Problem Chosen	2021 MCM/ICM Summary Sheet	Team Control Number
D		2125116

Rock the Music

Summary

The goal of our model is to understand and measure the influence of previously produced music on new music and musical artists, based on provided data. Specifically, the first step of our model is to establish a network of influencers and followers. After that, we need to establish an indicator to measure the influence of influencers, the definition of indicators is inspired from “Paper Evaluation System”, which has established complete evaluation influence system. In order to simplify the model, we did not define music influence like what the thesis evaluation system has done we simply define the music influence of Define the influencer’s musical influence as the number of his followers. This definition is simple and realistic. It is rough and inaccurate to directly address the data regardless of plenty of factors behind this sophisticated problem. In order to find how danceability, energy, valence, tempo and other factors contribute to the influence of music one appropriate method is to first find similar music based on musical characteristic, then compare how they influence the development of music.

For the Second part We define the similarity between genre and genre based on Euclidean distance and use these distances to construct the “genre distance matrix”. On similarity, we have judged if artists within genre more similar than artists between genres.

For the third part we use Projection selection method to choose factors which distinguish a genre form other. Projection selection method has the similar principle with “PCA”, both of them filter variables according to the variance contribution rate after projection. Using the selected variables, we test whether influencers really had an impact on followers and find out which musical characteristics are more contagious.

Finally, we take the time factor into consideration, researching music influence, music characteristics and genres changes over time. We visually display the music characteristics and the trend of genres over time through a line chart. We try to find the relationship between music characteristics and genres changes over time through graphs and we also try to use graphs to find the time nodes of major changes in music and explain them. We use “SVM” to establish the connection between music influence and music characteristics, and try to establish its connection with time through this indirect method, because he can't directly connect with time, At the end of this passage we explain how our model reacts to changes in the external environment and analyze our model

keywords: relational network diagram, SVR, PCA

CONTENT

1 Introduction

1.1 Background

Since ancient times, music has been a part of human society. But how previously produced music affects new music and music artists has always been a mystery. In order to understand and measure the impact of music on new music and music artists, using data on songs released by multiple famous artists and their relationships, we hope to find a way to quantify the impact of previously produced music on the development of new music. First, we use the data of artist relationships to establish an artist network, and establish a measure of music similarity based on the data of songs released by artists. Next, analyze different music characteristics and events such as whether "influencers" have influence on the music creation of "followers" through the network to see which of these characteristics are most relevant to "music influence", among which "music influence" "It is measured by the number of "influencers" that influence "followers." Then, based on the extracted features, we judge whether the "influencer" really has a great influence on the music creation of the "follower" and try to find out the artists and important events that have had a great influence in the development of music based on this . Finally, we try to explain the development laws of certain genres and the influence of music on other aspects of society (for example: culture) through our model.

2 Analysis of the problem

2.1 Data preprocessing

- Make a natural link between *full_music_data.csv* and *influence_data.csv* for easy reading.
- Eliminate a few chorus tracks to facilitate code writing.
- Since the data values of each dimension are quite different, normalize them.

$$x_k = \frac{x_i - \min(x_k)}{\max(x_k) - \min(x_k)}$$

2.2 Basic Assumption

In order to use the provided data to build a powerful music impact analysis model, we made several key assumptions:

- Use the Euclidean distance between the characteristics of different music as an indicator of music similarity. This is because if the two types of music are more similar, the values of certain music characteristics should be similar. According to the calculation formula of Euclidean distance, the Euclidean distance between them will be smaller. So, in order to reduce the complexity of calculation, we directly use the distance between music as a measure of its similarity.

- If the follower and the influencer are in the same genre, we think that the follower is affected by the influencer, and vice versa. other words, although the influencers and followers belong to different genre situations, the influencers may also have an influence on the followers, but in order to ensure that the model can better handle more critical issues, we ignore this part of the

influence, that is, we only consider the influence When the follower and the follower belong to the same genre, there is the influence of the influencer on the follower

2.3 Overview

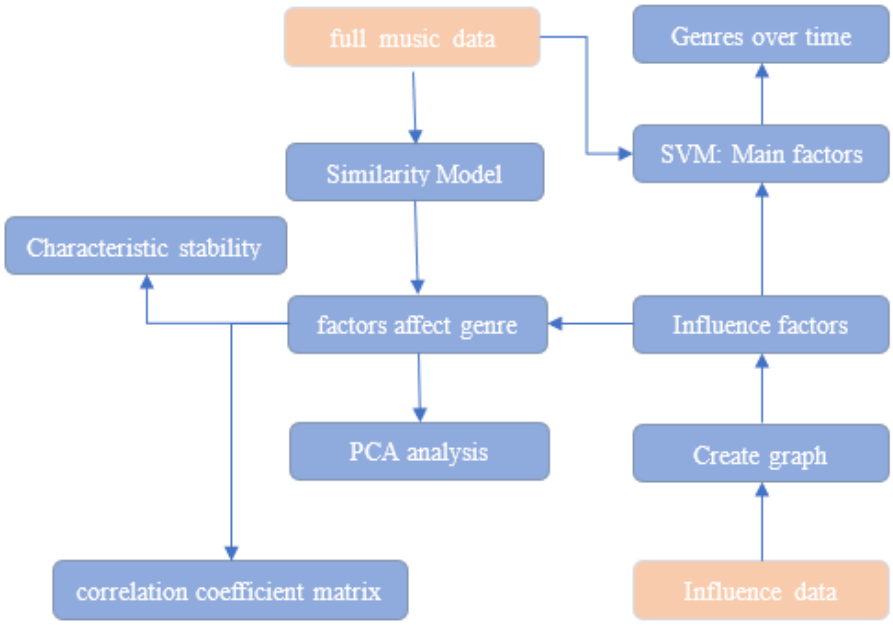

Figure 1: Model overview

3 The Model

3.1 Construct A Graph Theory Model to Measure the Artist's Musical Influence

In order to visually express the influence relationship between artists and the size of influence, we build a directed graph G between artists according to *influence_data.csv*, and describe the similarity between artists or music by Euclidean distance.

