Discovering the top-selling artist music network & audio feature variations

Joe Amedeo, Kenia M Way, Max Kanaskar
MS, Analytics, Georgia Institute of Technology
{jamedeo3, kway3, mkanaskar3} @gatech.edu

ABSTRACT

Music goes back as early as history has been recorded, and it surely won't be leaving anytime soon[1]. Music is fascinating, universal, and has become an outlet for activism and human connection across different demographics. Nowadays, technology has been increasingly used in the music industry, anywhere from creating algorithms to understand users' preferences, to predicting the optimal length or composition of a track. Music is a rich, dynamic system and there are some musicians that have truly understood the industry and sold millions of records. We are interested to explore the nature of this network of top performing artists and follow the path of Euler by applying graph theory to understand it.

We analyzed the music network graph between musicians that have sold more than 1 million certified albums, according to the Record Industry Association of America (RIAA), by leveraging metrics from Spotify's API.

Keywords

Music, Spotify, RIAA, Graph Theory, Dimensionality Reduction, NLP.

1. Problem Statement

Currently, there is no easy way to understand what makes a specific album track a top track. There could be multiple combinations of the features that make up a top track, and it can be difficult and time consuming for consumers to analyze all this in the search of what's trending in the industry while being consistent with their own music taste. Furthermore, we explored what happens if we capture the optimal combination of features for top tracks and artists, given a track and its feature set.

Lastly, we generated a recommendation system based on genres for top artists and used NLP to explore common traits for their lyrics.

2. Data Source

RIAA.com[9] was scraped for musicians that have sold over one million records on 10/01/2022 that consisted of 1,806 artists. Using this initial starting point, Spotify's API[7] was accessed to obtain two main data sets: a musician's related artist dataset and each musician's top tracks dataset described below. After several iterations of cleaning the data due to integrity issues, the number of artists reduced to 1,730. One reason for the reduction is some musicians, like Neil Young[8], refuse to allow Spotify to stream their music.

2.1 Top Tracks Dataset

Musician's top tracks dataset consisted of 17,307 songs where each artist has approximately 10 top songs. The main audio features in this dataset are: *danceability*, *energy*, *key*, *loudness*, *mode*, *speechiness*, *tempo*, *valence*, *acousticness*, *instrumentalness*, *valence*, and *liveness*. These characteristics are unique to each song and definitions, defined by Spotify, are provided in the Appendix's Data Dictionary section.

2.2 Related Artist Dataset

Musician's related artists dataset consists of 35,063 records where each artist has approximately 20.27 *Fans Also Like* artists, according to Spotify. An artist that sold over 1 million records was considered a node artist with their corresponding Fans Also Like artist was considered an edge artist.

Spotify's Fan also like page for The Beatles

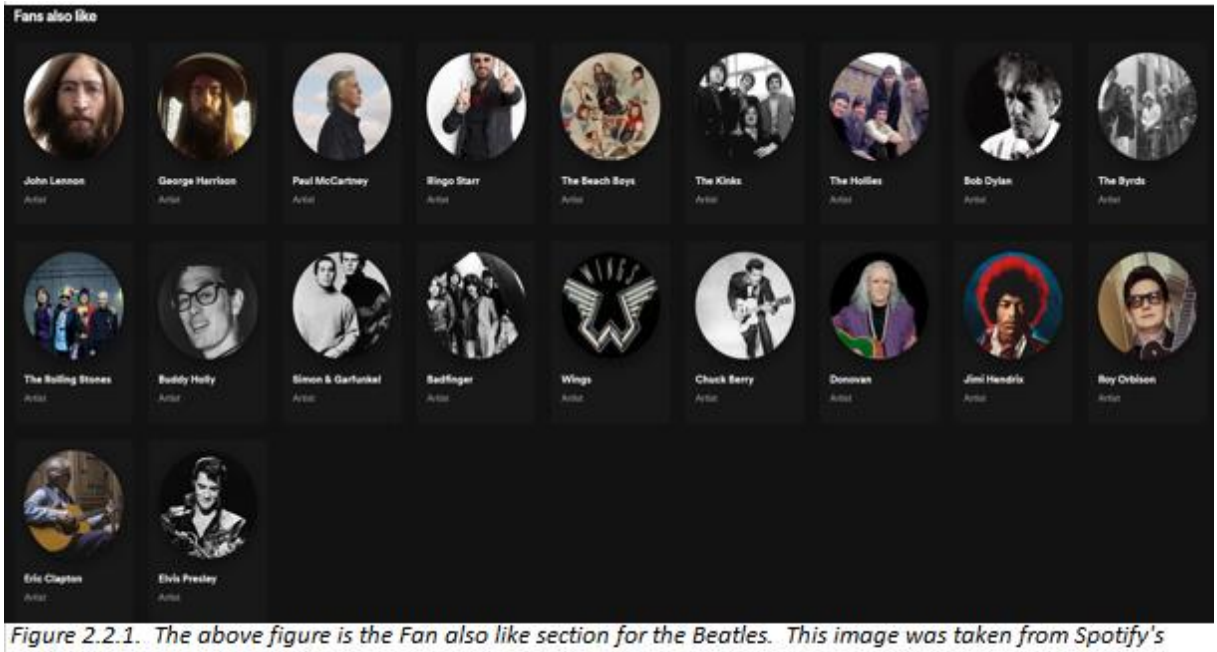

Figure 2.2.1. The above figure is the Fan also like section for the Beatles. This image was taken from Spotify's website on 12/01/2022.

