ariana Grande (From notebook)

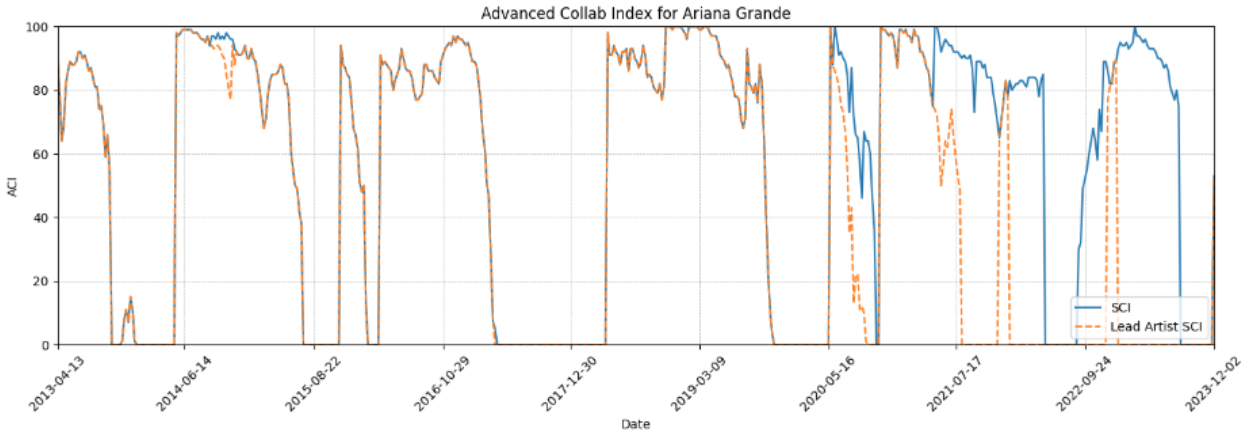


Figure 6: ACI(Advanced Collab Index) - BTS (From notebook)

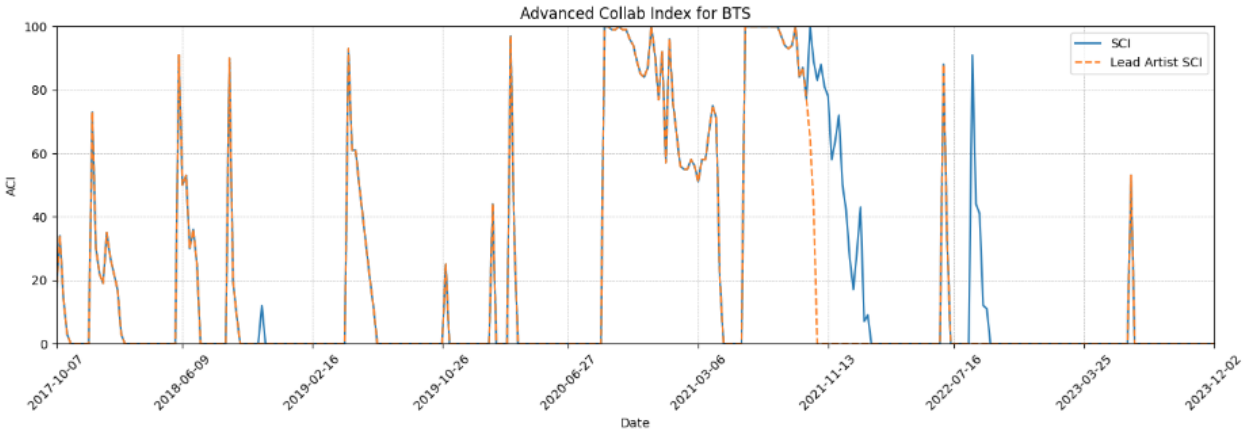
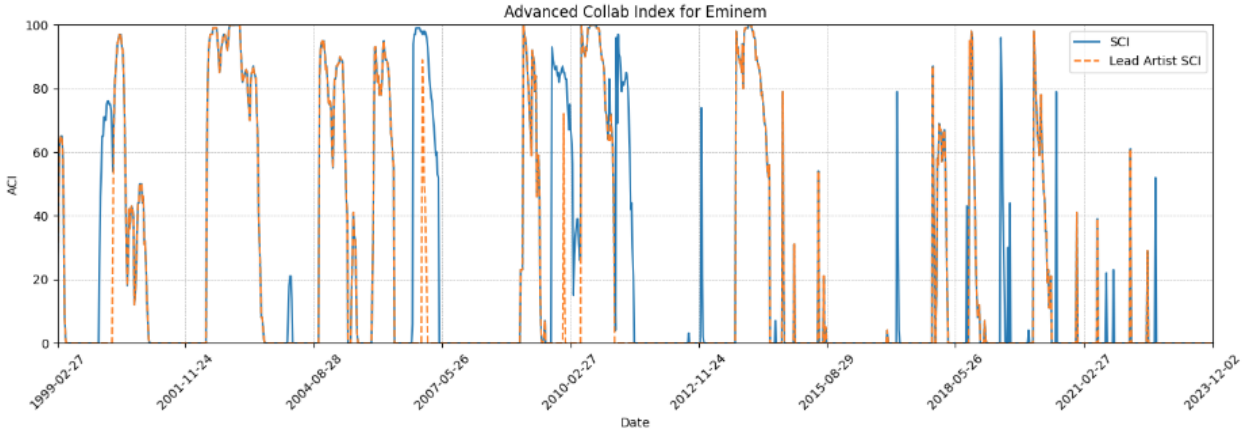


