Problem Chosen
D

2021
MCM/ICM
Summary Sheet

Team Control Number
2113454

The Analysis of Music Influence

Summary

In this paper, an analysis is conducted regarding the influence of music using the given data. Each music shows its unique characteristics, which explains why the types of music are diversified and enriched around the world. In order to find out the difference, relation and evolution of influence exerted by music, some networks and models were constructed to analyze the given data. The problem was broken down into seven tasks to be solved in steps.

For task1, we created a complex network based on directed graph so that we can establish the contact between all the influencers and followers. Afterwards, we calculated the outdegree and indegree of all the nodes. In order to analyze the music influence in our network, we built a **N-dimensional weighted distance model(NDWDM)**.Using this model, we can obtain the $score$, which represents the measurement of the influence degree in the network.

In task2, we used the concept of "centroid" and built a **Centroid of N-dimensional Influence Model(CNDIM)"**. Subsequently, we calculated the **Pearson Correlation Coefficient** between genres and artists and used the matrix of coefficient to draw a **Artists Similarity Heatmap(ASH)**.

To solve task3, firstly, we further discussed the similarity among genres using **K-Means Algorithm**. Afterwards, we analyzed the influence and relation between genres by building a **Genres Influence Heatmap(GIH)**. Then we uses **Adaboost Algorithm** with **Decision Tree** as its classifier for the task of classifying music's genre. We can retrieve the importance of each characteristic for distinguishing music from the model. Next, we used the methods of **Time series analysis(TSA)** and built an **Auto regressive Moving Average model(ARMA)** to analyze how genres change over time and make a prediction for the changes of genres in 2021-2026.

For task4, we utilized **N-dimensional weighted distance model(NDWDM)** to compare the distance within influencer and follower and follower to the others so as to analyze whether influencer did influence the follower. Meanwhile, we analyzed the distance of different characteristics within influencers and followers to see if some characteristics are more contagious.

In task5, to find out the revolution in music, we applied the **entropy weight method(EWM)** to analyze the difference degree of each characteristic of music per year. Afterwards, we can find out the revolution year by finding the max difference between the score calculated by weighted characteristic within adjacent year.Next, we calculated the revolution score of each music by analyzing the revolution score of the year and the distance between the song and the revolution's stream. The revolution score of an artist is the score of all the music he created.

To solve task6, by using the **entropy weight method(EWM)**, we found out which characteristic is more likely to reflect the change within and between music genres. Afterward, we made use of these characteristics to analyze how the genres change over time.

For the final task7, we found that the World War II and the improvement of technology after the war had a profound impact on the development of music.

Keywords: music influence, network, directed graph, centroid, distance, weight

Contents

1 Introduction

1.1 Background

As one of the splendid cultures created by human beings, music has played a significant part in the development of human society [1]. There are various influencing factors for artists when they write a new song, including the people who exerted influence on them, the genre which they belonged to, the impact of current social and political events, the appearance of newly invented instruments and tools, or their personal experiences. Due to the data showed in $influence\_data.csv$ and $full\_music\_data.csv$, the relationship between the artists and the the characteristics of their music works can be figured out.

Since the network science is a subject which studies the qualitative and quantitative relationship between complex objects, the relevant knowledge of network science could be exercised to capture the influences of these music artists with a network constructed to mark their influences relationship and to investigate the similar characteristics of their musical works.

When there is a revolution occurring to musical history, for example, the emergence of a new genre or the reinvention of a genre, there could be various and trivial reasons. To analyze the reasons, it is better to construct some models based on math for best fitting the data and reflecting the influence of music.

1.2 Restatement of the problem

To explore the difference, relation and evolution of music, the following problems were solved:

Firstly, it is necessary to create one or more directed networks of musical influences based on known data sets, thus connecting influencers to followers. Then, a subnetwork was built to explain what "musical influence" measures reveal.

Next, we need to develop our measure with consideration given to the similarity of artists within genres. Then, it is necessary to distinguish between different genres while understanding the changes over time and potential influences between the genres.

Based on the analysis and data as mentioned above, it is necessary to verify whether the identified influencers have genuine impact on the music created by followers, and to find out which characteristics are more "contagious" than others. From the data, it is necessary to identify the characteristics of the revolution that signify the evolution of music, and to find the artists who represents the revolutionaries in the network.

Finally, there is a necessity to analyze the influenctial process of musical evolution, find the indicators of the dynamic influences in a genre, and explain the changes occurring to genres and artists over time. In addition to the influences of music itself, we are also supposed to identify the influences of social, political or technological changes experienced by the network.

2 General Assumptions and Variable

2.1 Assumptions

- The networks are directed graph without circles.
- The followers will not influence the influencers.

- The data given can reflect the actual situation objectively and comprehensively.
- The given data include all the characteristics of music we need for analysis.

2.2 Variable Description

Table 1: Notation

Symbol	Definition
t	Time(year)
μ	The average value of a feature
σ	Standard deviation
$centroid$	The center point in space considering a set of data

3 Model Establishment and Solutions

3.1 Task1

3.1.1 Model Establishment and Solving

We built a directed network of music influences with weights using **influencer_id** and **follower_id** in *influence_data.csv*. Since the range of ID is [0,4000000] but there are only about 6000 artists, we performed discrete operation on ID and mapped it to [0,6000] in order to have higer space performances. In the process of edge construction, weights are given to each edge to measure the influence degree from father to children. The greater the influence of the influencer on the followers, the more similar the followers and the influencers should be. The similarity between two artist can be reflected by there distance in space using *artist_data.csv*. For each node, we calculated its four attributes: **Influential_Score**, **Outdegree**, **Be_Affected_Score** and **Indegree**. These attributes can be calculated during the process of building graph. The model has the advantages of efficiency in space utilization and calculation.

In order to analyze "music influence" in our network, we built a **N-dimensional weighted distance model(NDWDM)**, where N is the number of different kind of music characteristics. In this model, all the influential factors listed in the *data_by_artists.csv* contributes to the weight on edge.

Firstly, we normalized all the factors:

$$x_{Standardization} = \frac{x-\mu}{\sigma} \tag{1}$$

Next, we calculated the *distance* between each connected point in N-dimensional space where *a*,*b* indicated two separated artists and *n* indicated the number of the music characteristic which would be considered:

$$distance = \sqrt{(a_1-b_1)^2+(a_2-b_2)^2+...+(a_n-b_n)^2} \tag{2}$$

According to the network we had built and the distance we had calculated, we considered that weight of the edge between two artists should be large if one had a great influence on the other. In our network, the shorter distance between two nodes indicates the greater influence. In the model, the weight is calculated as:

$$weight = \frac{1}{distance} \tag{3}$$

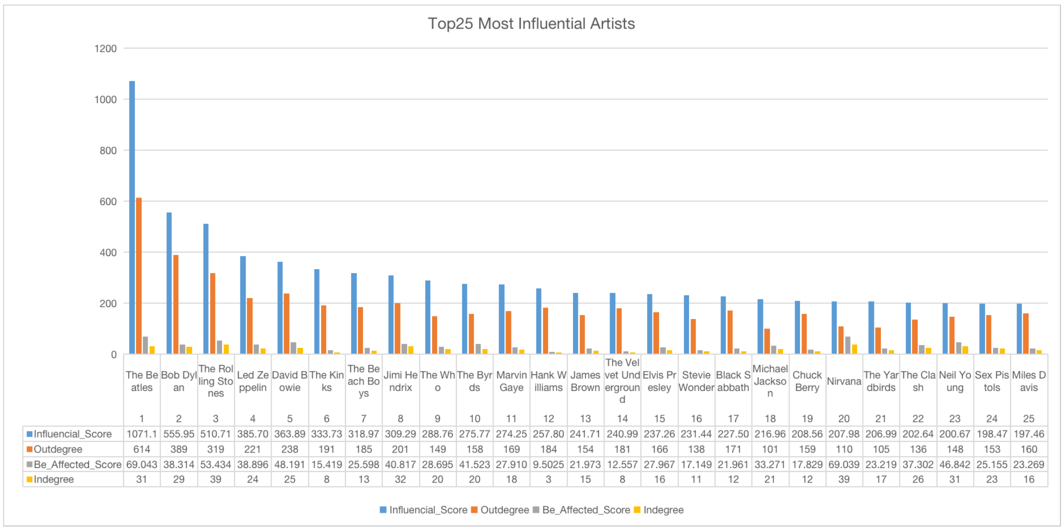
Then we could add up all the weight of each edge which directed out of a vertex to get a final score. The larger the *score* is, the greater the influence this artist had on others:

$$score = weight_1 + weight_2 + ... + weight_n \tag{4}$$

Where *n* indicates the number of edges directed out of a vertex in a directed graph.