3.1.1 A Directed Graph That Establishes an Artist's Influence Relationship

In order to better analyze the relationship between influencers and followers and the mechanism of their influence, we have generated an artist relationship network. The artist relationship network is a directed graph, in which each node is an artist, and there are directed edges between nodes Representatives from influencers to followers. Use adjacency to indicate the relationship between influencer A and follower B.

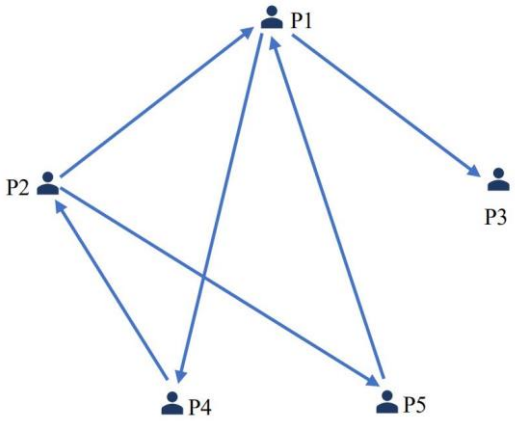

Figure 2: An example of directed graph

Its corresponding adjacency matrix G is as follows,

	$P1$	$P2$	$P3$	$P4$	$P5$
$P1$	0	0	1	1	0
$P2$	1	0	0	0	1
$P3$	0	0	0	0	0
$P4$	0	1	0	0	1
$P5$	1	0	0	0	0

Figure 3: An example of directed graph adjacency matrix

Because $p1$ has an effect on $p2$, $G[p1][p2]$ is equal to 1, otherwise it is 0.

3.1.2 The Definition of Music Influence

How do we judge the influencer's musical influence is closely related to the model's performance? In our model, an influencer's musical influence consists of one main factor:

- The number of followers he affects, which is the out degree of his node

Therefore, the music influence F of artist x is defined as

$$\sum_{i=1}^{n} G[x][i]$$

This is because if an influencer can influence more followers, his musical influence should be greater. Based on this, we define the influencer's music influence is the degree of output of each node in the network. By calculating the music influence of each artist, the following music influence table is obtained as follows (take the top 5)

Influencer_id	Influencer_name	Music_influence
754032	The Beatles	615
66915	Bob Dylan	389
894465	The Rolling Stones	319
531986	David Bowie	238
139026	Led Zeppelin	221

Table 1: The top 5 music influence artists

3.1.3 Choose A Subnet to Describe the Influence Relationship Between Artists

Since the entire relationship network is too large to be conducive to visualization, we selected one of the subnets for visualization. This subnet describes the influence network of an artist named Pestilenc and shows how an artist influences an artist and this How artists influence other artists, in our network, other subnets are roughly the same except for the number of artists involved.

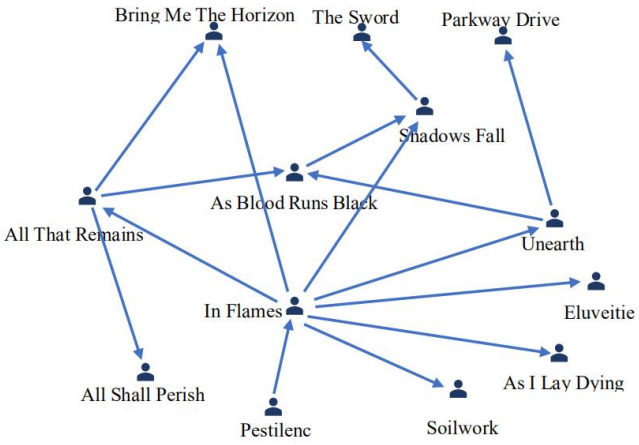

Figure 4: A subnet that affects the relationship network graph

In this network, Pesilenc is the root node with an music influence of 1, but its "child" In Flames affects 7 people, with an music influence of 7. We have also considered whether the influence of the offspring needs to be superimposed on the parent, but when the offspring become famous, the parent may no longer be in the music industry. Therefore, it is reasonable for every artist to use the outdegree to measure this influence.

3.2 Definition of Music Similarity

3.2.1 Distance Definition of High-dimensional Vector

In the definition of distance in high-dimensional space, common ones are Euclidean distance and

Manhattan distance.

For calculation and explanation, we use the definition of Euclidean distance as the definition of distance The Euclidean distance between two n-dimensional vectors $x_1 = [x_{11}, x_{12} \dots x_{1n}]^T$, $x_2 = [x_{21}, x_{22} \dots x_{2n}]^T$ is defined as follows:

$$d_{12} = \sqrt{\sum_{k=1}^{n} (x_{1k} - x_{2k})^2}$$

3.2.2 Construct the Distance Matrix between Different Genres

In order to judge whether the music of artists belonging to the same genre is more similar, we use their corresponding "music feature vector" distance to measure similarity. According to our hypothesis, if the average distance between the music of artists of the same genre is smaller than that of artists of different genres The average distance between us can be considered that the music of artists of the same genre is more similar than the music of artists of different genres. We use the *data_by_artist.csv* data set to calculate the distance between each artist and other artists through the characteristics of each artist, and then average according to the genre classification to obtain a 20-dimensional square matrix D. The square matrix D is the average distance of genre. matrix. Among them, the element $D[i][j]$ (i, j are not equal) of the square matrix D represents the average distance between the artists belonging to the i-th genre and the j-th genre, that is, the distance between the music of different genre artists Average distance, the diagonal element $D[i][i]$ represents the average distance between artists belonging to the i-th genre, that is, represents the average distance between music of artists of the same genre. The distance matrix D is obtained as follows:

Vocal	Stage& Screen	Relig-ious	Regg-ae	R&B	Pop/-Rock	New Age	Latin	Jazz	Inter-Natianl	Unk-Nown	Folk	Elect-ronic	Easy Liste-ning	Cou-ntry	Com-Edy/S Poken	Clas-sical	Chil-dre-n's	Blues	Avant-Garde
0.68	0.78	0.75	0.94	0.87	0.84	0.86	0.86	0.80	0.80	0.80	0.67	0.95	0.65	0.72	0.81	0.77	0.71	0.77	0.82
0.78	**0.75**	0.86	1.12	1.02	0.97	0.82	1.02	0.89	0.94	1.00	0.81	1.06	0.73	0.90	0.93	0.75	0.91	0.94	0.87
0.75	0.86	**0.68**	0.89	0.82	0.74	0.94	0.81	0.84	0.80	0.71	0.72	0.87	0.69	0.66	0.76	0.88	0.71	0.76	0.89
0.94	1.12	0.89	**0.80**	0.83	0.88	1.16	0.82	0.93	0.87	0.78	0.88	0.92	0.92	0.77	0.90	1.16	0.79	0.82	1.07
0.87	1.02	0.82	0.83	**0.82**	0.83	1.06	0.81	0.88	0.84	0.76	0.83	0.89	0.84	0.74	0.84	1.06	0.77	0.79	1.00
0.84	0.97	0.74	0.88	0.83	**0.75**	1.04	0.81	0.89	0.83	0.69	0.80	0.87	0.78	0.69	0.79	1.01	0.77	0.79	0.97
0.86	0.82	0.94	1.16	1.06	1.04	**0.85**	1.08	0.94	1.00	1.07	0.88	1.10	0.82	0.98	0.99	0.81	0.97	1.00	0.93
0.86	1.02	0.81	0.82	0.81	0.81	1.08	**0.79**	0.88	0.82	0.70	0.82	0.89	0.83	0.71	0.83	1.06	0.75	0.77	1.00
0.80	0.89	0.84	0.93	0.88	0.89	0.94	0.88	**0.84**	0.85	0.85	0.78	0.95	0.77	0.80	0.87	0.91	0.80	0.82	0.91
0.80	0.94	0.80	0.87	0.84	0.83	1.00	0.82	0.85	**0.82**	0.74	0.77	0.93	0.77	0.72	0.83	0.96	0.74	0.77	0.93
0.80	1.00	0.71	0.78	0.76	0.69	1.07	0.70	0.85	0.74	**0.40**	0.73	0.86	0.75	0.57	0.75	1.03	0.65	0.67	0.95
0.67	0.81	0.72	0.88	0.83	0.80	0.88	0.82	0.78	0.77	0.73	**0.63**	0.93	0.64	0.66	0.79	0.79	0.65	0.72	0.81
0.95	1.06	0.87	0.92	0.89	0.87	1.10	0.89	0.95	0.93	0.86	0.93	**0.90**	0.91	0.84	0.89	1.10	0.90	0.90	1.06
0.65	0.73	0.69	0.92	0.84	0.78	0.82	0.83	0.77	0.77	0.75	0.64	0.91	**0.59**	0.67	0.76	0.73	0.69	0.74	0.79
0.72	0.90	0.66	0.77	0.74	0.69	0.98	0.71	0.80	0.72	0.57	0.66	0.84	0.67	**0.55**	0.72	0.92	0.60	0.66	0.87
0.81	0.93	0.76	0.90	0.84	0.79	0.99	0.83	0.87	0.83	0.75	0.79	0.89	0.76	0.72	**0.78**	0.95	0.76	0.80	0.93
0.77	0.75	0.88	1.16	1.06	1.01	0.81	1.06	0.91	0.96	1.03	0.79	1.10	0.73	0.92	0.95	**0.68**	0.91	0.96	0.84
0.71	0.91	0.71	0.79	0.77	0.77	0.97	0.75	0.80	0.74	0.65	0.65	0.90	0.69	0.60	0.76	0.91	**0.49**	0.67	0.86
0.77	0.94	0.76	0.82	0.79	0.79	1.00	0.77	0.82	0.77	0.67	0.72	0.90	0.74	0.66	0.80	0.96	0.67	**0.71**	0.91
0.82	0.87	0.89	1.07	1.00	0.97	0.93	1.00	0.91	0.93	0.95	0.81	1.06	0.79	0.87	0.93	0.84	0.86	0.91	**0.88**

Table 2: distance matrix

3.2.3 Explore the Relationship between The Similarity of Music and Genre of Artists

According to the above distance matrix D, we can draw the following table:

Music Genres	Rate	Music Genres	Rate
Vocal	85%	Unknow	95%
Stage & Screen	85%	Folk	95%
Religious	90%	**Electronic**	**55%**
Reggae	80%	Easy Listening	95%
R&B;	70%	Country	95%
Pop/Rock	80%	Comedy/Spoken	70%
NewAge	80%	Classical	95%
Latin	75%	Children's	95%
Jazz	**60%**	Blues	80%
International	**55%**	**Avant-Garde**	**60%**
Amount of genres	**20**	**Threshold value**	**70%**

Table 3: the ratio of the number of differences between each genre and different genres to the amount of genres.

The table shows the ratio of the number of differences between each genre and different genres to the amount of genres. The larger the ratio, the greater the ratio of the music between artists belonging to this genre than the artists belonging to this genre. Similar to music that does not belong to this genre artist, here we assume the threshold is 70%. We can intuitively see from the data reflected in the root table that for most genre, their internal artists are more similar, except for Jazz, International, Avant-Garde and Electronic. Through the research on these four genres, we found that it may be caused by the following reasons lead to this phenomenon.