Figure 7: ACI(Advanced Collab Index) - Eminem (From notebook)



3. **FCI(Full Collab Index)** - Here we could also add the Album dataset information to provide an even full overview of the artist's popularity. We could see their album could push some artists to higher popularity like Taylor Swift. Her album is always topping a little bit higher than her solo/collab works in singles. Justin Biber did not leverage that much. Drake's performance of the Album included in the Full Collab Index is between Taylor Swift and Justin Bieber.

Figure 8: FCI(Full Collab Index) - Justin Bieber

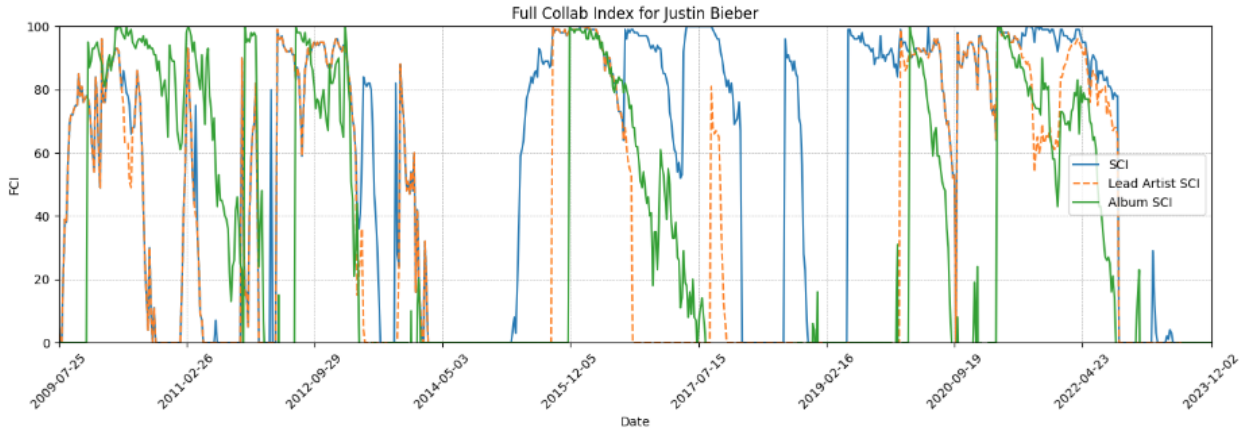


Figure 9: FCI(Full Collab Index)- Taylor Swift (From notebook)