3.1.2 Analysis and Evaluation of results

After calculated by our network, we got a rank of the top 25 most influential artists in our model showed in Fig 1:



	The Beatles	Bob Dylan	The Rolling Stones	Led Zeppelin	David Bowie	The Kinks	The Beach Boys	Jimi Hendrix	The Who	The Byrds	Marvin Gaye	Hank Williams	James Brown
	1	2	3	4	5	6	7	8	9	10	11	12	13
■ Influencial_Score	1071.1	555.95	510.71	385.70	363.89	333.73	318.97	309.29	288.76	275.77	274.25	257.80	241.71
■ Outdegree	614	389	319	221	238	191	185	201	149	158	169	184	154
■ Be_Affected_Score	69.043	38.314	53.434	38.896	48.191	15.419	25.598	40.817	28.695	41.523	27.910	9.5025	21.973
■ Indegree	31	29	39	24	25	8	13	32	20	20	18	3	15

	The Velvet Underground	Elvis Presley	Stevie Wonder	Black Sabbath	Michael Jackson	Chuck Berry	Nirvana	The Yardbirds	The Clash	Neil Young	Sex Pistols	Miles Davis
	14	15	16	17	18	19	20	21	22	23	24	25
■ Influencial_Score	240.99	237.26	231.44	227.50	216.96	208.56	207.98	206.99	202.64	200.67	198.47	197.46
■ Outdegree	181	166	138	171	101	159	110	105	136	148	153	160
■ Be_Affected_Score	12.557	27.967	17.149	21.961	33.271	17.829	69.039	23.219	37.302	46.842	25.155	23.269
■ Indegree	8	16	11	12	21	12	39	17	26	31	23	16

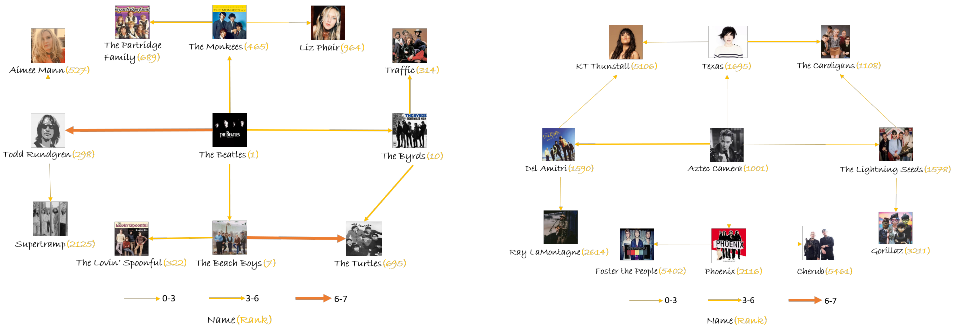
Figure 1: Top 25 Most Influential Artists

Influential_Score and **Be_Affected_Score** are the score calculated by the formula mentioned above. **Outdegree** and **Indegree** are provided by the network we built. From this figure, we could clearly tell that The Beatles is the most influential artist in music history.

For more details, we established two subnetwork and did a deeper analysis. Fig 2(a) is the subnetwork of The Beatles and its two-level child nodes.

Fig 2(b) is the subnetwork of Aztec Camera, an artist ranked at 1001 in our network, and its two-level child nodes. From the figures showed bellow, we could get the influence relationships between the fathers and children. The figure also showed the *weight* which is described by the depth of the color. The color is darker and the edge is thicker when the weight is larger. We divided the weight into three ranges 0-3,3-6,6-7 in the figure. Each weight is represented by different color and thickness of the edge.

So, the music influence between the artists in the network had been visualized clearly by our figure with the precise number.



(a) Three-level subnetwork of The Beatles

(b) Three-level subnetwork of The Aztec Camera

Figure 2: Subnetwork

3.2 Task2

3.2.1 Model Establishment and Solving

In this task, we used *data_by_artist.csv* and *influence_data.csv* to get the music characteristics and genres of each artist. Based on this information, we built a **Centroid of N-dimensional Influence Matrix(CNDIM)** to analyze the music characteristics of each genre. N represents the number of different kinds of music characteristics.

First of all, We assumed that there were n_a artists in genre A, n_b artists in genre B... and there were m kinds of music characteristics. We supposed that there was a genre called A and $a_1j, a_2j, ..., a_{n_a}j (j = 1, 2, ..., m)$ indicates the artists in genre A where j represents j-th music characteristics of an artist. Then we could get the "centroid" of each genre.

For genre A, the "centroid" is$(\frac{a_{11}+a_{21}+...+a_{n_a1}}{n_a}, \frac{a_{12}+a_{22}+...+a_{n_a2}}{n_a}, ..., \frac{a_{1m}+a_{2m}+...+a_{n_am}}{n_a})$. And the centroid of each genre can be calculated using the following matrix:

$$\begin{bmatrix} \frac{a_{11}+a_{21}+...+a_{n_a1}}{n_a} & \frac{a_{12}+a_{22}+...+a_{n_a2}}{n_a} & \dots & \frac{a_{1m}+a_{2m}+...+a_{n_am}}{n_a} \\ \frac{b_{11}+b_{21}+...+b_{n_b1}}{n_b} & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ \frac{s_{12}+s_{22}+...+s_{n_s1}}{n_s} & \dots & \dots & \frac{s_{1m}+s_{2m}+...+s_{n_sm}}{n_s} \end{bmatrix}$$

After that, we used **Pearson Correlation Coefficient** r to represent the correlation of artists with the "centroid" of genres:

$$r = \frac{N\sum x_iy_i - \sum x_i \sum y_i}{\sqrt{N\sum x_i^2 - (\sum x_i)^2}\sqrt{N\sum y_i^2 - (\sum y_i)^2} \tag{5}$$

where correlation coefficient r between:

- 0.8-1.0 are strongly correlated
- 0.6-0.8 are strong correlation
- 0.4-0.6 are Moderate correlation
- 0.2-0.4 are weak correlation

- 0.0-0.2 are very weak correlation or no correlation

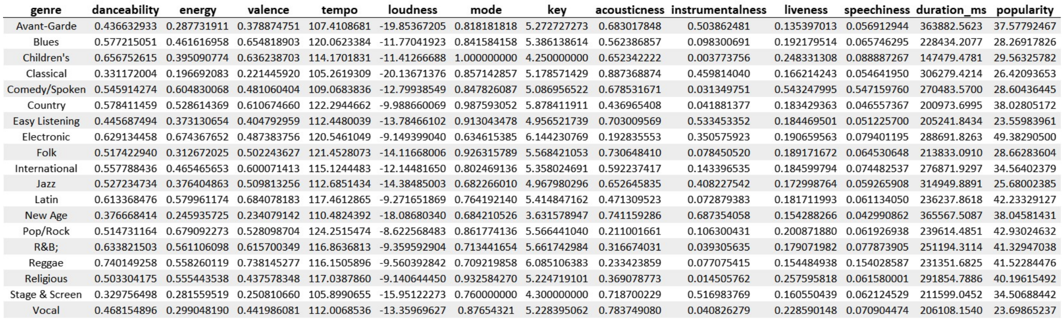
We define m_i as the i-th genre's centroid. $r(x, y)$ to represent the coefficient between x and y, $belong(x)$ as the i-th genre that the x belongs to. *mat* to represent our *Artists Similarity Matrix*, where mat_{ij} indicates the coefficient of all the artists that belong to the i-th genre and the centroid of the j-th genre.

$$mat_{ij} = \sum_{belong(x)=i} r(x, m_j) \tag{6}$$

After the calculation, we normalized each row with *min-max* to map the value into [0,1]. Then, we visualized the data of the coefficient with a **Artists Similarity Heatmap(ASH)**.

3.2.2 Analysis and Evaluation of results

After the calculation and analysis, we got the centroids of all the genres, shown in Fig 3: We ignored the genre "Unknown" because doing this could improve the accuracy of the calculation.