Electronic (0.55)

"Electronic" belongs to the category of "modern pop music", a genre of modern pop music. Before the 1990s, most of the early "electronic" music works were suitable for dancing; with the development of the times, more and more styles of "electronic music" were assembled. Its creative characteristics have also become "attention to the rhythm of dancing music, very loose song structure (if there is a structure)", and other musical characteristics are not obvious.

International (0.55)

International incorporates the characteristics of many other music genres. "World music" refers to music that has not fallen into the "pop" or "folk" (Folk) traditions of North America and the United

Kingdom, and is derived from the fusion of local music around the world. Styles like Jamaica's "Reggae" or "Latin Pop" music have developed and grown large enough to be attributed to their own music genres, but for everything from traditional Chinese music to African "folk" (Folk) Music and all other local music, have to be collectively referred to as "world music." Therefore, the characteristics of its own music are not much different from other music types.

Avant-Garde (0.60)

The core of "Avant-Garde" is to break through the boundaries of music and create new sounds. At the same time, during the development of "Avant-Garde", a variety of new music genres, such as Wagner, Debussy, etc., have been produced, which has also continuously expanded the boundaries of music. Therefore, the musical characteristics of Avant-Garde are not obvious.

Jazz(0.60)

Improvisation is the core element of "jazz" music. With its own evolution, "jazz" music has evolved many different styles, from "Be-Bop" (emphasizing fast and powerful rhythm), "Cool Jazz" (creating serene, Soft harmony) to "Free Jazz" (emphasizing tension and atonality of musical impact) and "Soul Jazz" (using a simple rhythm pattern). The factor that sustains all these different styles is that they are all built on the basis of "blues" music-relying on the interaction between the members of the orchestra and unpredictable impromptu performance, in the entire development process and all the different styles, the positive It is these musical characteristics that define "jazz" music, and have little to do with the musical characteristics of Jazz itself.

All in all, artists with similar works have similar styles

3.3 Choose Some Characteristics: Main factors Affecting Genre

3.3.1 Principal component analysis

Principal component analysis (PCA) is the process of computing the principal components and using them to perform a change of basis on the data, sometimes using only the first few principal components and ignoring the rest.

3.3.2 Determine Distinguished Characteristics: Projection Selection

Our idea comes from PCA. In PCA, in order to ensure that the loss of information after dimensionality reduction processing is as small as possible, the projection dimension selected in PCA is the dimension with relatively large variance after projection, because the greater the variance, the greater the implication in this dimension The more information is, the more it contributes to the overall. We adhere to this idea to project the 7-dimensional characteristic vectors of different genres into one dimension, and then calculate the mean value of the variance contribution rate of each characteristic.

The table obtained is as follows:

characteristics	Variance rate
tempo	0.2
energy	0.2
valence	0.1
loudness	0.1
danceability	0.1

Table 4: the mean value of the variance contribution rate of each characteristic

From the above table, we can conclude that tempo and energy play a relatively large role in distinguishing different genres, and loudness, valence and danceability will also affect the division of genres, but the impact is significantly weaker than the first two indicators

3.3.3 Use the Correlation Coefficient Matrix to Measure the Correlation between Genres

In order to judge whether there is a correlation between genre, we need to calculate the correlation coefficient between different genre, here we use Pearson correlation coefficient, given by the following formula:

$$\rho_{XY} = \frac{E[XY] - E[X]E[Y]}{\sqrt{E[X^2] - [E[X]]^2}\sqrt{E[Y^2] - [E[Y]]^2}}$$

Here we assume that the correlation coefficient is greater than 0.8 as a strong correlation. If the absolute value of the calculated correlation coefficient is greater than 0.8, it indicates that the two genres are positive/negative. Through the method just now, we have selected several of the most important factors as follows: tempo, energy, valance, loudness and danceability

The correlation coefficient matrix calculated by these influencing factors is as follows:

1.00	0.68	-0.04	0.95	0.74	0.22	0.95	-0.28	0.59	0.88	0.96	-0.01	0.95	0.87	0.65	-0.56	-0.10	0.67	0.73	0.98
0.68	1.00	0.08	0.77	0.17	0.86	0.67	-0.48	0.67	0.91	0.74	-0.15	0.85	0.91	0.96	-0.33	0.44	-0.54	0.40	0.54
-0.04	0.08	1.00	-0.08	0.34	0.19	0.12	0.71	0.56	0.22	-0.24	0.82	-0.12	-0.14	0	0.76	0.73	0.65	0.56	-0.11
0.95	0.77	-0.08	1.00	0.52	0.38	0.83	-0.50	0.46	0.88	0.97	-0.23	0.95	0.95	0.81	-0.66	-0.10	-0.79	0.54	0.89
0.74	0.17	0.34	0.52	1.00	-0.26	0.82	0.41	0.62	0.56	0.55	0.56	0.52	0.34	0.06	-0.03	0.02	-0.08	0.93	0.79
0.22	0.86	0.19	0.38	-0.26	1.00	0.24	-041	0.49	0.61	0.31	-0.15	0.47	0.61	0.84	-0.02	0.67	-0.23	0.06	0.01
0.95	0.67	0.12	0.83	0.82	0.24	1.00	-0.06	0.78	0.90	0.88	0.22	0.91	0.79	0.57	-0.33	0.11	-0.47	0.85	0.94
-0.28	-0.48	0.71	-0.50	0.41	-0.41	-0.06	1.00	0.31	-0.24	-0.52	0.94	-0.46	-0.61	-0.63	0.85	0.39	0.87	0.44	-0.22
0.59	0.67	0.56	0.46	0.62	0.49	0.78	0.31	1.00	0.80	0.46	0.61	0.61	0.51	0.48	0.26	0.69	0.07	0.84	0.51
0.88	0.91	0.22	0.88	0.56	0.61	0.90	-0.24	0.80	1.00	0.85	0.11	0.93	0.91	0.85	-0.30	0.35	-0.49	0.73	0.79
0.96	0.74	-0.24	0.97	0.55	0.31	0.88	-0.52	0.46	0.85	1.00	-0.26	0.98	0.94	0.75	-0.72	-0.18	-0.83	0.53	0.94
-0.01	-0.15	0.82	-0.23	0.56	-0.15	0.22	0.94	0.61	0.11	-0.26	1.00	-0.17	-0.31	-0.33	0.81	0.58	0.75	0.67	0
0.95	0.85	-0.12	0.95	0.52	0.47	0.91	-0.46	0.61	0.93	0.98	-0.17	1.00	0.97	0.82	-0.60	0.03	-0.74	0.58	0.89
0.87	0.91	-0.14	0.95	0.34	0.61	0.79	-0.61	0.51	0.91	0.94	-0.31	0.97	1.00	0.93	-0.64	0.06	-0.79	0.43	0.78
0.65	0.96	0	0.81	0.06	0.84	0.57	-0.63	0.48	0.85	0.75	-0.33	0.82	0.93	1.00	-0.46	0.29	-0.65	0.26	0.51
-0.56	-0.33	0.76	-0.66	-0.03	-0.02	-0.33	0.85	0.26	-0.30	-0.72	0.81	-0.60	-0.64	-0.46	1.00	0.69	0.97	0.14	-0.58
-0.10	0.44	0.73	-0.10	-0.02	0.67	0.11	0.39	0.69	0.35	-0.18	0.58	0.03	0.06	0.29	0.69	1.00	0.52	0.33	-0.23
-0.67	-0.54	0.65	-0.79	-0.08	-0.23	-0.47	0.87	0.07	-0.49	-0.83	0.75	-0.74	-0.79	-0.65	0.97	0.52	1.00	0.02	-0.66
0.73	0.40	0.56	0.54	0.93	0.06	0.85	0.44	0.84	0.73	0.53	0.67	0.58	0.43	0.26	0.14	0.33	0.02	1.00	0.72
0.98	0.54	-0.11	0.89	0.79	0.04	0.94	0.22	0.51	0.79	0.94	0	0.89	0.78	0.51	-0.58	-0.23	-0.66	0.72	1.00

Table 5: The correlation coefficient matrix of each genre

3.4 Re-discussion on Influencers and Followers

3.4.1 Do the 'Influencers' Actually Affect the Music Created by the Followers?

Through the artist relationship network established above, we can calculate the ratio P of the total followers who are really affected by the influencer to the total followers. Among them, all followers are the total number of directed edges in the network, and the sum of followers that are really affected by the influencer is the total number of the same directed edges of the child node and the parent node. Here we assume that if the P is greater than 0.7, it indicates that the influencer has an influence on the followers. It is calculated that 80.2% of the followers are affected by the influencers, which proves that the influencers do influence the music creation of the followers.

3.4.2 Are Some Music Characteristics More 'Contagious' than Others?

Depth-first search (DFS) is an algorithm for traversing or searching tree or graph data structures. The algorithm starts at the root node (selecting some arbitrary node as the root node in the case of a graph) and explores as far as possible along each branch before backtracking.

Influence can play the role of assimilation, so the more stable the attribute along an influence chain, the stronger the aggressiveness. If a certain musical feature is more infectious than others, then the indicators of influencers and followers affected by it should be more similar, so we use the variance of each musical feature to measure the infectious power of this musical feature. The smaller the variance, the greater the infectivity, and vice versa, the smaller the infectivity. For each music feature, calculate the mean variance of its followers and influencers to get the following results:

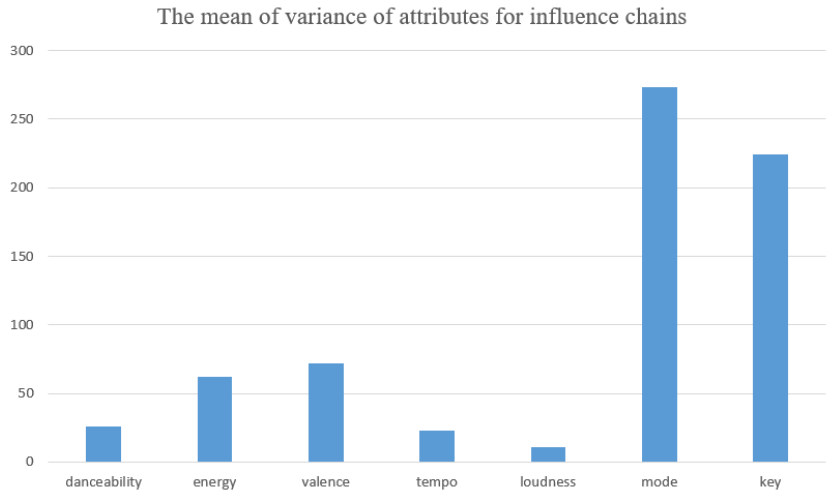

Figure 5: The mean of variance of attributes for influence chains

3.4.3 Results of the Analysis

It can be seen from the above table that the mean variance of the three characteristics of tempo, loudness, and danceability in the influence chain is significantly smaller than the other four indicators, which shows that *tempo, loudness, and danceability* are more contagious in all music characteristics, while energy, *valence, mode, and key* are not so contagious.