3. Methodology

3.1. Graph Analysis

The first step of our analysis was constructing an undirected, unweighted adjacency matrix *A* from the musician's related artists dataset. *A* is a symmetric matrix defined by

$$A_{i,j} = A_{j,i} = \begin{cases} 1 \; if \; a \; node \; artist \; is \; connected \; to \; an \; edge \; artist \\ 0 \qquad\qquad\qquad otherwise \end{cases},$$

where *i* is the node artist index, *j* is the edge artist index, and $A \in R^{8,108x8,108}$. The Laplacian matrix L defined by $L = D - A$ where D is the degree matrix of A defined by

$$D_{i,j} = \begin{cases} \deg(A_i) & if \; i = j \\ 0 & otherwise \end{cases},$$

where $L \in R^{8,108x8,108}$. The number of connected components of the Laplacian matrix is denoted by the multiplicity of the 0 eigenvalue[13] where the eigenvalues and eigenvectors are obtained by solving

$$(L - \lambda I)V = 0\,,$$

where λ is the eigenvalue corresponding to the eigenvector V and I is the identity matrix. The nonzero index of an eigenvector corresponding to a 0 eigenvalue denotes the components, or in the context of our problem the artists, that are uniquely connected within the entire network. Three main connected components of artists were discovered, discussed more in detail in section 4.1, and analyzing these three clusters is now the focus of the paper.

3.2. Dimensionality Reduction

After identifying the three largest clusters of artists, the next step is finding a meaningful low-dimensional structure hidden in their high-dimensional space. Isometric feature mapping, or *Isomap*, is an effective three step algorithm, summarized below, for representing nonlinear structures typically undetectable to classical techniques such as PCA and MDS[14].

For each cluster:

1. Construct an adjacency matrix similar to step 1 in section 3.1, limiting the number of node-to-edge artist connections to the first 9 nearest neighbors.
2. Compute the shortest path graph search.
3. Construct a three-dimensional embedding with partial eigenvalue decomposition with the embedding encoded in the three largest eigenvectors.

Once complete, the components were combined back with the original artists and K-Means clustering was implemented on the three-dimensional structure with K=9 clusters to obtain discrete labeling for each artist unique to each cluster. Next is combining the cluster assignment and K-Means label for each artist with the tracks data to discover audio feature characteristics popular with each graph.

3.3. Exploratory Data Analysis (EDA) - Top Tracks

The EDA included data visualization and Principal Component Analysis, *PCA*, to unveil any hidden patterns within each graph and reduce the number of features required to capture variability when analyzing the track features. We applied clustering algorithms on the reduced feature space to explore any further groupings that might emerge and find its relationship to the original dimensions.
Each of the three artists clusters were analyzed one by one. For each cluster, we analyzed all songs for artists in that cluster for various track features (see Appendix for the list of the track features) to explore how the track features vary.

To analyze the sub-clusters of tracks, we first applied PCA to reduce the dimensionality - a total of 11 track features was reduced to 7. We then used k-means clustering to identify sub clusters of similar songs within each artist cluster of songs.

3.4. Natural Language Processing (NLP)

We incorporated a final layer of analysis by running a couple of NLP techniques on the results obtained by the graph analysis and artists traits for generating recommendations.

For recommendations, we used the TF-IDF (Term Frequency - Inverse Document Frequency) feature selection approach to model the text-based features using cosine similarity.

Cosine similarity: computes the L2-normalized dot product of vectors and could be interpreted based on the closer the cosine value to 1, the smaller the angle the greater the match between vectors.

In addition, we explored adding another data source levering python's *lyricsgenius* package, a python client for the Genius.com API. This allowed to explore any pattern that top artists within the defined clusters might share in their popular songs.

4. Evaluation and Final Results

4.1 Related Artist Analysis

After implementing steps outlined in section 3.1 and 3.2, three clusters emerged consisting of 764, 170, and 110 musicians. As expected, due to the nature of how we consume music, one large cluster emerged while the remaining clusters are relatively smaller in size. Figure 4.1.1, Figure 4.1.2, and Figure 4.1.3 below are three-dimensional projections of each cluster after combining the artists back with the output of the Isomap algorithm.

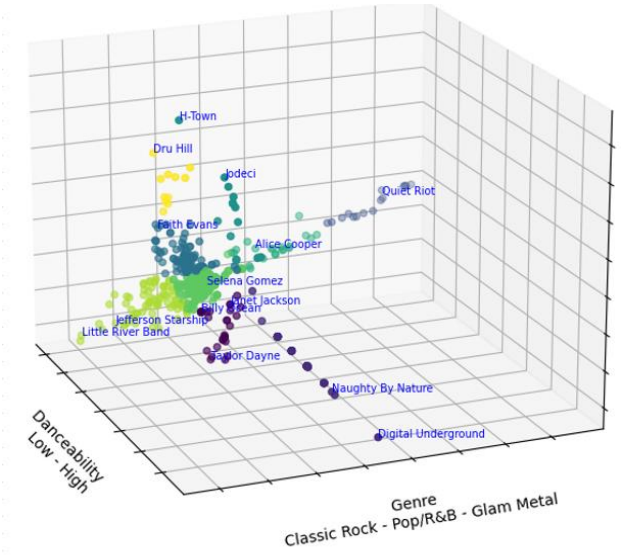

Figure 4.1.1. The above figure is the 3-dimensional representation of cluster 1 after performing an Isomap dimensionality reduction. The different colors represent the Kmeans label, helpful for viewing purposes.

Figure 4.1.1 to the left represents the artists in the largest cluster. Only two components had a high correlation with features that define the cluster: danceability and genre. This was expected for the first cluster due to individual's enjoying a wide variety of music spanning different generations and genres. As seen in the preceding clusters, a recommendation is to filter cluster 1 by genre or time period prior to performing dimensionality reduction techniques.