genre	danceability	energy	valence	tempo	loudness	mode	key	acousticness	instrumentalness	liveness	speechiness	duration_ms	popularity
Avant-Garde	0.436632933	0.287731911	0.378874751	107.4108681	-19.85367205	0.818181818	5.272727273	0.683017848	0.503862481	0.135397013	0.056912944	363882.5623	37.57792467
Blues	0.577215051	0.461616958	0.654818903	120.0623384	-11.77041923	0.841584158	5.386138614	0.562386857	0.098300691	0.192179514	0.065746295	228434.2077	28.26917826
Children's	0.656752615	0.395090774	0.636238703	114.1701831	-11.41266688	1.000000000	4.250000000	0.652342222	0.003773756	0.248331308	0.088887267	147479.4781	29.56325782
Classical	0.331172004	0.196692083	0.221445920	105.2619309	-20.13671376	0.857142857	5.178571429	0.887368874	0.459814040	0.166214243	0.054641950	306279.4214	26.42093653
Comedy/Spoken	0.545914274	0.604830068	0.481060404	109.0683836	-12.79938549	0.847826087	5.086956522	0.678531671	0.031349751	0.543247995	0.547159760	270483.5700	28.60436445
Country	0.578411459	0.528614369	0.610674660	122.2944662	-9.988660069	0.987593052	5.878411911	0.436965408	0.041881377	0.183429363	0.046557367	200973.6995	38.02805172
Easy Listening	0.445687494	0.373130654	0.404792959	112.4480039	-13.78466102	0.913043478	4.956521739	0.703009569	0.533453352	0.184469501	0.051225700	205241.8434	23.55983961
Electronic	0.629134458	0.674367652	0.487383756	120.5461049	-9.149399040	0.634615385	6.144230769	0.192835553	0.350575923	0.190659563	0.079401195	288691.8263	49.38290500
Folk	0.517422940	0.312672025	0.502243627	121.4528073	-14.11668006	0.926315789	5.568421053	0.730648410	0.078450520	0.189171672	0.064530648	213833.0910	28.66283604
International	0.557788436	0.465465653	0.600071413	115.1244483	-12.14481650	0.802469136	5.358024691	0.592237417	0.143396535	0.184599794	0.074482537	276871.9297	34.56402379
Jazz	0.527234734	0.376404863	0.509813256	112.6851434	-14.38485003	0.682266010	4.967980296	0.652645835	0.408227542	0.172998764	0.059265908	314949.8891	25.68002385
Latin	0.613368476	0.579961174	0.684078183	117.4612865	-9.271651869	0.764192140	5.414847162	0.471309523	0.072879383	0.181711993	0.061134050	236237.8618	42.23329127
New Age	0.376668414	0.245935725	0.234079142	110.4824392	-18.08680340	0.684210526	3.631578947	0.741159286	0.687354058	0.154288266	0.042990862	365567.5087	38.04581431
Pop/Rock	0.514731164	0.679092273	0.528098704	124.2515474	-8.622568483	0.861774136	5.566441040	0.211001661	0.106300431	0.200871880	0.061926938	239614.4851	42.93024632
R&B;	0.633821503	0.561106098	0.615700349	116.8636813	-9.359592904	0.713441654	5.661742984	0.316674031	0.039305635	0.179071982	0.077873905	251194.3114	41.32947038
Reggae	0.740149258	0.558260119	0.738145277	116.1505896	-9.560392842	0.709219858	6.085106383	0.233423859	0.077075415	0.154484938	0.154028587	231351.6825	41.52284476
Religious	0.503304175	0.555443538	0.437578348	117.0387860	-9.140644450	0.932584270	5.224719101	0.369078773	0.014505762	0.257595818	0.061580001	291854.7886	40.19615492
Stage & Screen	0.329756498	0.281559519	0.250810660	105.8990655	-15.95122273	0.760000000	4.300000000	0.718700229	0.516983769	0.160550439	0.062124529	211599.0452	34.50688442
Vocal	0.468154896	0.299048190	0.441986081	112.0068536	-13.35969627	0.87654321	5.228395062	0.783749080	0.040826279	0.228590148	0.070904474	206108.1540	23.69865237

Figure 3: Centroid of N-dimensional Influence Matrix

Then, we drew the heatmap of the music similarity in Fig 4 and compared the coefficient between genres. If the coefficient is close to 1, we could tell that the artists within this genre(showed in the right side of the figure)are high related to the genres(showed at the bottom of the figure). We took row 4,column 2 as an example. In this node, the coefficient is 0.67 which indicates that the average coefficient of all the artists in "Classical" and "Blues" is strong. But it is still less than the coefficient of "Classical" and itself. So, after we analysed all the situation showed in the figure, we can draw a conclusion that all genres are obviously highly related to itself. There are more conclusions that come from Fig 4 in task3 when we further discuss the similarity among genres.

3.3 Task3

3.3.1 Model Establishment and Solving

For Task3, we mainly measured the similarities and influences between and within genres and analyzed the development of genres.

To analyze the similarity among genres, we calculated the centroid for each genre which is shown in The centroid is calculated with the artists that belong to the genre. Also, after acquiring each centroid, we could cluster genres with **K-Means Algorithm** to form four clusters [3]. The genres in the same cluster are highly similar.

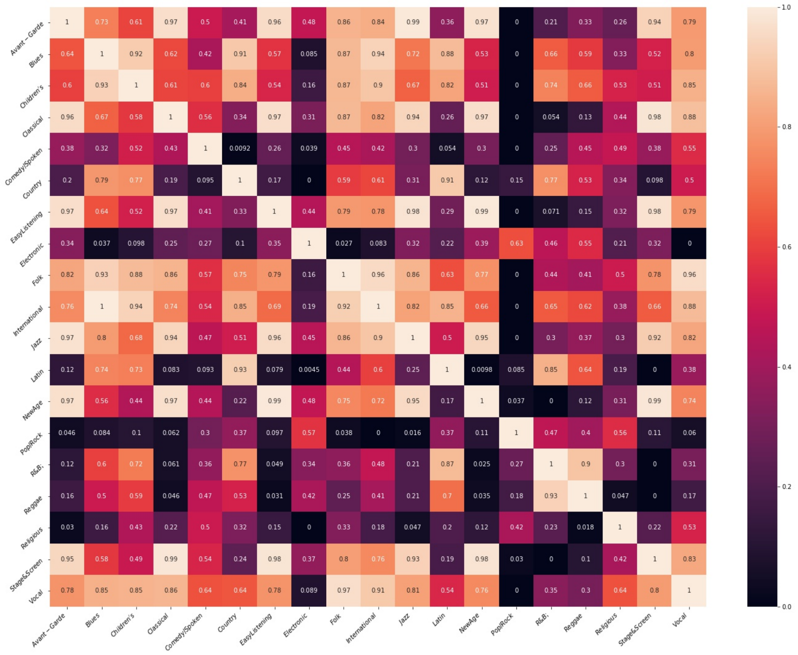


Figure 4: Artists Similarity Heatmap(ASH)

To analyze the influence between and within genres, we utilized the data in *influence_-data.csv* and drew a line from the genre that the influencer belongs, to the genres t hat the follower belongs. We defined $matf$ as the influence matrix.

$$matf_{ij} = \sum_{belong(influencer)=i \ \&\& \ belong(follower)=j} 1 \tag{7}$$

The matrix reveals the impact of one genre had on others.

To distinguish a genre, we utilized machine learning models. We performed the task of classifying each music within the dataset *full_music_data.csv*. Each music is labeled by its artist and the label of the artist could be found in *influence_data.csv*. However, since the song that were written by more than one artists from different genres is not able to be labeled, we dropped this kind of music which is a small part of the data. We adopted **Adaboost Algorithm** and took two hundreds **Decision Tree** as its classifiers [4, 5].

To analyze how genres change over time, we group all the music by its genre and year. We calculated how each genre changes every 5 years. The current circumstances of a genre in 5 years could be inferred by its songs in the 5 years. Therefore, for each genre, we calculated the centroid of it every 5 years with the songs within. Afterwards, we took 9 genres as examples and visualized the result using line chart to show and analyze how genres change over time. The visualization is shown in fig.a and fig.b .

We took "popularity" as an example which is an important factor of music influence and used the methods of **Time series analysis(TSA)** to analyze a specific characteristic and make a prediction [6]. That's because **TSA** highlight the role of time factors in predicting which could help us make the predictions more accurately. We randomly picked up 2 genres of music to built the **Auto regressive Integrated Moving Average model(ARIMA)** or **Auto regressive moving average model(ARMA)** and analyzed which model best fitted our data [6]. We took the interval of 2 years, 3 years, 4 years, 5 years, 10 years as the units and analyzed them separately. In particular, we separated the last two units into **testing set** and the others into **training set**. In other words, we used the training set to obtain two predicted values of the last two units so that we could compare them with the actual values of the last two units and calculate the $Error$:

$$Error = |\frac{ActualValue - PredictedValue}{ActualValue}| \times 100\% \tag{8}$$

In this way, we can tell whether the predicted value in the future is reliable or not, and how much is the accuracy. Our specific steps to do the analysis is showed in the following:

1) Select 2 genres to analyze the change of "popularity" randomly.

2) Choose the interval of 5 years as a unit.

3) Take the last two points(units) into testing set whose abscissa 2016(the average "popularity" of 2011-2016) and 2021(the average "popularity" of 2016-2021).

4) Calculate the average $Error$ of two units in step 3.

5) Erase the testing set and turn them into training set.

6) Predict the values of "popularity" in the next five years and tell how much may be the $Error$ of the predictions.

3.3.2 Analysis and Evaluation of results

In Fig 3, we could find characteristic with similar value among different genres such as the loudness of Blues an Country. Similar conclusions could be draw from the figure. Afterwards, using **K-Means Algorithm**, the genres are clustered into four groups. The result is shown as Fig 5 and Table 2. Let's refer to Fig 4 once again. It's obvious that the genres in the same cluster are also highly related. The similar result of different methods realize mutual verification. The **K-Means** shows only the relationship in a cluster and the **ASH** shows the degree of coefficient between every genres.