3.5 Exploration of Music Development in the Past Hundred Years

3.5.1 The Relationship of Genres over Time

Integrate some niche genres, and the changes of each genre over time are shown in the figure

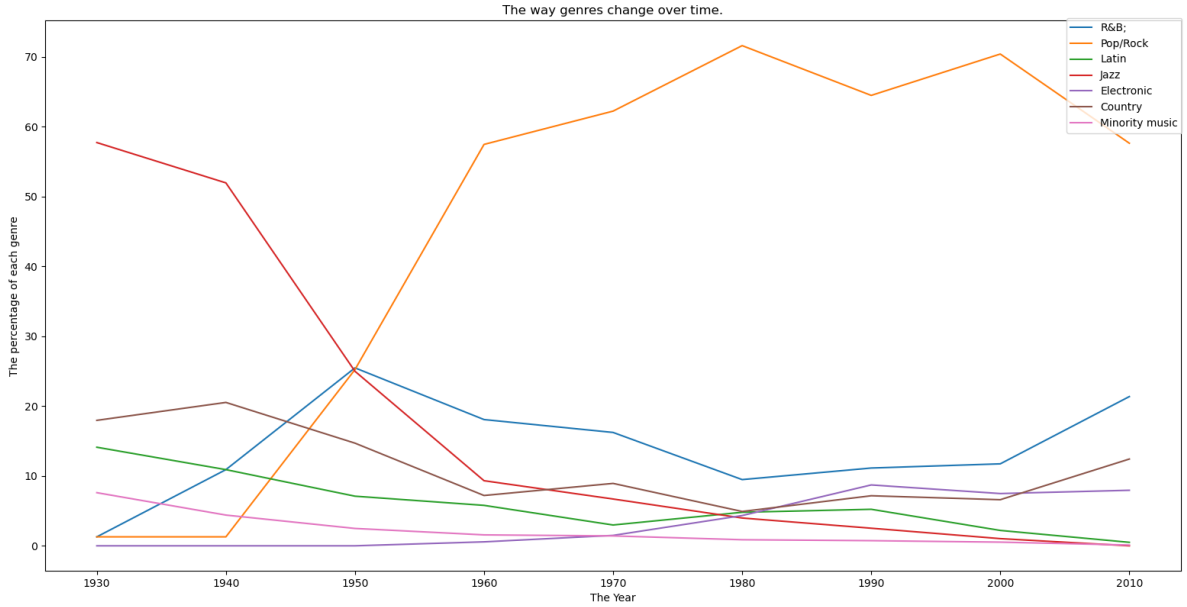

Figure 6: The way genres change over time

3.5.2 The Relationship of Characteristics over Time and What Artists Represent Revolutionaries?

Through the *data_by_year.csv* data set and the *full_music_data.csv* data set, we can clearly observe the change trend of the ratio of different music characteristics and music types over time. We draw it into a graph to get the following figure:

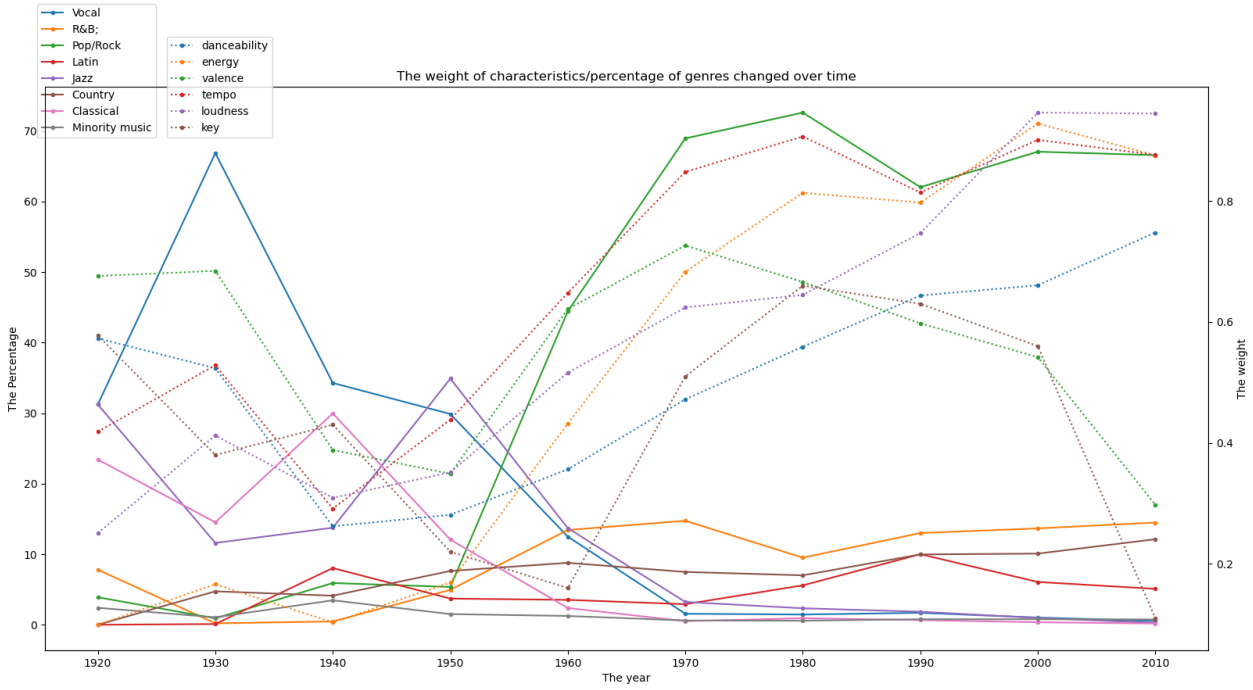

Figure 7: the way characteristic & percentage of genres changed over time

After normalizing each feature, since the proportion of style is also in the range of 0~1, the image shows them surprisingly clear.

Among them, the relatively niche music types have less data and the effect of individual analysis is not obvious. Therefore, we re-merge all the niche music types into a new music type called niche music for analysis. Here we pass each the number of singles released in ten years is used as a criterion for judging whether a genre is a niche. If the number of singles released per ten years for a music genre is always 650, we consider the music genre to be niche music, and the music characteristic mode is not obvious. So, we choose to eliminate mode and use the remaining 6 music features for analysis.