Figure 4.1.2 to the right represents the artists in the second cluster consisting of primarily Country music artists. This cluster presented three clearly defined axes: genre, gender of the singer, and tempo. The majority of musicians in this category tend to be male and are centered around the country-pop genre. The Tempo axis will have artists that sing more ballads at the top and quicker, country-rock songs at the bottom of the axis. According to Spotify, the fans that listen to this cluster the most live closest to U.S. cities Dallas, Houston, and Atlanta.

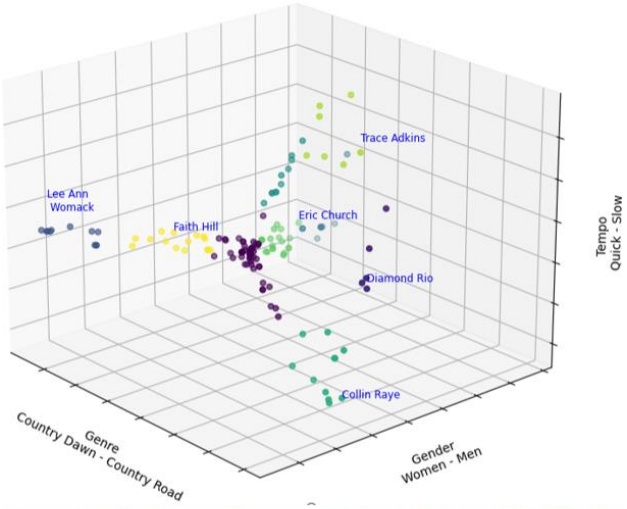

Figure 4.1.2. The above figure is the 3-dimensional representation of cluster 2 after performing an Isomap dimensionality reduction. The different colors represent the Kmeans label, helpful for viewing purposes.

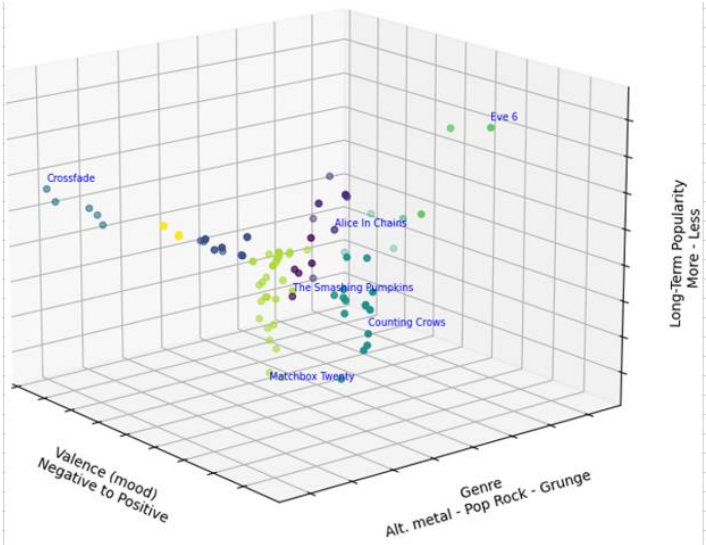

Figure 4.1.3. The above figure is the 3-dimensional representation of cluster 3 after performing an Isomap dimensionality reduction. The different colors represent the Kmeans label, helpful for viewing purposes.

Figure 4.1.3 to the left are artists in cluster 3 and are primarily Pop-Rock bands popular from late 1980s to early 2000s. Interestingly valence, a scale in which a song's mood is measured, had a high correlation with the cluster. Popular bands like Crossfade, Nirvana, and Alice In Chains typically scored lower on this scale. The Long-term Popularity scale were bands that typically were popular throughout this time period and had several hit albums and songs. Bands like the Goo Goo Dolls, Matchbox Twenty, and Smashmouth all scored high on this scale.

4.2. EDA – Track Analysis by Cluster

We looked at over 17,000 top tracks across all artists to see what track characteristics are associated with the top songs. The first analysis we did was to look at the track characteristics - for example, speechiness, instrumentalness, etc. to try to see if there are any natural clusters that emerge across these characteristics. Please see the appendix for the definitions of these characteristics. We applied PCA treating the 11 dimensions as "observations" across the top tracks and plotted the distribution in 2 dimensions.

4.2.1. EDA – Track Analysis by Cluster

We also analyzed the tracks by clusters of connected artists to see if there were any common themes. We analyzed three clusters of connected artists.

Artist Cluster #1

This is the largest cluster of artists with genres spanning: pop, rock, R&B and soul. There are multiple artists spanning various decades that feature numbers across these genres and sub-genres of music, and crossovers such as pop rock. There are a total of 3872 songs in this cluster.

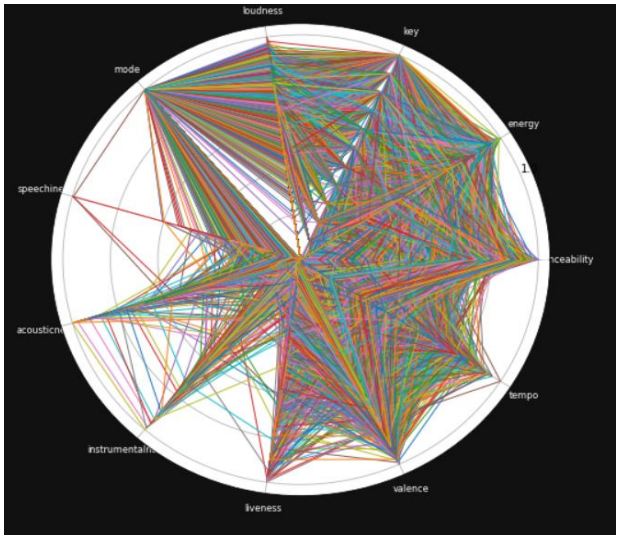

Figure 4.2.1. Artist cluster 1.