With influence matrix *matf*, we built the **Genres Influence Heatmap(GIH)** in Fig6. The larger the weight of $matf_{ij}$ is, the more impact the i-th genre had on the j-th genre. From

Table 2: Influence value

Cluster	Genres
$Cluster1$	Comedy/Spoken
$Cluster2$	Classical, Avant-Grade, New Age, Stage&Screen
$Cluster3$	Pop/Rock, R&B, Country, Electronic,Latin, Reggae, Religious
$Cluster4$	Blues, Jazz, Folk, Vocal, Children's, Easy Listening, International

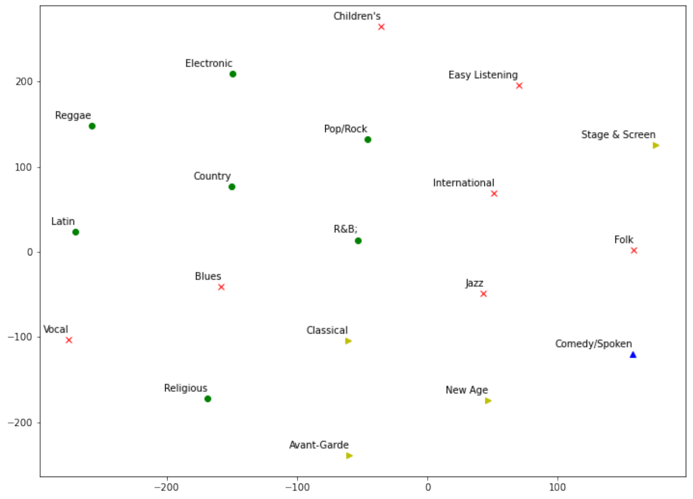


Figure 5: Classify Figure

the figure, we can tell most genre has a great influence on itself and especially Pop/Rock. On the other hand, Pop/Rock is a great influencer to many genres besides itself. However simultaneously, it is also being affected by a great deal, the greatest impact within different genres is R&B to Pop/Rock. Other conclusion can be drawn from the heatmap similarly.

The data used in Adaboost is splitted into training dataset and testing dataset in a proportion of 0.7 and 0.3 . The result of our model is as follows.

$$Accuracy = 0.6728587780018532$$

We didn't perform operation such as K-Fold, the accuracy can reach above 0.7 with several optimizations.

We can get the importance degree of each characteristics from **Adaboost Model(AM)**. The importance degree indicates the importance of the characteristic in classifying music. Therefore, we can distinguish each genre with the characteristic that has large importance. The importance of each characteristic is shown in Fig 7. We could draw a conclusion that the duration and the acousticness of a song is most significant in classifying music, while explicit, key and mode of a song are useless.

Fig 8 and Fig 9 is the visualization of how genres change over Time. As we needed only the changing rate of each characteristic, we conducted manipulation on some characteristics before they are shown for better visualization of small values. We divided the number of *popularity* by 60, the number of *tempo* by 100, the number of *duration_ms* by 200000, the number of *cnt* by 1000. With visualization, we could see that the popularity of each genre is growing steadily and they all had a time from 1940s-1980s when the number of new works belonging to that genre boosted. Also there is a big drop on most characteristics of Country music in 1944 which may be caused by the end of World War 2.

Next, we analyzed how do genres change over time. We took the popularity in "Classical" and "International" as the examples. The results of the two genres are shown in Fig 10. In Fig