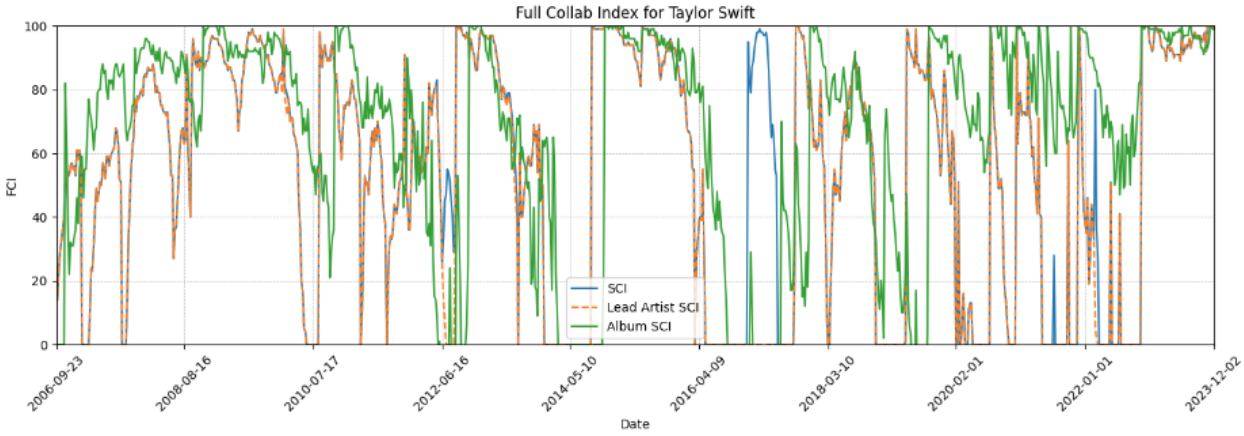
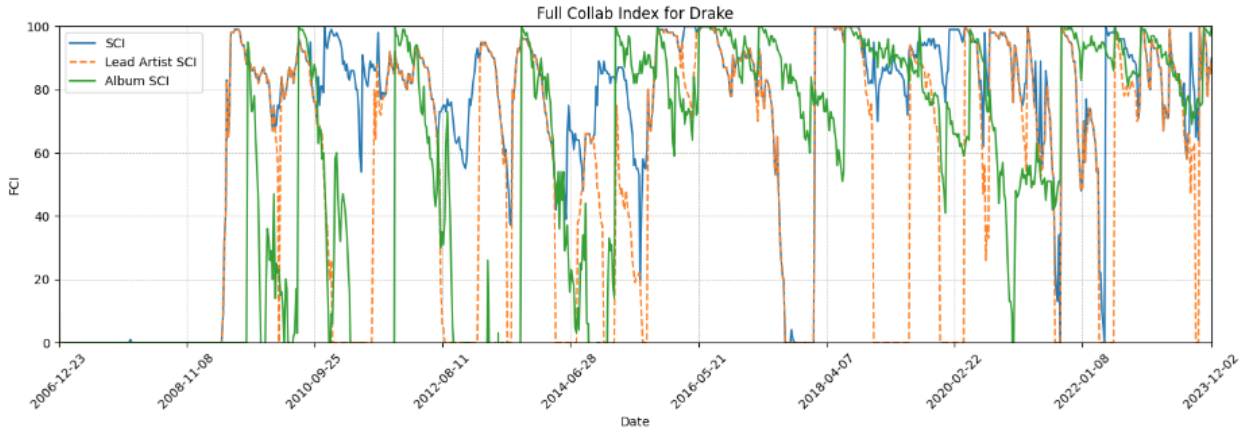
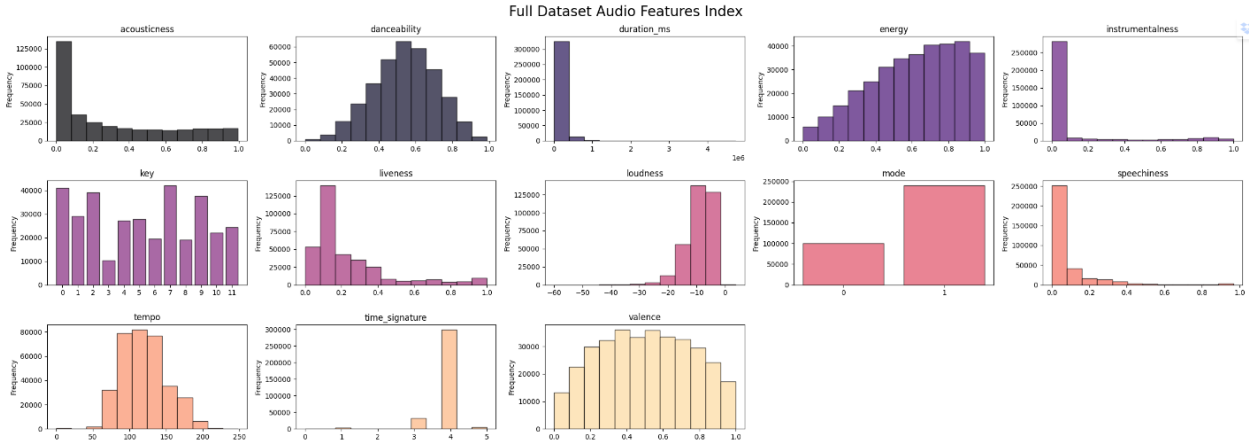


Figure 10: FCI(Full Collab Index) - Drake (From notebook)



4. **AFI(Audio Features Index)** - The Audio Features Index includes 10 audio features to have an overview of our whole dataset. We may say it's the distribution charts to offer us for all the major audio features involved.

Figure 11: AFI(Audio Features Index) (From notebook)



5. **AAFI(Artist Audio Features Index)** - This index offers the best individual artist overview for their specific audio feature distributions. When we compare two famous artists, we can still find out the major difference between Justin Bieber and Rihanna is that Justin Biber's loudness is skewing to the right more. Both of them had a wider range in tempo than the whole dataset. It implies the success of Justin Bieber and Rihana partially contributes to their capability to offer a much wider range of the music for audience to enjoy.

Figure 12: AAFI(Artist Audio Features Index) - Justin Bieber (From notebook)