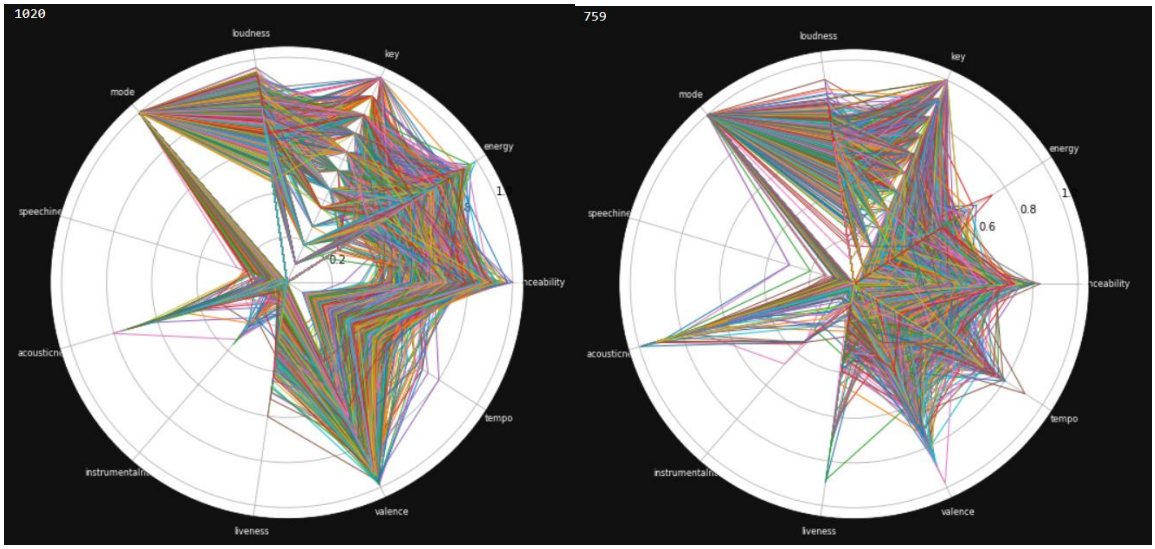

Figure 4.2.2. (left) Artist cluster 1- Sub-cluster 1, (right) Artist cluster 1- Sub-cluster 2

To the left, one of the largest sub-clusters of songs for artist cluster #1. This subcluster of connected artists trends medium to high on danceability and valence, with loudness and energy being a mix. This sounds like a cluster of tracks that are positive/cheerful and are great dance tunes. To the right, there is another cluster which seems to be a collection of live performances with high acousticness and low to medium energy level, tempo and valence. This seems to be a collection of songs with no electric/electronic sound production, but more of the natural acoustic sound in a live setting.

Artist Cluster #2

Artist Cluster 2 is a collection of country music songs, the second largest cluster with 980 songs. Country music often consists of ballads and honky-tonk dance tunes with generally simple form, folk lyrics, and harmonies often accompanied by string instruments such as electric and acoustic guitars, steel guitars (such as pedal steels and dobros), banjos, and fiddles as well as harmonicas.

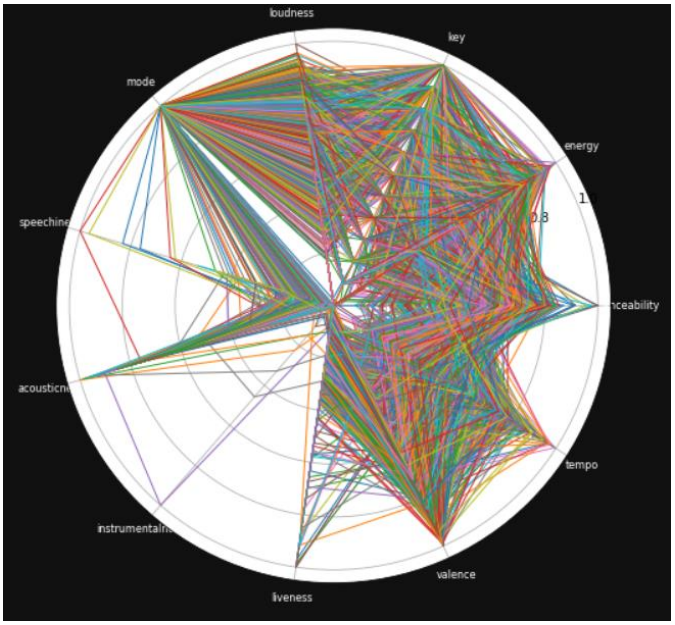

Figure 4.2.3. Artist cluster 2.