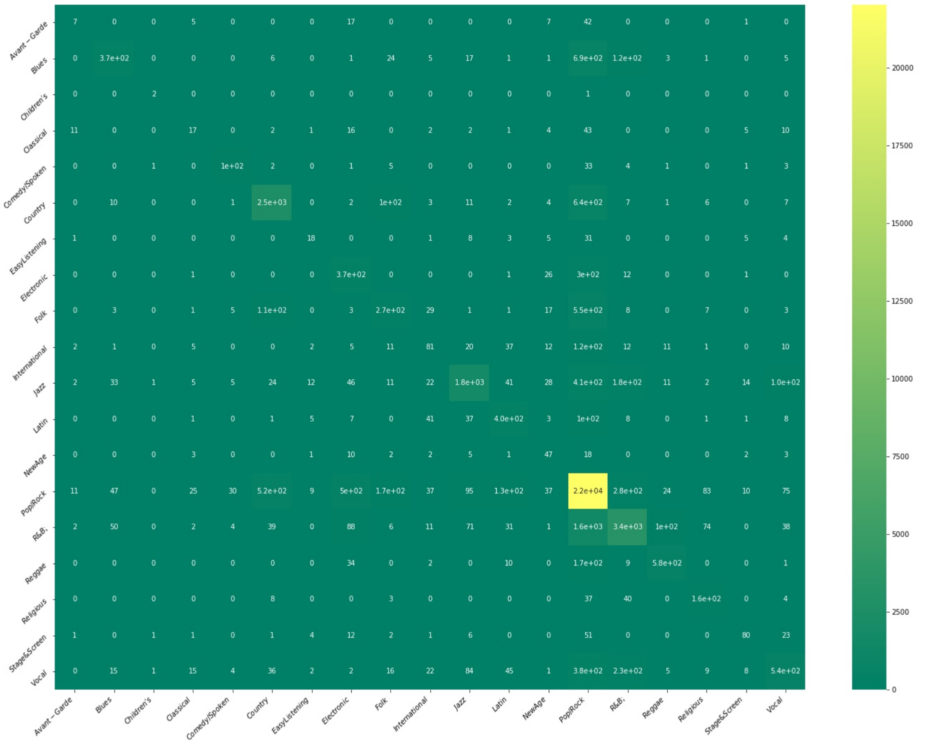


Figure 6: Influence Heatmap

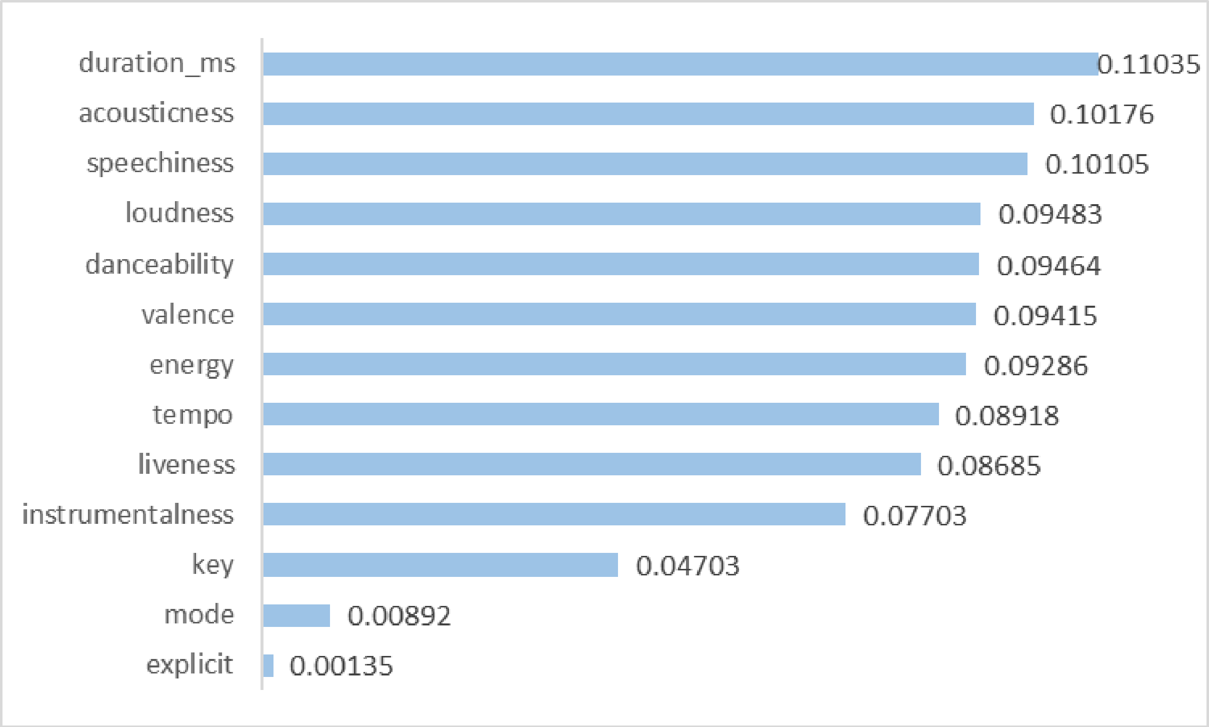


Figure 7: Rank Of Characteristic

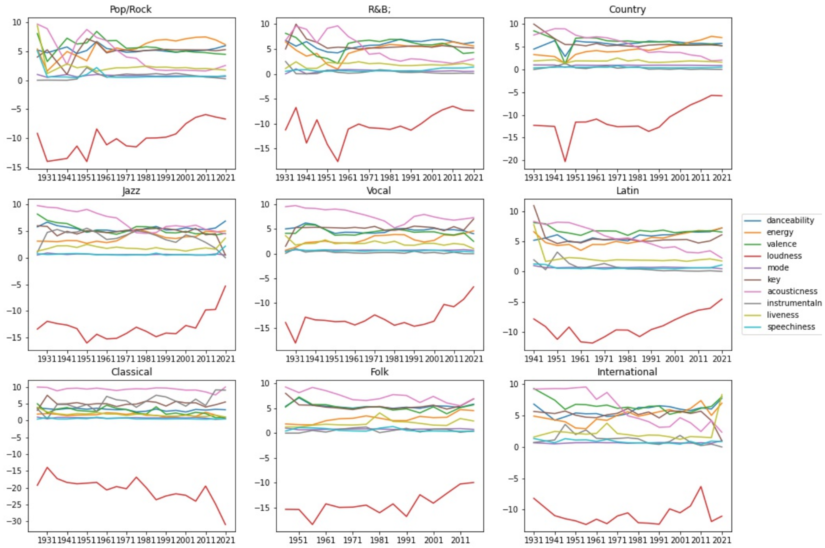


Figure 8: Line Chart1

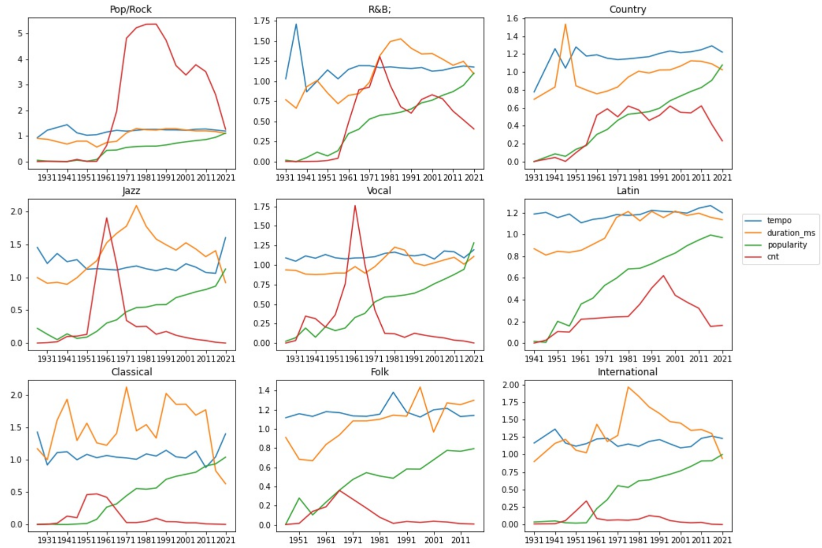
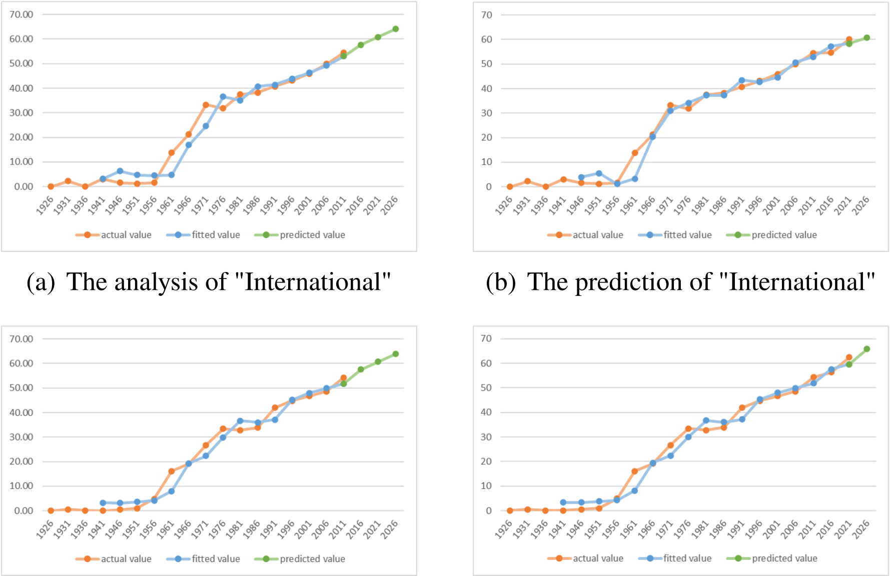


Figure 9: Line Chart2



(a) The analysis of "International"

(b) The prediction of "International"

(c) The analysis of "Classical"

(d) The prediction of "Classical"

Figure 10: The analysis and predictions of "popularity" of two genres

10(a) and (c) , we chosen the interval of 5 years as a unit, so we took 2011-2016 and 2016-2021 into a testing set to calculated the $Error$. For "International" the $Error$ we got was 3.51%, and for "Classical" the $Error$ was 2.474% which were both totally acceptable. Then, we erased the testing set and turned them into training set in order to predict the average values of 2021-2026. The results are showed in Fig 10(b) and (d). The optimal model of "International" was found to be: ARMA(1,1,2), and its model formula is showed in the following:

$$y(t) = 3.001 - 0.535 * y(\frac{t-1921}{5} - 1) + 0.895 * \varepsilon(\frac{t-1921}{5} - 1) + 1.000 * \varepsilon(\frac{t-1921}{5} - 2) \quad (9)$$

We predicted that in 2021-2026 the average value of "popularity" will be 60.668 and the $Error$ was 2.474%. So the predicted value would be between 59.203 and 62.207.

The optimal model of "Classical" was found to be: ARMA(0,1,0), and its model formula is showed in the following:

$$y(t) = 3.289 * \frac{t-1921}{5} \quad (10)$$

We predicted that in 2021-2026 the average value of "popularity" will be 65.789 and the $Error$ was 3.51%. So the predicted value would be between 63.558 and 68.182.

3.4 Task4

3.4.1 Model Establishment and Solving

In order to analyze that the identified influencers in fact influence the respective artists, we used the **N-dimensional weighted distance model(NDWDM)** built above. We used the formula 1 to normalize each features.

Then, by using formula 2, we calculated the distance between a specific influencer and his follower and added it up together. We named the value as $distance_{if}$. Also, for a specific influencer and follower, we calculated the average distance between the follower and other artists beside the influencer and the follower itself. We named the value as $distance_{fo}$. We define

$dis(x, y)$ as the distance between x and y.

$$distance_{if} = dis(influencer, follower) \tag{11}$$

$$distance_{fo} = \frac{\sum_{j!=influencer\&\&j!=follower} dis(follower, j)}{n-2} \tag{12}$$

Then we performed the manipulation on every influencer and follower pair and receive the sum of $distance_{if}$ and $distance_{fo}$ as $total\_distance_{if}$ and $total\_distance_{fo}$. We could compare the number of these two values where the smaller one represents the greater influence.

In order to analyze which characteristic is more "contagious" than others, we calculated the distance within the music characteristic separately for every influencer and follower pair. The results indicate the contagious degree of the characteristic. The shorter the distance is, the more similar the influencer and follower are on the characteristic, inferring that the characteristic is highly contagious. After that we calculated the weight of the distance of each characteristic using the formula 3. Then we used ***Topsis*** to evaluate the weight of each character [2]:

$$x_{normalization} = \frac{x_i - x_{min}}{x_{max} - x_{min}} \tag{13}$$

At last, we ranked the "contagious" of music characteristic.

3.4.2 Analysis and Evaluation of results

$total\_distance_{if}$ and $total\_distance_{fo}$ are showed in table2. As 35661 is much smaller than 51812, we can tell that the influencers indeed influenced the followers.

The rank of the "contagious" of music characteristic is showed in Fig 11. So from the rank showed in the figure, we found that the music characteristic *speechiness* is highly contagious while key has the lowest contagious degree. The conclusion is consistent with our knowledge because the key of a singer is highly determined by his talent.

Finally we could conclude that some music characteristics more "contagious" than others.

3.5 Task5

3.5.1 Model Establishment and Solving

For task5, in order to analyze which characteristics might signify revolutions.Firstly, we used **The entropy weight method(EWM)** to calculate the weight of each characteristic in order to determine how much does the music change in a year and also to realize how much should a specific characteristics contributes to the revolutions [9]. Afterwards, we normalized the values of each characteristic in music and calculated the the centroid of each year with songs within. The definition and calculation of centroid is stated above. The centroid is represented as $\overline{x_1}, \overline{x_2}, \overline{x_3}, ..., \overline{x_n}$. The character of n represents the number of the characteristics we considered.