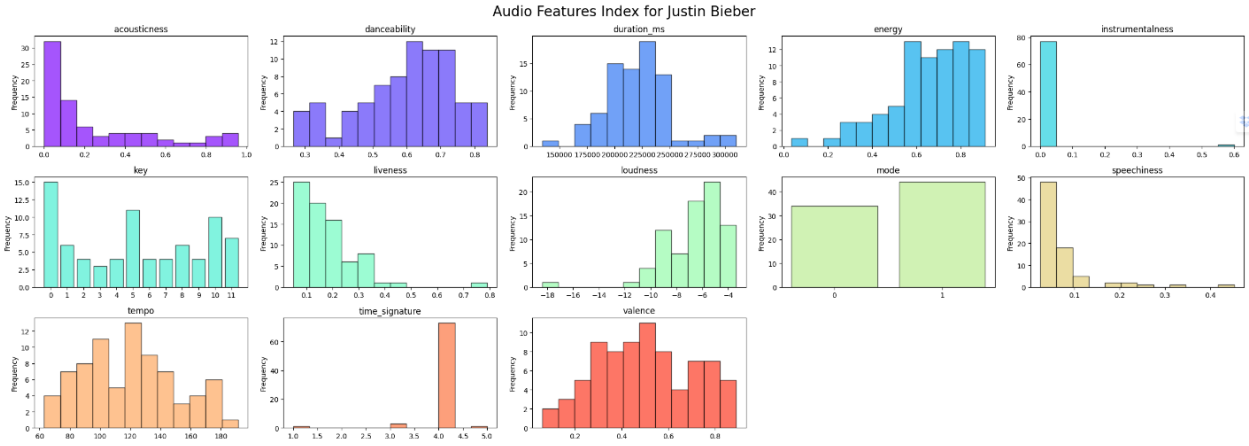
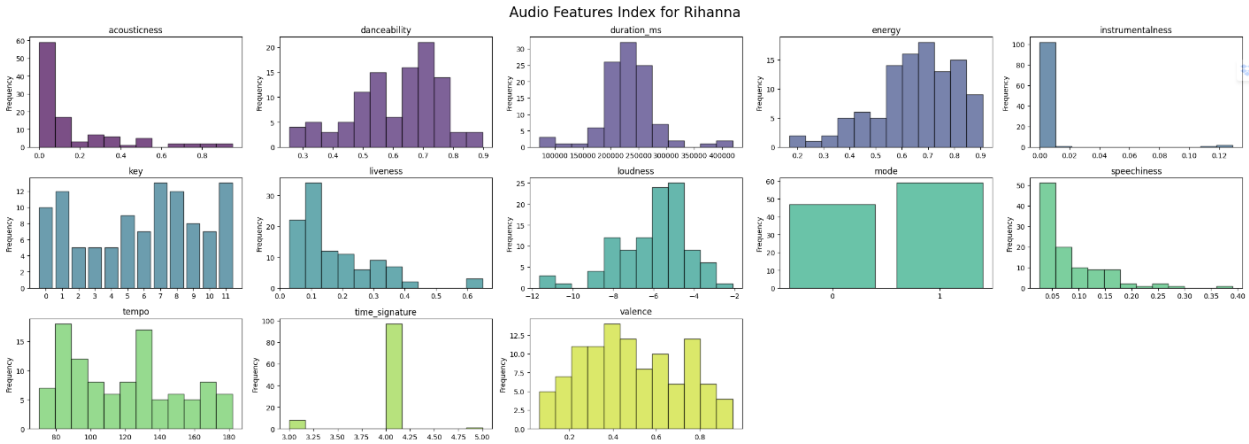


Figure 13: AAFI(Artist Audio Features Index) - Rihanna (From notebook)



6. **CCAFI(Current Collab & Audio Features Index)** - With the current (the most updated) collab index and our overall artist-specific audio features sets, we could generate the one-row data frame for a specific artist. Which feeds LLM and prepares the next chart for a better overview.
7. **PDAFI(Power-Law Distribution Audio Features Index)** - With the breakdown of two different major types of distributions, one of them is power-law distribution. So here we create a PDAFI(Power-Law Distribution Audio Features Index) To compare 2 artists' power-law-distribution-oriented audio features (acousticness, instrumentalness, liveness, and speechiness) similarity between 0 to 100.
8. **NAFI(Normal-Distribution Audio Features Index)** - With the breakdown of two different major types of distributions, one of them is closer to normal distribution. So here we create a NAFI(Normal Distribution Audio Features Index) to compare 2 artists normal-distribution-oriented audio features (danceability, duration in million seconds, energy, loudness, mode, tempo, time signature, valence) similarity between 0 to 100.
9. **RAFI(Radar Audio Features Index)** - Radar chart is the visual we believe is the best to provide multi-dimension in a glance to understand the strength between two artists in collaboration. We also offer a conclusion for the score between 0 to 100. In the chart to demonstrate RAFI(Radar Audio Features Index), we offer an easy 10-audio feature overview including acoustics, instrumentalness, liveness, speechiness, danceability, duration in million seconds, energy, loudness, tempo, and valence.

   That comparison has a great story to tell: In the Weeknd and Ariana Grande collaboration, the score is 90.97. We could easily know why they can successfully launch two BB100 songs due to their perfect match almost in 9 out of 10 dimensions: acoustics, liveness, speechiness, danceability, duration in

million seconds, energy, loudness, tempo and valence scoring 90 to 100. The only huge difference is in instrumentalness. This collaboration created a huge potential for their audience to accept their collab works. The producer only needs to focus on mitigating the gap in one and only audio features to make it work.

We specifically chose Kenny G, the Saxophone player, and Cardi B to highlight the differences between them. So their collab score is 79.42 even can't reach 80 on average. Many audio features have a big problem like danceability, speechiness, and instrumentalness.

Another pair to watch is the Norah Jones and Enimem. Norah Jones is famous for her soft and slow style of Jazz music but Enimem is the flagship of rap songs which is fantastic with a strong beat. It's absolutely not an easy one scoring 80.61 to explain the difficulties, but multiple audio features to fix their collaboration including speechless, loudness, instrumentalness, and energy - all below 80, which require a lot of work to make the collab successful.