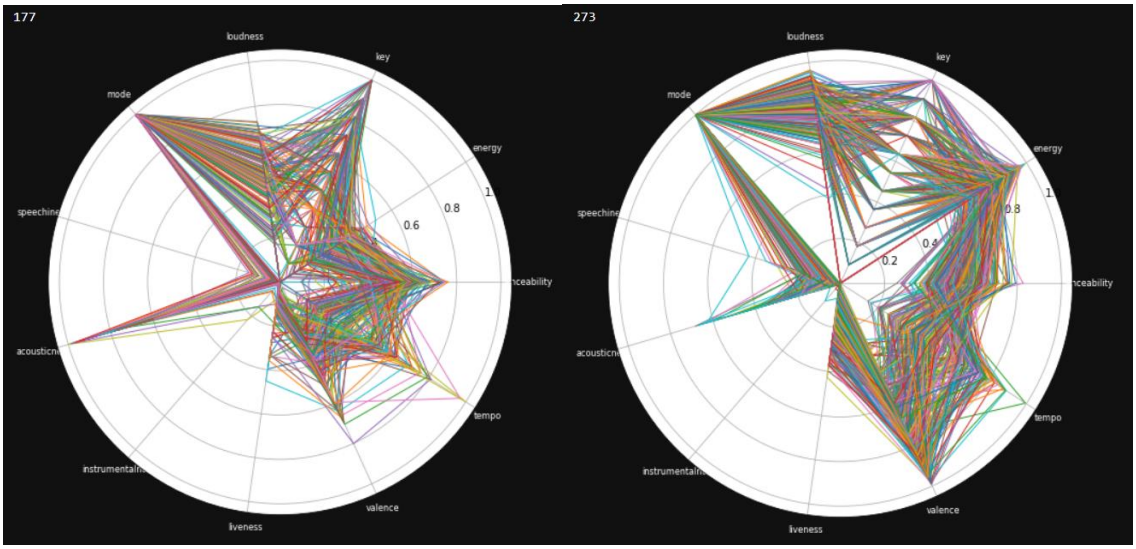

Figure 4.2.4. (left) Artist cluster 2- Sub-cluster 1, (right) Artist cluster 2- Sub-cluster 2

To the left of image 4.2.4, we find one subcluster of songs within this cluster which is low on energy, valence and tempo, but high on acousticness – this seems to be a collection slow, acoustic, probably melancholic/sad numbers. Subcluster 2, is a substantially large cluster of songs that seem to have good energy, loudness, and valence - perhaps a collection of energetic numbers with positive themes.

Artist Cluster #3

This cluster is a collection of artists in Rock, Alternative rock, Pop rock, Post-grunge, Power pop, Pop-punk categories. This cluster appears to be the smallest of the three artists clusters, with only 690 songs. Overall, these music tracks seem to be low on speechiness and acousticenss, and medium to high on instrumentalness dimensions.

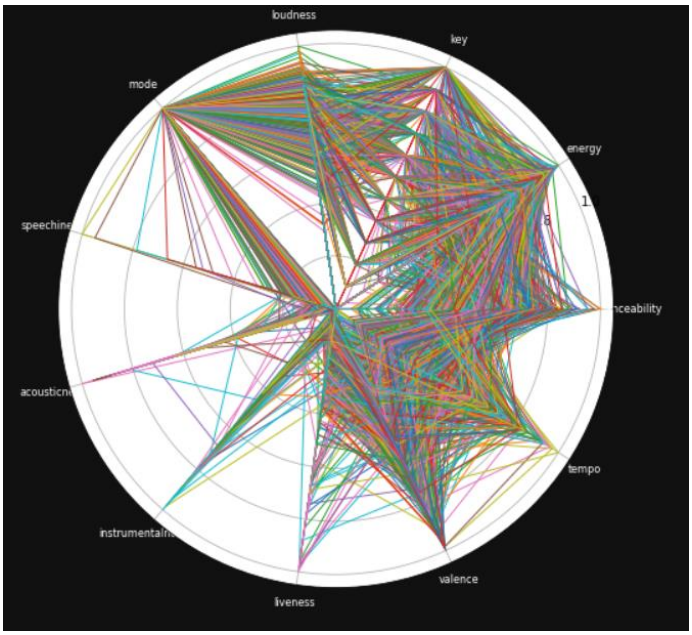

Figure 4.2.5. Artist cluster 3.

Within this cluster of songs, there is a sub cluster (left) which is low on tempo and valence, and medium to high on energy and loudness – these probably seem to be a collection of songs that are not very upbeat/positive, perhaps have somewhat darker or sadder in themes.

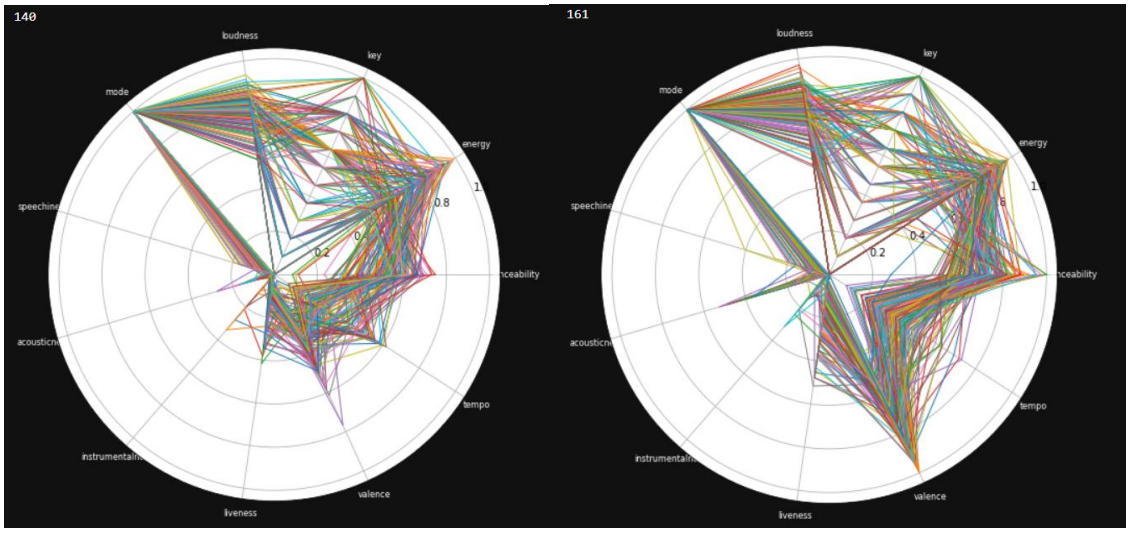

Figure 4.2.6. (left) Artist cluster 3- Sub-cluster 1, (right) Artist cluster 3- Sub-cluster 2

Interestingly, there is another subcluster (right) of songs that is the opposite: high in valence and danceability and energy - so perhaps a collection of upbeat/positive themed songs that also seem to have danceability.