Table 3: Influence value

Name	Value
$total\_distance_{if}$	35661.670131
$total\_distance_{fo}$	51812.009625

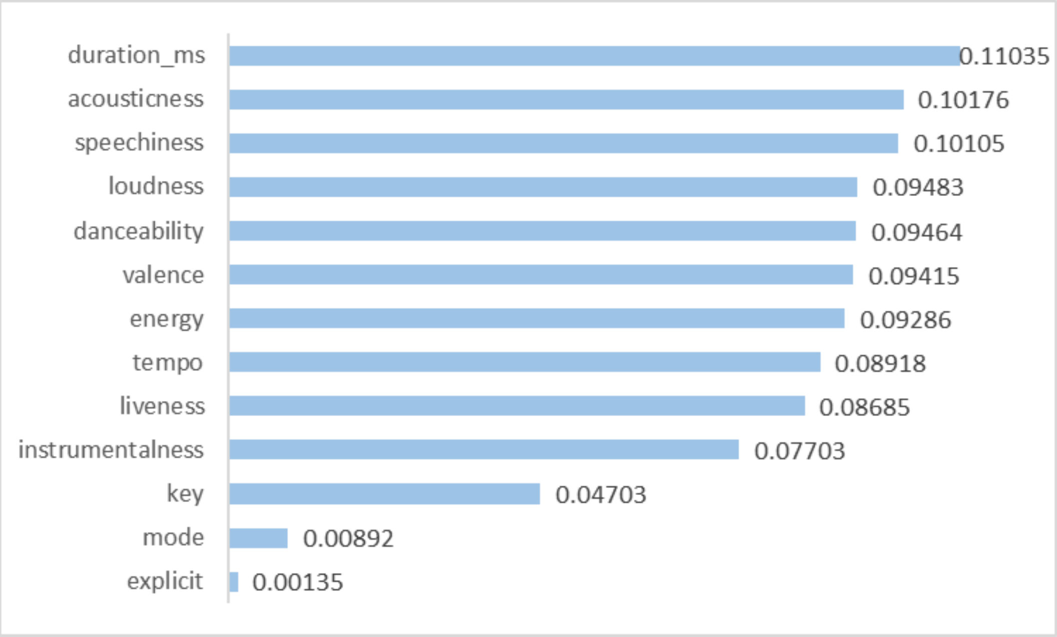


Figure 11: The rank of the "contagious" of music characteristic

Supposed that $w_1, w_2, w_3, w_4, ..., w_n$ represented the weight of each characteristic we calculated using **EWM** and $\sum_{i=1}^{n} w_i = 1$. We defined $S_c(i)$, which represents the score of all the weighted characteristic in year i. The calculation of $S_c(i)$ is in the following:

$$S_c(i) = \sum_{i=1}^{n} \overline{x_i} * w_i \tag{14}$$

We took $|S_c(i-1) - S_c(i)|$ as the value to indicate the difference for the i-th year and years from 1935 to 2020 as the domain to built a coordinate system. We removed the values before 1935 because the songs before were too rare and it would bring up high $noises$. In this coordinate system, the range of y represents the changing degree of each year. The higher $|S_c(i-1) - S_c(i)|$ indicates the bigger change of music in this year.

In order to find artists that can represent revolutionaries in our network, we defined $S_a(j)$ as the score of the artist in a specific year, where j represents the j-th music. We assumed c_i represents the i-th characteristic of the centroid of this year, where the centroid represents the mainstream of the music in this specific year. The distance between the music and the "centroid" indicates the relevance and influence the music has on the mainstream. The shorter distance between the "centroid" and the music means the more involved for the music in the revolution, and would therefore contribute more scores to its artists. However, different characteristics with different weights should contribute differently to the distance. The more weight a characteristic has, the more related the characteristic and the score should be. Also, the score of music should correlate positively with the the revolution score of the year. Therefore, the score of each music could be calculated as follows, where we define *cur_year* as current year.

$$S_a(j) = \frac{|S_c(cur\_year) - S_c(cur\_year - 1)|}{\sqrt{\sum_{i=1}^{n}(x_i - c_i)^2 * w_i}} \tag{15}$$

The score of an artist is the total score of every music he created.

3.5.2 Analysis and Evaluation of results

The results of analysis using **EWM** are showed in Table3. There were 14 characteristics we considered in our model. Using above formulas, we could retrieve a line chart indicating the changing degree of each year. We found that there were big changes in the years of 1936,1944,1945, 1954,1956, 2020 which means revolutionaries occurred in these years. In our research, 1944 and 1945 are the ending years of World War 2. In 2020, we have just experienced severe epidemic.

We took the year of 2020 as example and calculated the score of all the artists who released songs in this year. We drew a **radar map** for visualization to find out the most revolutionary artist in 2020. In Fig 13. The numbers shown at the bottom are the range of score.

Each artist has been marked as a black point in the map and we labeled the names and score of the top 10 artists who influenced the trend most. The artist with higher score will locate closer to the center. From the map, we can tell that Kygo, The weekend , Dua Lipa have top 3 greatest contribution to the revolution in 2020.

3.6 Task6

3.6.1 Model Establishment and Solving

In task6, we analyzed the influence processes of musical evolution that occurred over time in some genres. First of all, we still used **EWM** to obtain the weight of each characteristic in genres. According to the definition of **EWM**, we can say that the greater the weight, the greater the variation of the characteristic over the years. So we mainly analyzed these characteristic because they are more likely to reflect the change of music genres. Afterwards, we combined the weight with figure 8 and 9 so as to find out which and how characteristics lead to the evolution of genres. We selected three genres and the average value of all the genres as 4 objects to draw a histogram. Then we made use of the data and trend showed in the Table 5 to explain how the characteristic affected the change of genres over time.

Table 4: The weight of each characteristic

Characteristic name	Information entropy	Information utility value	Weight
danceability	0.9989	0.0011	0.53%
energy	0.9864	0.0136	6.39%
valence	0.9985	0.0015	0.72%
tempo	0.9998	0.0002	0.10%
loudness	0.9674	0.0326	15.29%
mode	0.9995	0.0005	0.24%
key	0.9999	0.0001	0.06%
acousticness	0.9652	0.0348	16.32%
instrumentalness	0.9574	0.0426	20.00%
liveness	0.9989	0.0011	0.52%
speechiness	0.9965	0.0035	1.62%
duration_ms	0.9987	0.0013	0.63%
popularity	0.9544	0.0456	21.44%
total_cnt	0.9656	0.0344	16.17%

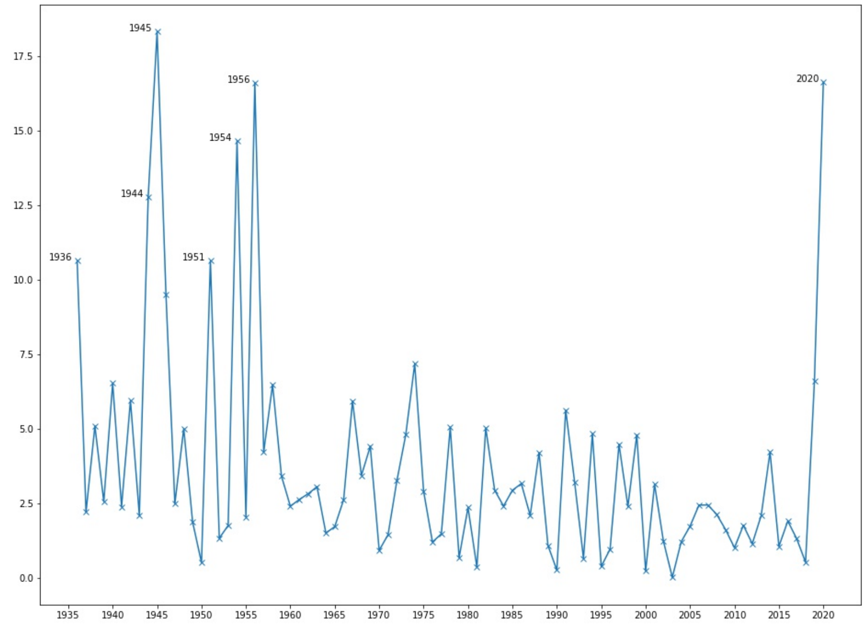


Figure 12: The rate of change of the music characteristic

3.6.2 Analysis and Evaluation of results

To analyze the change within a genre, we used three genres as examples for analysis. The histogram is showed in Fig 14. In this figure, y ordinate represents the percentage of the number of weight. From the figure, we can draw the following conclusions. For Pop/Rock, the top 3 greatest characteristics of music influence are "speechiness", "popularity", "count". For Jazz, they are "loudness", "popularity", and "count". And for Classical, they are "instrumentalness", "popularity", and "count".Taken as a whole, we found that the weight of "danceability" and "tempo" are small which indicated that they are rarely contributed to the change of the genres.