From the above picture, we observe that the proportion of Pop/Rock in all music genres has risen sharply since 1950, while the proportion of Vocal has dropped sharply, so we think that 1950 is a year of major changes in music. , It can be intuitively found from the figure that the change trend of music characteristics danceability, energy, valence, tempo, loudness is similar to that of Pop/Rock, that is, these music characteristics are all Pop/Rock different from Vocal, which is different from us. The results obtained in 3.3 are consistent, that is, pop/Rock and Vocal are negatively correlated, and the characteristics of energy, valence, tempo, and loudness are just the characteristics of dividing different genres. Therefore, we believe that these five characteristics can be used as indicators to judge whether music has undergone revolutionary changes.

In order to explore the people who led this trend, we extracted music released as Pop/Rock between 1950 and 1960 and analyzed their authors. Because the greater the music influence, the easier it is to lead the trend, so you can find the artists with the greatest influence among these authors, and they can be trend leaders.

Sequence	ArtistId	Artist Name	Music influence
1	**180228**	**Elvis Presley**	**166**
2	120521	Chuck Berry	159
3	631774	The Byrds	158
4	538677	Buddy Holly	97
5	824022	Little Richard	88
6	55128	Bo Diddley	65
7	233066	Lou Reed	58
8	46699	The Everly Brothers	53
9	332141	Jerry Lee Lewis	53
10	852007	Roy Orbison	47
Grand total			944

Table 6: the greatest influence artists among Pop in 1950

The results found that Elvis Presley, Chuck Berry, The Byrds and others have the largest impact factors, that is, these three artists promoted the popularity and development of Pop/Rock between 1950 and 1960. After 1970, The Beatles will promote the popularity and development of Pop/Rock. Development reached its peak.

3.6 Measure the dynamic influencers of Pop/Rack

3.6.1 Introduction to SVM

In machine learning, support-vector machines (SVMs, also support-vector networks) are supervised learning models with associated learning algorithms that analyze data for classification and regression analysis. Hyperplane was given by $w^T \cdot x + b = 0$where w is the normal vector to the hyperplane and w is the points on the plane. A valid hyperplane satisfies $\begin{cases} w^T \cdot x + b \geq 0, for\ y_i = \quad 1 \\ w^T \cdot x + b < 0, for\ y_i = -1 \end{cases}$ Support vectors are points closest to hyperplane, which are edges of two different sets.

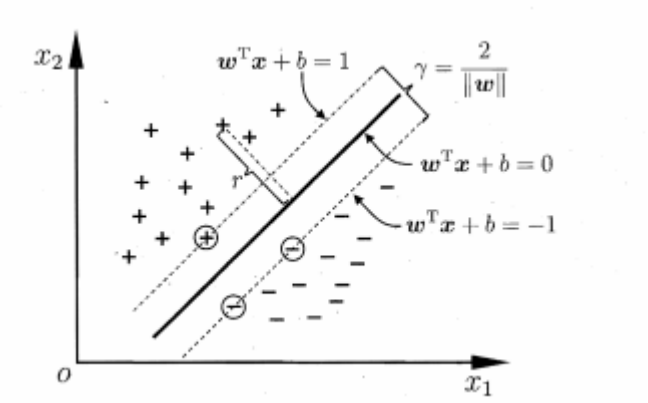

Figure 8: Support vector machine classifier

The distance between the two heterogeneous support vectors to the hyperplane is

$$\gamma = \frac{2}{||\boldsymbol{w}||},$$

we call it margin.

The purpose of classification is to make the distance of the same category as close as possible, and the distance of different categories as far as possible, that is, to make the maximum interval as large as possible. Which is

$$\begin{aligned} &\min_{\boldsymbol{w},b} \ \frac{1}{2}\|\boldsymbol{w}\|^2 \\ &\text{s.t. } y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b) \geqslant 1, \quad i = 1, 2, \ldots, m. \end{aligned}$$

3.6.2 Use the Idea of SVM to Find the Weight of Different Characteristics

Data preprocessing

- Read the *data_by_artist* feature X as a feature vector and the influence factor of the artist to which it belongs.
- Musicians whose extraction type is only Pop/Rock
- Set the threshold of the influence size. If it is greater than 20, it will be 1, which means that the musician has great influence, and 0 is not so great.
- Divide 80% of the data into the test set, and divide the remaining 20% into the prediction set

Choice of kernel function

After many attempts, we chose a *linear* function with a relatively large correct rate (*81%*). The classification accuracy rate is not very high, and an increase in the amount of data may have a good effect. But it still gives us enlightenment, which dimensions are more weighted? Another advantage of choosing a linear kernel function is that the weight of the split hyperplane can be directly output.

Perform the above preprocessing on part of (40000) data in the full_music_data.csv file, and find the artist corresponding to each song, then the model can be trained. After our training, we get the parameters w and b respectively as

$$w = [-15.28, 3.14, 4.67, -0.21, -0.47, 7.21, -0.23]^T \cdot 10^{-3} \qquad b = -0.99087797$$

We know that the hyperplane is the standard for dividing different categories. When $w^T \cdot x + b \geq 0$, this category is considered to have a large impact. In other words, the distance between the hyperplane $w \cdot x^T + b = -1$ and points as far as possible from the hyperplane and score as high as possible. Based on this, we can define an artist's score S as

$$w^T \cdot x + b + 1$$

S is an indicator to measure the ability to influence. Different from the previous impact factor, when S is relatively large, it is easier to produce excellent industry leaders. This is because the weights used are generated by this definition. Statistics of Pop/Rock songs and artists in each year can draw the dynamic influence curve of Pop/Rack each year.