4.2.2. PCA on Top Track Features

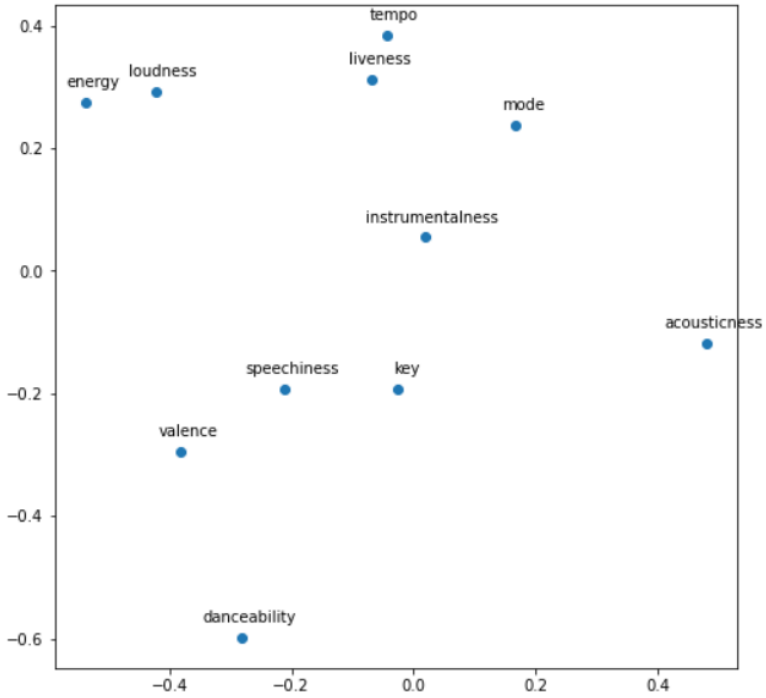

Figure 4.2.7. PCA results

As might be expected, there is some clumping of these characteristics we observe:

1. Energy and Loudness: Loudness is the decibel level, the physical strength of the song, and it goes together with the energy, which is the intensity and activity of the song.

2. Valence, Speechiness, and Key: There is association of Key with different emotions, valence indicates if the number is sad or cheerful, and is perhaps correlated with that as well e.g., sad songs perhaps have low speechiness, while cheerful songs the opposite.
3. Liveness and Tempo: It is a well-known fact that singers and artists perform their numbers at a higher tempo when performing live, perhaps to maintain the energy level and to keep the crowd engaged and entertained.
4. Danceability and Acousticness: These characteristics are isolated and not really associated with other characteristics.

4.4. NLP - Recommendation System

Here is the result of applying text analysis on the genres found in each of the analyzed clusters.

You can easily see the differences and similarities of these artists at a glance, which were described in more detail within the EDA artists analysis.

In each cluster we can see which genres stand out and leverage this to try to recommend similar artists to users when they want to find popular artists while being consistent with their preferences.