In order to analyze the change between genres, we drew a table of the "count" of the three genres from the year of 1926 to 2021 showed in 5. From the table we found that Pop/Rock genre has a big number of pop songs released in recent years. The "golden years" of Jazz was in the decade of 1956-1966. For Classcial, we can see that average songs released each years during 1951-1966 is more than the others. Overall, it showed a downward trend. At last, we can concluded that each kind of genre has its own "The golden age". Due to more "speechiness", "popularity" and less "loudness" Pop/Rock brings, more and more generations prefer this kind of music.

3.7 Task7

In figure 8 and 9, we can see that the number of creations of most genres increased significantly after 1945. The World War II broke out from 1939 to 1945, this war brought great disaster to the people of the world, and the political pattern of the world changed fundamentally. All these are bound to have a profound impact on people's thoughts and psychology, and must be reflected in the music creation. Whether it is the pain brought by the war or the joy of people

Table 5: The "count" of the three genres

Year/Rock	Pop/Rock	Jazz	Classcial
1926	1	2	2
1931	8	8	4
1936	0	21	20
1941	1	101	130
1946	83	107	105
1951	6	134	462
1956	8	1098	474
1961	630	1902	422
1966	1952	1200	224
1971	4815	344	30
1976	5222	250	30
1981	5358	256	49
1986	5363	135	96
1991	4730	178	45
1996	3745	120	43
2001	3377	85	26
2006	3778	58	26
2011	3509	40	10
2016	2608	14	6
2021	1260	2	2

Table 6: The "count" of three genres from 1926-2021

Year	Pop/Rock	Information utility value	Weight
danceability	0.9989	0.0011	0.53%
energy	0.9864	0.0136	6.39%
valence	0.9985	0.0015	0.72%
tempo	0.9998	0.0002	0.10%
loudness	0.9674	0.0326	15.29%
mode	0.9995	0.0005	0.24%
key	0.9999	0.0001	0.06%
acousticness	0.9652	0.0348	16.32%
instrumentalness	0.9574	0.0426	20.00%
liveness	0.9989	0.0011	0.52%
speechiness	0.9965	0.0035	1.62%
duration_ms	0.9987	0.0013	0.63%
popularity	0.9544	0.0456	21.44%
total_cnt	0.9656	0.0344	16.17%

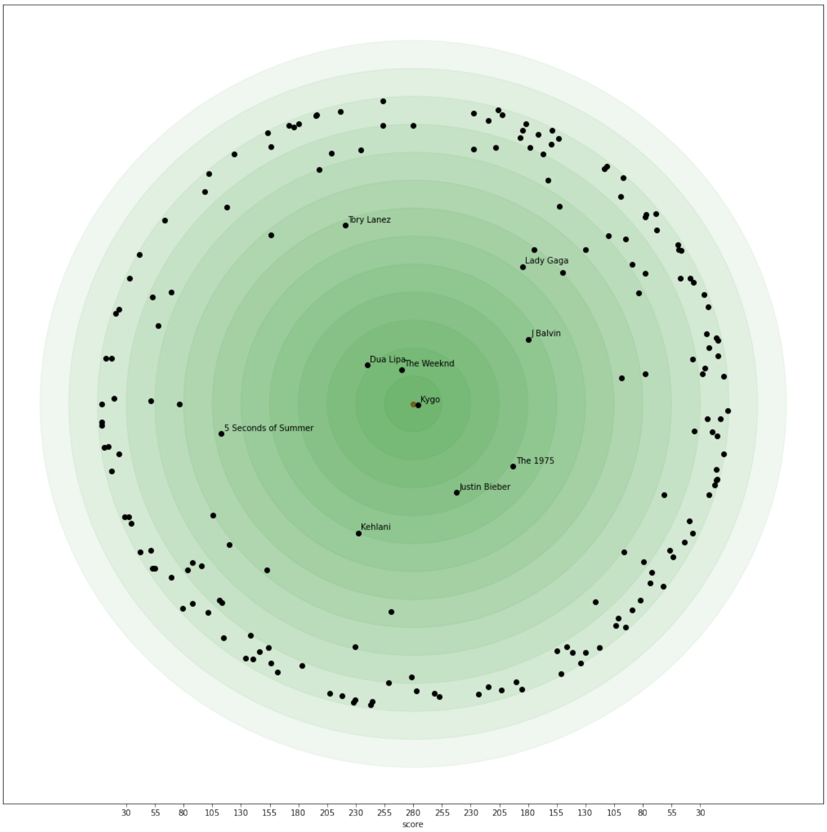


Figure 13: The Radar Map of the influence of the artists have on revolutionary in 2020

in the peace period after the war, it has greatly stimulated the artists' creation.

The World War II objectively promoted the development of science and technology. The application of new technology and scientific had a great impact on the creativity of musicians and the creation of musical works [10]. Technology has greatly improved the speed and accessibility of communication, which helps to form the similarity and unity of music taste.

The recording and broadcasting industries have developed to a point where long musical works can be recorded without interruptions, recordings of any and all types of music are readily available and music is accessible everywhere. Electronic instruments are easier to use and help artists explore new timbres. Tone manipulation provides composers with new choices. When the complexity of rhythm exceeds people's ability, it can be solved by computer technology. The overall noise level of the environment is much higher than that of previous times, and some types of music (Pop / Rock) have higher decibel levels.

In a word, the changes of society and science and technology have stimulated the creation of music and have a great impact on the development of music.

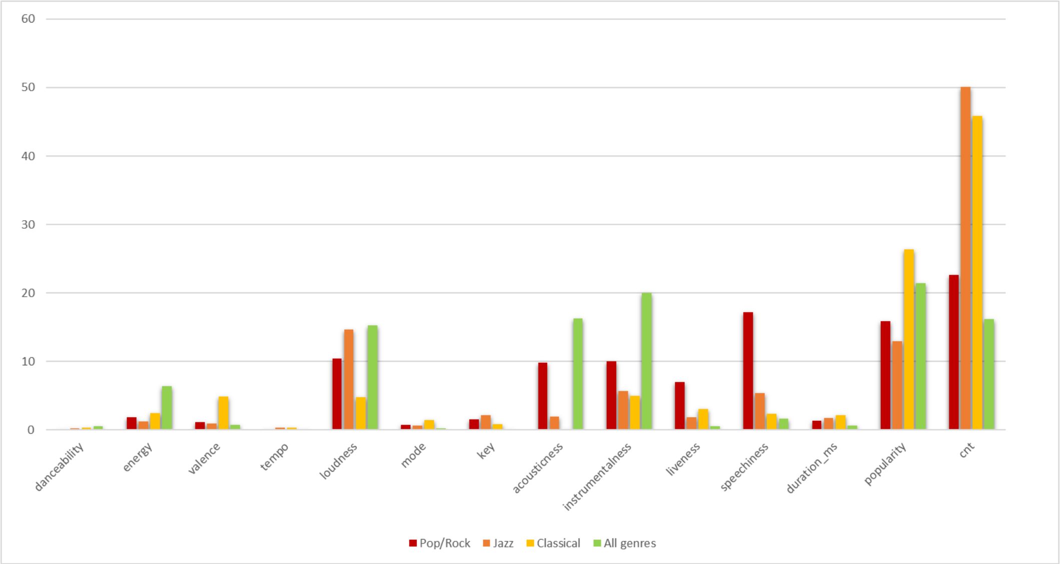


Figure 14: The change of music characteristic of three genres

4 Conclusions

In order to analyze the difference, relation and evolution of music. In the paper, we adopted several models. Some models are introduced by our analysis, such as **N-dimensional weighted distance model (NDWDM)** and **Centroid of N-dimensional Influence Model (CNDIM)**, but we also adopted existed algorithms and models such as **K-Means Algorithm**, **AdaBoost with Decision Tree Model**, **Entropy Weight Algorithm** and also **Topsis Algorithm** and **Auto regressive Moving Average Model (ARMA)**.

Started from **NDWDM**, the model is aimed to calculate the influence score of each artist and reveal the influence among influencer and follower and influence within genres as well. The result of the model refers that the Beatles is the most influential artist in history and also reveals the close relationship of influencer and follower, indicating that the influencer indeed had great impact on followers. Besides, it outputs the **Genres Influence Heatmap(GIH)** which reveals how genres influence each other.

As for **CNDIM**, the model is introduced to calculate similarities within artists. To form a better solution, we also adopted **K-Means Algorithm** in analyzing the similarities. CNDIM output **Artist Similarities Heatmap(ASH)** as result, and utilize **Topsis Algorithm** to indicate the correlation between artists within and between genres. K-Means Algorithm outputs the result of clustering for genres as result, which supports the result of CNDIM since the genres in the same cluster has high correlation with each other where genre and itself has the highest correlation and we therefore obtain the similarities. Besides, the intermediate product of the product is the key to analyze how genres and music change over time.

While finding indicators to distinguish a genre, I employed **AdaBoost with Decision Tree Model** to perform classification task and let the model to decide which characteristic is the best indicator and which is worst and assign weight to each characteristic.