Figure 14: AAFI(Artist Audio Features Index) - Weeknd & Ariana Grande (From notebook)

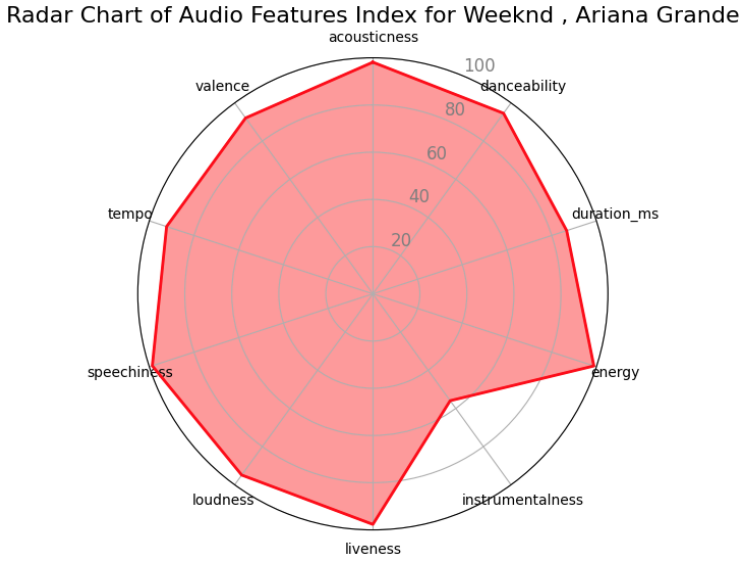


Figure 15: RAFI(Radar Audio Features Index) - Kenny G & Cardi B(From notebook)

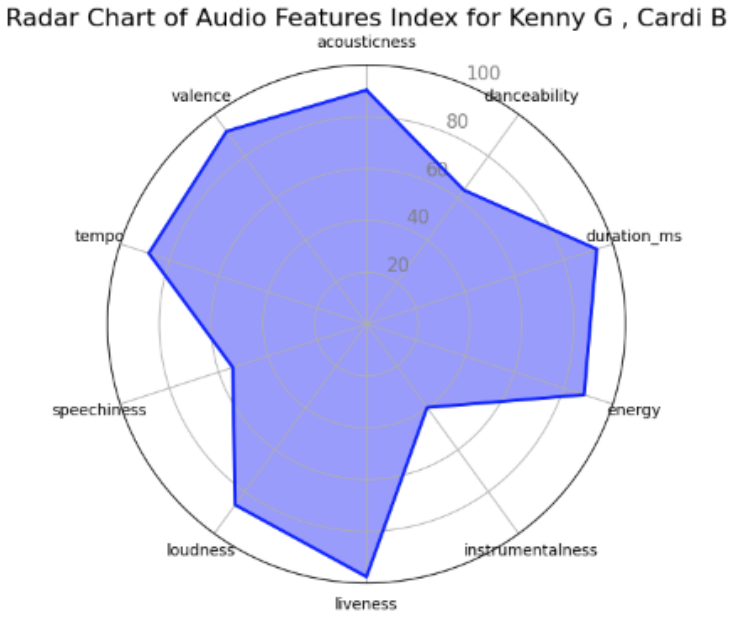
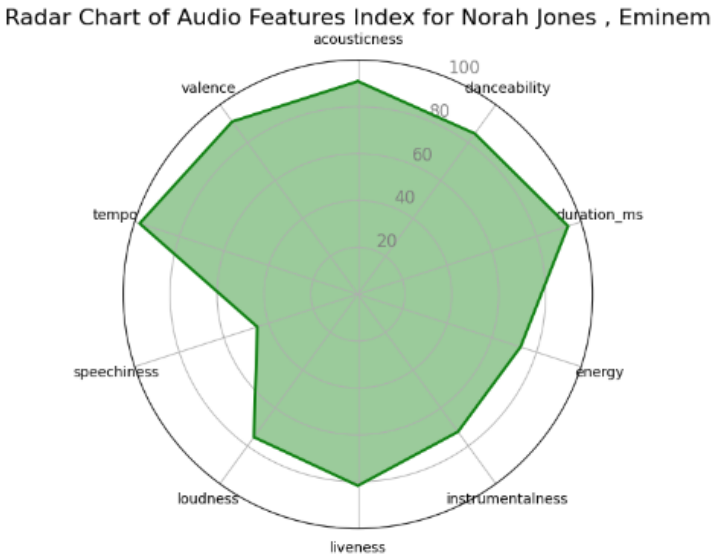


Figure 16: RAFI(Radar Audio Features Index) - Norah Jones & Eminem (From notebook)



Part II

Within Part II Our goal focuses on recommendation and question answering from our data after mining. We found that Streamlit offers a very simple approach to transform the data modeling and mining results to an easy-to-understand interface and operation

that both benefits the data scientist/analyst and the users. After our audio feature is chosen, it will be like this:

These features were collected from Spotify, for further explainability, see below for brief definitions:

- *acousticness: A confidence measure from 0.0 to 1.0 of whether the track is acoustic. 1.0 represents high confidence the track is acoustic..*
- *danceability: Danceability describes how suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity. A value of 0.0 is least danceable and 1.0 is most danceable.*
- *duration_ms: The duration of the track in milliseconds.*
- *energy: Energy is a measure from 0.0 to 1.0 and represents a perceptual measure of intensity and activity. Typically, energetic tracks feel fast, loud, and noisy. For example, death metal has high energy, while a Bach prelude scores low on the scale. Perceptual features contributing to this attribute include dynamic range, perceived loudness, timbre, onset rate, and general entropy.*
- *instrumentalness: Predicts whether a track contains no vocals. "Ooh" and "aah" sounds are treated as instrumental in this context. Rap or spoken word tracks are clearly "vocal". The closer the instrumentalness value is to 1.0, the greater likelihood the track contains no vocal content. Values above 0.5 are intended to represent instrumental tracks, but confidence is higher as the value approaches 1.0.*
- *liveness: Detects the presence of an audience in the recording. Higher liveness values represent an increased probability that the track was performed live. A value above 0.8 provides strong likelihood that the track is live.*
- *loudness: The overall loudness of a track in decibels (dB). Loudness values are averaged across the entire track and are useful for comparing relative loudness of tracks. Loudness is the quality of a sound that is the primary psychological correlate of physical strength (amplitude). Values typically range between -60 and 0 db.*
- *tempo: The overall estimated tempo of a track in beats per minute (BPM). In musical terminology, tempo is the speed or pace of a given piece and derives directly from the average beat duration.*

And the Radar chart is also suggested here:

Figure 17: The Streamlit Demonstration (To-be-continued) - Audio Features Index without LLM (From Streamlit)

These features were collected from Spotify, for further explainability, see below for brief definitions:

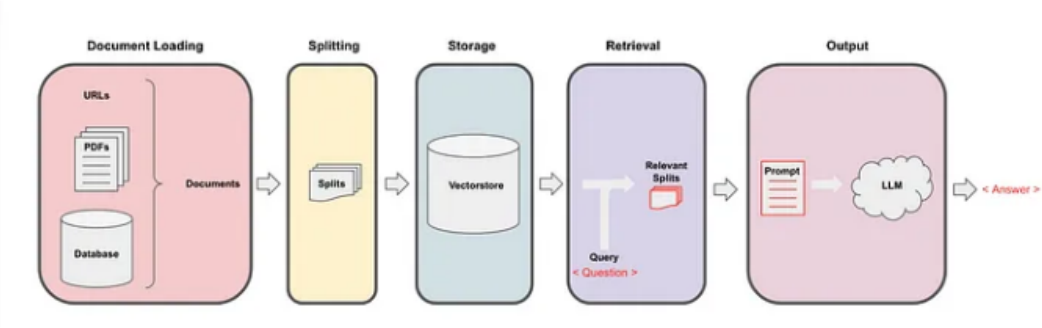
- **acousticness**: A confidence measure from 0.0 to 1.0 of whether the track is acoustic. 1.0 represents high confidence the track is acoustic..
- **danceability**: Danceability describes how suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity. A value of 0.0 is least danceable and 1.0 is most danceable.
- **duration_ms**: The duration of the track in milliseconds.
- **energy**: Energy is a measure from 0.0 to 1.0 and represents a perceptual measure of intensity and activity. Typically, energetic tracks feel fast, loud, and noisy. For example, death metal has high energy, while a Bach prelude scores low on the scale. Perceptual features contributing to this attribute include dynamic range, perceived loudness, timbre, onset rate, and general entropy.
- **instrumentalness**: Predicts whether a track contains no vocals. "Ooh" and "aah" sounds are treated as instrumental in this context. Rap or spoken word tracks are clearly "vocal". The closer the instrumentalness value is to 1.0, the greater likelihood the track contains no vocal content. Values above 0.5 are intended to represent instrumental tracks, but confidence is higher as the value approaches 1.0.
- **liveness**: Detects the presence of an audience in the recording. Higher liveness values represent an increased probability that the track was performed live. A value above 0.8 provides strong likelihood that the track is live.
- **loudness**: The overall loudness of a track in decibels (dB). Loudness values are averaged across the entire track and are useful for comparing relative loudness of tracks. Loudness is the quality of a sound that is the primary psychological correlate of physical strength (amplitude). Values typically range between -60 and 0 db.
- **tempo**: The overall estimated tempo of a track in beats per minute (BPM). In musical terminology, tempo is the speed or pace of a given piece and derives directly from the average beat duration.

Figure 18: The Streamlit Demonstration- Audio Features Index without LLM in Radar Chart (From Streamlit)

Part III

The demonstration is here. LongChain is a framework for building LLM applications such as chatbots, question-answering, and summarization with the idea that we can chain different components to create specific use cases. Of the many components within LongChain, Retrieval augmented generation allows for a specific dataset to be included to add more relevant information and prevent LLM hallucinations and out-of-date training.

Figure 19: Retrieval Augmented Generation(RAG) Flow Chart- (From Medium blog)



Retrieval Augmented Generation(RAG)

We are currently focusing on the possibility of if we can use our data after mining to make recommendations for the user based on input of audio features and the collaborative risk index from part I. Ideally, this should allow users to then ask questions about the recommended artist to "collaborate with" and answer them by providing insights into the type of features and value of the features for the artist recommended and how it closely aligns with the user stylistically. We were able to achieve a simple example trained on some mock data.