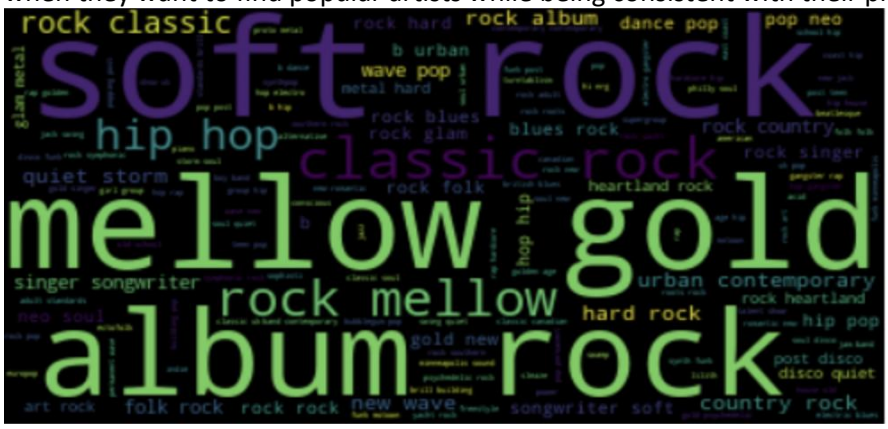

Figure 4.4.1. Wordcloud for genres found in Cluster 1

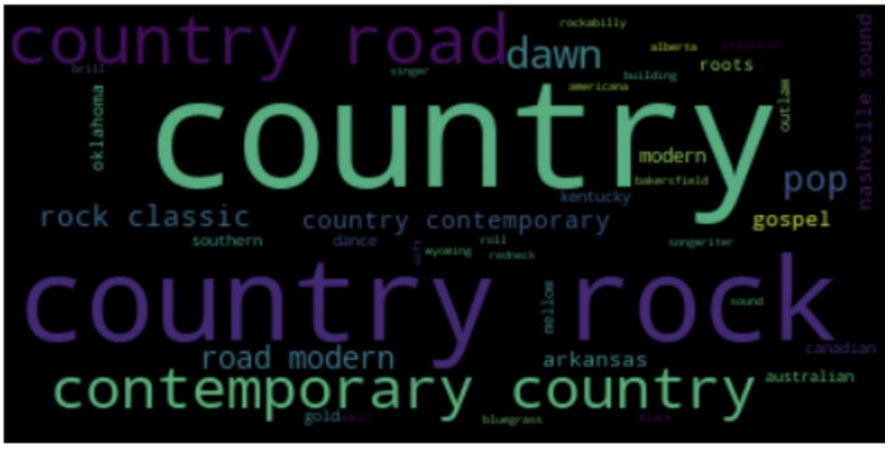

Figure 4.4.2. Wordcloud for genres found in Cluster 2

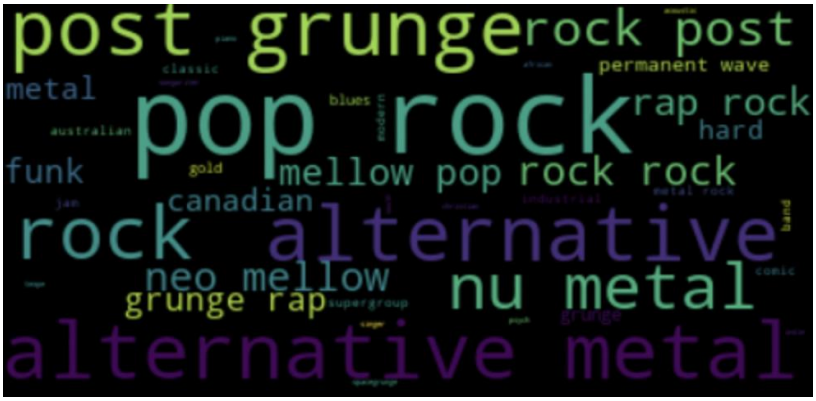

Figure 4.4.3. Wordcloud for genres found in Cluster 3

In this part of the project, we applied TF-IDF with linear kernel to generate a recommendation system on the artist data filtered by cluster. Based on the genre that users like, they will get recommended the top 20 most related artists.

Here are examples of the output obtained when applying this algorithm by cluster.

Recommended artists by selected artist in Cluster 1

selected_artist	recommended_artist	cluster
Joe	[Mýa, Aaliyah, Brandy, Toni Braxton, Amerie, M...	1
Berlin	[Thompson Twins, Howard Jones, Corey Hart, Mr....	1
702	[Brownstone, Changing Faces, Xscape, Jaheim, D...	1
Blaque	[3LW, Danity Kane, Janet Jackson, The Pussycat...	1
H-Town	[Shai, Hi-Five, Tevin Campbell, Soul For Real,...	1

Recommended artists by selected artist in Cluster 2

selected_artist	recommended_artist	cluster
Crystal Gayle	[Eddie Rabbitt, The Oak Ridge Boys, Dan Seals,...	2
Josh Turner	[Mark Wills, Chris Young, Thomas Rhett, Eric C...	2
Doug Stone	[Ty Herndon, David Ball, BlackHawk, Sawyer Bro...	2
Kane Brown	[Josh Turner, Mark Wills, Chris Young, Thomas ...	2
Brooks & Dunn	[John Michael Montgomery, Neal McCoy, Trace Ad...	2

Recommended artists by selected artist in Cluster 3

selected_artist	recommended_artist	cluster
Chevelle	[Puddle Of Mudd, Crossfade, Staind, Theory of ...	3
Crossfade	[Breaking Benjamin, Shinedown, Trapt, Staind, ...	3
Default	[Three Days Grace, Theory of a Deadman, Eve 6,...	3
Hoobastank	[Trapt, Staind, Hinder, 3 Doors Down, Scott St...	3
Finger Eleven	[Theory of a Deadman, Crossfade, Three Days Gr...	3

We ran a couple of experiments during this analysis: TF-IDF similarity kernels (cosine vs linear) and the trust of our evaluation approach.

TF-IDF similarity kernels

we applied TF-IDF with Cosine Kernel and TF-IDF with Linear kernel. We obtained the same similarity scores in both cases, and the difference between kernels were around their performance.

Linear kernel was computationally less expensive and we prioritized performance.

System trust

We ran an experiment to evaluate the system based on its trust and compared the recommendations obtained by our approach with those found on Spotify.

The artist selected for this experiment was Gwen Stephanie. We observed that – Pink, Fergie and Christina Aguilera – were usually included in the recommended artists by sites like imdb, spotify and all music, and these artists were also recommended by our approach.

Like every recommended system, this one has its own advantages and disadvantages. Our biggest disadvantage is accuracy in the results as we're only using one feature ("genre"). An advantage of our approach is to provide recommendations when we don't have a lot of information from the users (no cold start).

4.5. NLP - Lyrics Analysis

In this part of the project, we used the python's *lyricsgenius* package to explore any pattern that top artists within the defined clusters might share in their popular songs.

There is a limit on how many requests per second and numbers of requests per month when using the genius-song lyrics API free tier. Due to this limitation, we were not able to run a word cloud by all artists in each cluster, instead, we selected 5 artists by cluster and retrieved their top 3 most popular songs.

The following figures are the wordclouds generated by each of the clusters.

Figure 4.5.1 to the right represents the lyrics from artists in the largest cluster. It's interesting to notice how the cluster 1's wordcloud for popular songs is very consistent to the track's descriptions extracted from Spotify's API. This seems to be a cluster with strong traits of positive and cheerful songs.

Figure 4.5.1. Wordcloud for lyrics found in Cluster 1

Figure 4.5.2 to the right is the wordcloud for Cluster 2. This is the cluster for country music, and the lyrics analysis was again matching the song's traits analysis. Here, we find songs that one can interpret as having a simple vocabulary.