Entropy Weight Algorithm is the algorithm that actually solve the problem of analysing the changes in music and genres over time. The output of Entropy Weight Algorithm directly indicates which characteristic changes the most among years. Using Entropy Weight Algorithm, we find out the revolution year such as 1944 and 1945, which is the ending year of World

War Two. Meanwhile, we can see how each genres changes differently with the output of Entropy Weight Algorithm. Through the analysis of the weight of music characteristics of each genre, we can identify indicators that reveal the dynamic influencers. Take Pop/Rock, Jazz and Classical for example, for Pop/Rock, the top 3 greatest indicators are "speechiness", "popularity", "count", for Jazz, they are "loudness", "popularity", and "count", and for Classical, they are "instrumentalness", "popularity", and "count". The difference of changes in different genres can be explained in the history.

In the end, after acquiring the changes of genres over years, we utilized **Auto regressive Moving Average Model (ARMA)** to predict the future changes according to the history.

5 Strengths and weaknesses

5.1 Strengths

- **Extensibility and Flexibility**
  The model consists of several separate models and algorithms. All of them are able to handle different task and they can run independently or cooperatively. While there is a need to handle different task, we can find the best combination of models and algorithms. On the other hand, our model dynamically calculate the number of characteristic. Therefore, there is no worry when the characteristic of the music enrich.

- **Reliability**
  While measuring the similarities within genres, the result of **Centroid of N-dimensional Influence Model** and the result of **K-Means Algorithms** support and verify each other, indicating the reliability of the result. Also, the result of models and algorithms are reasonable

- **Innovative**
  Our solution, such as the calculating of revolution score of each artist and applying machine learning model in finding the indicators of distinguishing genres are creative.

- **Simplicity**
  The calculation of our model and the logic behind is rather simple.

- **Objectivity**
  All the weights and score are base on calculation without subjective factors.

- **Coverage**
  We have solved all the problems proposed.

5.2 Weakness

- While considering the influence the influencer has on follower, we didn't take the genre of the influencer and the follower into consideration. The impact within and between genres should be weighted differently.

- While computing the similarity of the two artists by calculating the distance in space, we dropped four characteristics including popularity, duration, key and mode of the artist because we thought even the influencer has a great impact on the follower, these attributes may still differ greatly. However, it lacked further confirm.

6 Document to ICM

TO:ICM

FROM:Team 2113454

DATE:February 8, 2021

RE:Target of learning about the influence of music

Today, music has become a necessary cultural activity in our life. There are various influencing factors for artists when they write a new song. In an age with boosting information, we need a model to analysis the influence factor in music.

Firstly, our model is based on data. It means that we can achieve higher accuracy with more data our model gets. This characteristic should be more and more useful as the data can be obtained more and more easily nowadays. However, the model does not over dependence on large dataset. Most of the model and the algorithms adopted in the model can still achieve reasonable accuracy with small dataset except for the **AdaBoost Algorithm with Decision Tree** for calculating how each characteristic contributes in classifying music's genre. However, the weight of each characteristic for classifying music changes slowly and can be therefore pre-calculated or pre-defined.

Besides, our model achieve high reliability. While measuring the similarities within genres, the result of **Centroid of N-dimensional Influence Model** and the result of **K-Means Algorithms** support and verify each other, indicating the reliability of the result. Moreover, our calculation is simple and can be explained clearly in logic, which makes it hard to be incorrect. In the result of our calculation, we can always draw the expected conclusion from the result with high confidence level. Even the prediction error using **Auto regressive Moving Average Model (ARMA)** for the future popularity of each genres can reach below 4%.

On the other hand, our model is fairly powerful. The function of the model has covered all the challenges proposed. However, our model can do even more. To analyze the influence and the similarity of music, we separately introduced **N-dimensional weighted distance model (NDWDM)** and **Centroid of N-dimensional Influence Model (CNDIM)**. Also, we utilized **K-Means Algorithms** while analyzing similarity. As for the change degree of each characteristic of music, we adopted **Entropy Weight Algorithm**. Meanwhile, the **AdaBoost Algorithm with Decision Tree** can easily tell the importance of the characteristic in classifying music and **Auto regressive Moving Average Model(ARMA)** can predict future data with current trend. These algorithms are the base of our model and every proposed challenges are solved with the combination of these algorithms. These algorithms have already handle the basis analysis needed for music and could therefore extend easily to solve more problems. That's the powerful point of our model.

Since our model is based on data, the more data we get result in better accuracy we can achieve. On the other hand, our algorithms compute rather quickly and consume few resources because time and space optimization while implementing with computers such as memorization and discrete operation.Therefore, there is no need to worry about the enrich of data.

In conclusion, our model is simple but powerful with high expansibility and reliability. Moreover, since our model is based on data, the enrich of data can only benefit our model.

References

[1] Daniel Rager, "The Role of Music in Society Past, Present and Future",2008.

[2] Wikipedia, "Normalization(statistics)"[Online], Retrieved from https://en.wikipedia.org/wiki/Normalization_(statistics),2021, 1 8.

[3] Aristidis Likas, Nikos Vlassis, Jakob Verbeek, "The global k-means clustering algorithm",2001.

[4] Trevor Hastie, Saharon Rosset, Ji Zhu, Hui Zou, "Multi-class AdaBoost",2009.

[5] S.R. Safavian, D.Landgrebe, "A survey of decision tree classifier methodology",1991.

[6] Mohammad Valipour, Mohammad Ebrahim Banihabib, Seyyed Mahmood Reza Behbahani, "Comparison of the ARMA, ARIMA, and the autoregressive artificial neural network models in forecasting the monthly inflow of Dez dam reservoir",2012.

[7] William W.S. Wei, "Time Series Analysis",2013.

[8] Ian Cross, "Music, Cognition, Culture, and Evolution",2001.

[9] Qiyue Chen, "Structure entropy weight method to confirm the weight of evaluating index",2010.

[10] Marek Korczynski, "Music at Work: Towards a Historical Overview",2003.

Appendices

Due to the requirement of 25-page limit, we showed one of our code in the following.

The code of Adaboost algorithm:

```
#!/usr/bin/env python
# coding: utf-8
import pandas as pd
import math
import matplotlib as mpl
import matplotlib.pyplot as plt
import numpy as np
import seaborn as sns
import networkx as nx
from sklearn.ensemble import AdaBoostClassifier
from sklearn.tree import plot_tree
from sklearn.tree import DecisionTreeClassifier
import random

music_pd=pd.read_csv('full_music_data.csv')
music_pd

music_pd[['artists_id']]=music_pd[['artists_id']]
.apply(lambda x: x[0].lstrip("[")
.rstrip("]") if "," not in x[0] else np.nan,axis=1)
music_pd=music_pd.dropna(how='any').reset_index().drop(['index'],axis=1)

drop_list=['artist_names','popularity','year','release_date'
,'song_title (censored)']
music_pd=music_pd.drop(drop_list,axis=1)
music_pd

influence_pd=pd.read_csv('influence_data.csv')
influence_pd1=influence_pd[['influencer_id','influencer_main_genre']]
influence_pd1.columns=['artists_id','genre']
influence_pd2=influence_pd[['follower_id','follower_main_genre']]
influence_pd2.columns=['artists_id','genre']
influence_pd1=pd.concat([influence_pd1,influence_pd2],axis=0)
influence_pd1=influence_pd1.groupby('artists_id').agg({
    'genre':'first'
})
influence_pd1=influence_pd1.reset_index()

influence_pd1[['artists_id']]=influence_pd1[['artists_id']].astype(np.int64)
music_pd[['artists_id']]=music_pd[['artists_id']].astype(np.int64)

music_pd=pd.merge(music_pd,influence_pd1,on='artists_id',how='left')
music_pd=music_pd.dropna(how='any')
music_pd=music_pd.reset_index().drop(['index','artists_id'],axis=1)

rd=np.random.permutation(music_pd.shape[0])
music_pd=music_pd.loc[rd]

music_pd=music_pd.reset_index().drop(['index'],axis=1)

music_pd
```

```
x=0.8
trX=music_pd[0:int(music_pd.shape[0]*x)]
teX=music_pd[trX.shape[0]:music_pd.shape[0]]

trY=trX[['genre']]
trX=trX.drop(['genre'],axis=1)
teY=teX[['genre']]
teX=teX.drop(['genre'],axis=1)

trX=np.array(trX)
teX=np.array(teX)
trY=np.resize(np.array(trY),(trY.shape[0],))
teY=np.resize(np.array(teY),(teY.shape[0],))

bdt =
AdaBoostClassifier(DecisionTreeClassifier
(max_depth=100,min_samples_split=10,min_samples_leaf=5),
algorithm="SAMME",
n_estimators=200, learning_rate=0.8)

bdt.fit(trX,trY)

bdt.score(trX,trY)

bdt.score(teX,teY)

ar=bdt.feature_importances_
print(ar)
```