From this mock data we imported, embedded, and stored this data into a vector store using Longchain and experimented with 3 types of queries

Figure 20: LLM - GPT Reply (From notebook)

```
[ ] query = "if my danceability is 9, what artist should I collaborate with"
    response = chain({"question": query})
    print(response['result'])

    It looks like Calvin Harris has a collaboration score of 89 with a danceability of 9, so you may want to consider collaborating with him.

[ ] query = "if my danceability is 8, what artist should I collaborate with"
    response = chain({"question": query})
    print(response['result'])

    I don't know.

    query = "if my danceability is 8, what artist would be close to this, what artist should I collaborate with, recommend the artist and give all reasons"
    response = chain({"question": query})
    print(response['result'])

    You might want to consider collaborating with Calvin Harris. His songs have a danceability score of 9 and a collab_score of 89, which is close to your desired danceability of 8 and collab_score of 60.
```

As we can see from the examples if we enter the exact danceability of 9 from the mock training data it outputs Calvin Harris. If we ask about a danceability of 8 which does not exist in the mock data then the output is "I don't know". An additional test was done asking if the danceability was 8, what artist would likely be close to this, and best to collaborate with. Interestingly with more follow-up questions the LLM(GPT) was able to infer Calvin Harris and also explain why it made this choice

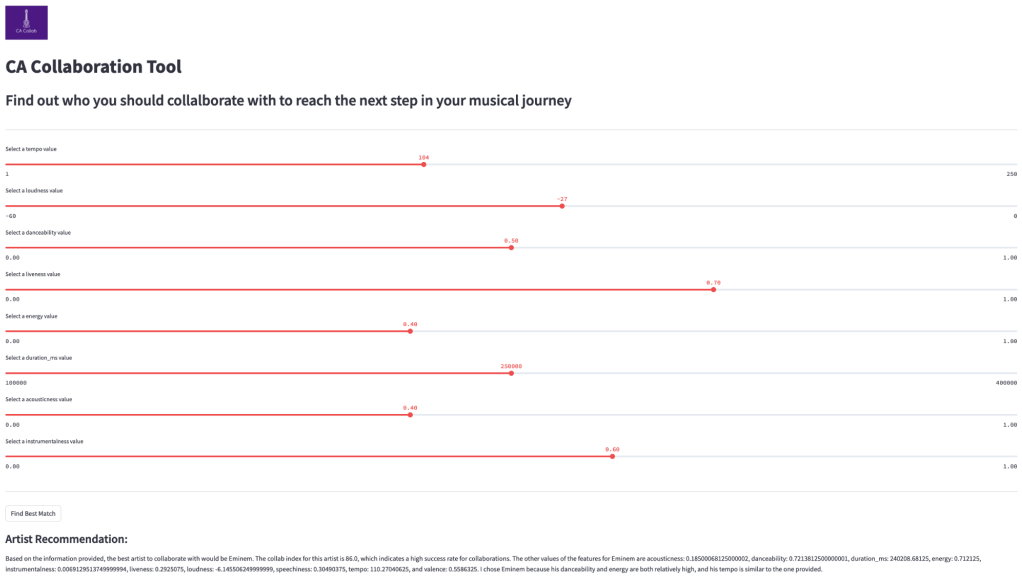
Part IV

We eventually found the comparison between GPT and Mistral, the speed of GPT stands out! The output from the API for users to see the results, and experience of the speed is the top factor from our experiments because users always are desperate to get the output due to they are expecting conversational results that require time to digest and think. If the wait time is too long, the model may not offer the value users want. In our example below, we reduce the scope of our models and the output could be less than 1 second, which is nice for users to stick to this interface and keep the operation for the next query. HERE is the site (API key needed)

In the example below, we can see the user could successfully finetune the parameters, then the streamlit will demonstrate according to our part 1 result, using part 2 streamlit experience, the part III LangChain, and LLM feedback, as our Part IV demonstration for the output as below:

"Based on the information provided, the best artist to collaborate with would be Eminem. The collab index for this artist is 86.0, which indicates a high success rate for collaborations. The other values of the features for Eminem are acousticness: 0.18500068125000002, danceability: 0.7213812500000001, duration_ms: 240208.68125, energy: 0.712125, instrumentalness: 0.0069129513749999994, liveness: 0.2925075, loudness: -6.145506249999999, speechiness: 0.30490375, tempo: 110.27040625, and valence: 0.5586325."

Figure 20: The Streamlit Demonstration - Audio Features Index with LLM (From Streamlit)



V. RESULTS AND CONCLUSIONS:

1. Collaboration deserves more study from the data - From our result of various dashboard-like indexes to monitor or realize the status and easy visuals, we could digest the complex collaboration impact toward artists in a simpler and faster approach. Audio features may not be the all, but at least a valuable direction to mine for more potential values for collaboration choices and success.

2. Collaboration deserves more application to demonstrate the exploration - In our demonstration, the interface of Streamlit becomes an intuitive access to translate hard-core models and output for general consumers' hands-on experience.

3. Collaboration deserves more tools with nuance to analyze - In our demonstration of the LangChain, or the LLMs like GPT, it's believed with a huge benefit to leverage the powerful capabilities to deliver the data insight for the users to understand more about the potential and all kinds of possibilities from data input to output.