Figure 4.5.2. Wordcloud for lyrics found in Cluster 2

There were not evident patterns from analyzing top tracks in cluster 3 (see figure 4.5.3).

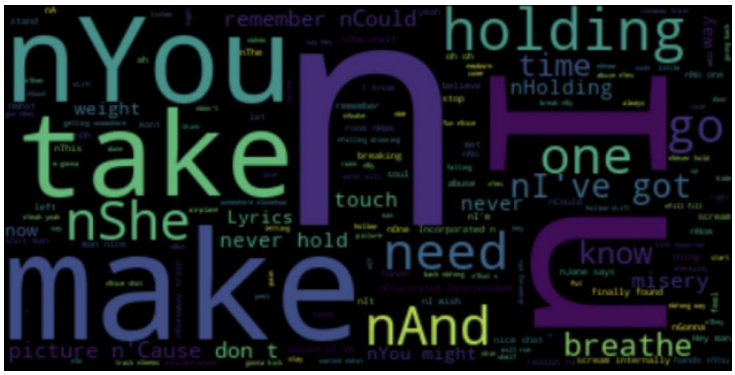

Figure 4.5.3. Wordcloud for lyrics found in Cluster 3

5. Conclusion

Graph analysis is an interesting form of data analysis with applications in many fields to help understand complex relationships among entities in networks.

By the end of this project, we were able to collect hundreds of observations and implement different algorithms successfully on a combination of real datasets. We were able to analyze the network of top artists by the audio features of their top tracks and developed a simple recommendation system for users.

One of the obstacles during the data collection phase, was that the genius api had a limit of requests for their free tier version. This presented a challenge when analyzing lyrics from all songs involved in the study. Nevertheless, for cluster 1 there was a pleasant grouping of upbeat and happy lyrics.

Track features of top songs vary by the genre, however, we found that across all genres, there was a large cluster of songs that have high valence, energy and tempo - so perhaps a recommendation to artists, singer/song writes to develop tracks that are energetic with positive/happy themes if they want to create hits. For future analysis, perhaps certain track features can be dropped – these require deeper

domain knowledge in music, and but certainly something that can be leveraged to further streamline and simplify the analysis.

All the resources used during data collection, modeling and analysis are open source. Our activities were divided in 4 main groups: Data acquisition and pre-processing, Graph Analysis, DEA, and NLP with all team member carrying similar research and development efforts.

Appendix - Audio Features Data Dictionary

- 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.
- 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.
- Key - The key the track is in. Integers map to pitches using standard Pitch Class notation. E.g. 0 = C, 1 = C♯/D♭, 2 = D, and so on. If no key was detected, the value is -1.
- 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.
- Mode - Mode indicates the modality (major or minor) of a track, the type of scale from which its melodic content is derived. Major is represented by 1 and minor is 0.
- Speechiness - Speechiness detects the presence of spoken words in a track. The more exclusively speech-like the recording (e.g. talk show, audio book, poetry), the closer to 1.0 the attribute value. Values above 0.66 describe tracks that are probably made entirely of spoken words. Values between 0.33 and 0.66 describe tracks that may contain both music and speech, either in sections or layered, including such cases as rap music. Values below 0.33 most likely represent music and other non-speech-like tracks.
- 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.
- 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.
- Valence - A measure from 0.0 to 1.0 describing the musical positiveness conveyed by a track. Tracks with high valence sound more positive (e.g. happy, cheerful, euphoric), while tracks with low valence sound more negative (e.g. sad, depressed, angry).
- 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.
- Time_signature - An estimated time signature. The time signature (meter) is a notational convention to specify how many beats are in each bar (or measure). The time signature ranges from 3 to 7 indicating time signatures of "3/4", to "7/4".

References

1. https://www.livelifemusicok.org/evolution-of-music
2. https://vrs.amsi.org.au/wp-content/uploads/sites/84/2018/04/tobin_south_vrs-report.pdf
3. http://www.diva-portal.org/smash/get/diva2:1570223/FULLTEXT01.pdf
4. https://graphaware.com/blog/engineering/building-and-exploring-music-knowledge-graph.html
5. https://emily.louie.ca/sixdegrees/
6. https://nielsdejong.nl/neo4j%20projects/2020/09/23/spotify-playlist-builder.html
7. https://www.spotify.com/au/about-us/contact/
8. https://www.nytimes.com/2022/01/26/arts/music/spotify-neil-young-joe-rogan.html
9. https://www.riaa.com/
10. Vecchio F, Miraglia F, Piludu F, et al. "Small world" architecture in brain connectivity and hippocampal volume in Alzheimer's disease: a study via graph theory from EEG data. *Brain Imaging Behav*. 2017;11(2):473-485. [PubMed] [Google Scholar]
11. F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor, Eds., Recommender Systems Handbook. Boston, MA: Springer US, 2011. doi: 10.1007/978-0-387-85820-3.
12. "A global geometric framework for nonlinear dimensionality reduction" Tenenbaum, J.B.; De Silva, V.; & Langford, J.C. Science 290 (5500)
13. *Cioabă, Sebastian M. (2011), "Some applications of eigenvalues of graphs", in Dehmer, Matthias (ed.), Structural Analysis of Complex Networks, New York: Birkhäuser/Springer, pp. 357–379, doi:10.1007/978-0-8176-4789-6_14, MR 2777924*; see proof of Lemma 5, p. 361

GT ids:

- Joe Amedeo: 903549683
- Kenia Way: 903549619
- Max Kanaskar: 903660366