VI. BROADER IMPACTS:

1. App-like dashboard for artists in the music industry- In our demonstration of the LangChain, or the LLMs like GPT, it's believed with a huge benefit to leverage the powerful tools to communicate the collaboration decisions.

a. ***For record companies -*** it could be an easy tool to manage their status and watch for the industry dynamics.
b. ***For the artist -*** it could be their source of the collab opportunities.
c. ***For the fresh music learners -*** it could be their fields of experiments and inspiration.

2. Audio-feature analysis - Our radar chart and various indexes demonstrate the potential to investigate in detail with more features with visuals

a. ***For record companies -*** it could offer them to invest and focus on the technical details of the performances and qualities of the audio features with better equipment and measurements
b. ***For the artist -*** it could be their source of tools to instantly reflect their creations and business potential.
c. ***For fresh music learners -*** it could offer a brand new field for them to practice with more easy-access established data and examples.

3. Facility of LLM and platform -

a. ***For record companies -*** it could be a common platform to integrate cross-department opinions for productions.
b. ***For the artist -*** It's natural for the artist to get feedback in a faster and more convenient tool.
c. ***For the fresh music learners*** - it's a complete kick-start for learning.

VII. STATEMENT OF WORK

1. **Content** - Alvin leads the Part I content and Sundar leads the Part II, III, and IV.
2. **Delivery** - Alvin and Sundar co-work for stand-up 1. Alvin leads the stand-up II, and Sundar leads the stand-up III. Alvin and Sundar take turns evenly for the remaining deliverables
3. **Final works** - Alvin leads the report and Sundar leads the poster.

VIII. INCORPORATED FEEDBACK

1. **Don't use chart-only data -** It's super helpful to light up when we're stuck in the tunnel. We incorporated the data off-chart from Billboard and on-the-chart of Billboard both could demonstrate the value of finding the similarity. We eventually decided to utilize all the audio features from both datasets. The benefit is we increase our portfolio to find out the most comprehensive audio features from an artist. It also directly avoids the sparse universe to find limited data for our goal. Eventually the performance of leveraging the whole audio features works very well specifically for Radar visualization.

2. **Narrow down the scope, and check the data imbalance nature** - It's absolutely a very insightful advice. We started with a very large scope. At that moment our neural network model shows the imbalance nature will lead us to amazing accuracy, even fabulous recall, precision, and the F1 score. However, when we test it with AUROC, it shows the vulnerable nature of the super-imbalanced data problem. We did try any possibility from all kinds of SMOTE: Adasyn, SMOTE, borderline SMOTE, and Random Over Sampling -

none of them could help us to improve in any significant way. It demonstrated that we should immediately change the approach to lead to our goal of finding similarities in collaboration. Though we completely threw away our NN models and all the SMOTE-related experiments - it's a wonderful learning and eventually we're so grateful for the incorporated feedback from instructors.

3. **Considering the trend and popularity could be different in various time spans** - We did think it was fabulous advice for us. Our action item is to incorporate the parameter for functions of the index even if the default is the whole dataset. There's a parameter to choose the year like after 2000, or 2010 to get the most recent years to make the recommendation useful depending on consideration or request for the shorter or longer term on top of the default.

IX. REFERENCES

[1]Billboard 200: The Lessons of Musical Success in the U.S. (Gourévitch, Boris, 2023), Sage Journal

[2]Collaborative Song Dataset (CoSoD): An annotated dataset of multi-artist collaborations in popular music(Duguay, M., Mancey, K., Devaney, J., 2023), 24th International Society for Music Information Retrieval Conference

[3]Detecting Collaboration Profiles in Success-based Music Genre Networks(Oliveira, P., Silva, M., Seufitelli, D., Lacerda, A., & Moro, M.,2020) 21st Int. Society for Music Information Retrieval Conf

[4] Using langchain for Question Answering on Own Data (Mishra, O, 2023), Medium

APPENDIX

Course #	Course Title	Technical Indepth
501	Being a Data Scientist	Utilize the three-phrase approach to conduct project
502	Mathematics Methods for Applied Data Science	Apply math methods into our project
503	Data Science Ethics	Take Ethical consideration into projects
505	Data Manipulation	Master the data manipulation
515	Efficient Data Processing	Attemp to make the operarion more efficient
516	Big Data: Scalable Data Processing	Try to make our big data operation scalable
521	Visual Exploration of Data	Master the visuals to explain our findings
523	Communicating Data Science Results	Demonstrat the consise reulst to the point
524	Presenting Uncertainty	Consider the uncertainty into our modelling
532	Data Mining I	Utilize the similairty score
542	Supervised Learning	Try supervised moedlling prediction in NN models
543	Unsupervised Learning	Try unsupervised to group in genre first
593	Milestone I	Team collaboration experiences
632	Data Mining II	Consider the retrieval system into our plan
642	Deep Learning I	Utilize NN model to run the first attemp of predicdtion
643	Machine Learning Pipelines	Apply our pipeline operaiton with GitHub
652	Network Analysis	Apply networking into our first-stage project scope
655	Applied Natural Language Proccessing	Applying NLP-oriented LLM to our output
685	Search and Recommender Systems	Utilize the recommendation system concept
691	Independenty Study	Apply the project management and SMOTE experiences
696	Milestone II	Team work specializatio and time management