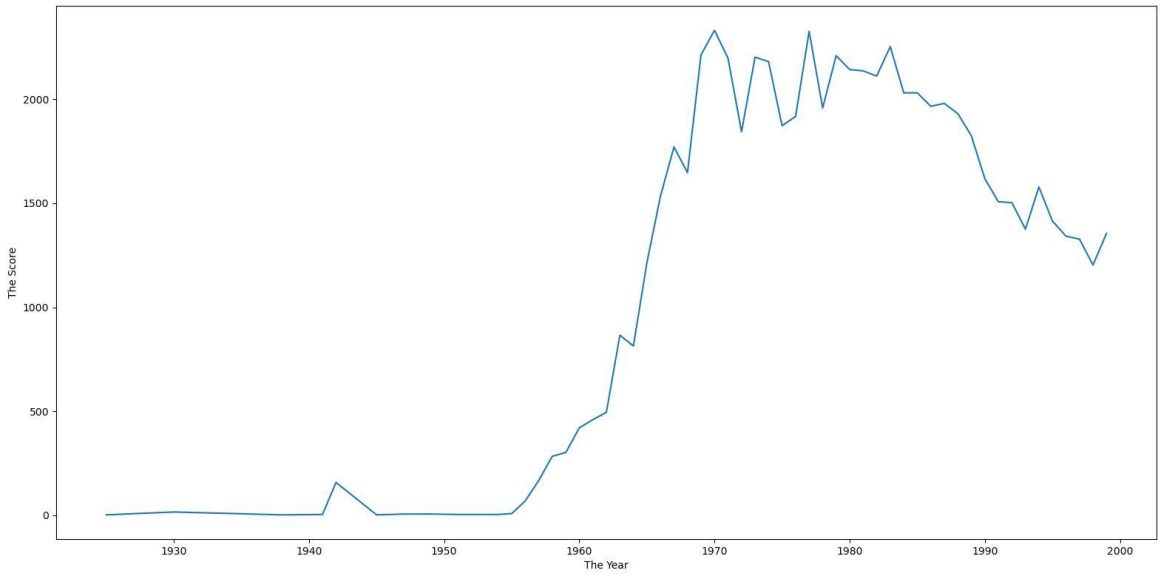

Figure 9: indicators of dynamic influencers over time

As can be seen from the above picture, as time goes by, our scores as a whole show an increase, which is basically consistent with the trend of Pop/Rock. This shows that the Pop/Rock type has a great influence on artists, reflecting the changes in the ability of the entire music industry to influence. And there is a tendency to assimilate to other types.

3.6.2 Algorithm flow pseudo code

This is the pseudo code of our algorithm

Algorithm：Gets the total dynamic influencer of the POP/Rock for a given year with svr regression

```
# 4000 × 7 input data
features←Empty dictionary
Obtain the total dynamic influencer of POP in a certain year to obtain the weights and constants w,b of SVR regression training.
for Each year yi do:
    feature←empty set
    #pi is the character vector of the current music
    for Each Pop/Rock pi do:
        Use [w*pi+b+1+feature] as the current feature
    end for
    store features[yi][pi]←[feature]
end for
#Show how features change over time
chart←[y,features]
```

3.7 The influence of the new century on the music industry

In our model, the development of music is closely related to characteristics such as danceability, energy, valence, tempo, loudness, etc. Therefore, if social, political, and technological changes affect these characteristics and cause them to undergo major changes, this change will make music Revolutionary changes have taken place, such as a certain type of music suddenly becoming the mainstream, or the emergence of a certain new type of music, etc. We can observe this change through network models, and then trace the source to find the year of change, and look for the external occurrence at that time What has changed has revolutionized music.

The mechanism of how our model recognizes external influences is as follows. Whenever a new follower is generated, we will record the generation time and calculate the similarity between its music and the existing music of each genre through his music characteristics. Here We set a threshold of 0.7. If its similarity with the existing genres is less than the threshold, we will not include it in the existing network system and generate an unincluded record. If it appears within a period of time Many times it has not been included, which means that the external environment has changed and one or more new genre types different from the existing genres have been produced. In the same period of time, most of the new followers are included in the same genre, and the genre is not the most

mainstream genre before, we will also think that the external environment has changed. However, due to the robustness of the model itself, the model is not sensitive to small changes in music characteristics, so if the external environment has a small impact on danceability, energy, valence, tempo, and loudness, our model may not be able to recognize this effect.

4 Model Analysis

Even if we fully consider the rationality of the model, some situations cannot be avoided.

- According to some artists, there are many influencers, which is not in line with reality. How many influencers can a person generally remember? Our model does not consider the difference brought about by this aspect.
- Established images among 4000 artists and performed depth-first search, achieving good efficiency. But when the data is larger, the efficiency of the algorithm will decrease.

5 Document

Dear Sir

We are honored to introduce our model to you, you can easily and accurately understand the influence of music through it. Firstly, using the given data set, we create a network that can visually reflect the relationship between influencers and followers and then We establish a music similarity measurement system through music characteristics. Through this system, we can judge whether the music is similar or not, and we can also judge whether the genres are similar. Then we used this system to find out what distinguishes a genre and how do genres change over time. Finally, we established an indicator to measure the influence of influencers on followers, from which we can judge whether influencers really influence followers, and further find out how influencers influence followers.

According to our research, we found that most genres have obvious characteristics and only a few genres, such as jazz, have no obvious characteristics. These uncharacteristic genres are gradually declining. This is related to the way in which influencers influence followers. When influencers have an influence on followers, the contagious musical characteristics of influencers are often easily imitated by followers. The indicator of strong influence happens to be the characteristic of pop/rock distinction, which has caused pop and other genres to become more and more popular while jazz is in decline.

Considering that some genres are not taken into consideration, our model's analysis of the influence of music may be different from the actual situation. At the same time, due to the limitations of the data used by our model, it is not possible to measure the impact of some external factors on music well. If we can use more data,

we can take more genres into consideration so that our model can better explain the influence of music, and more data can also take into account changes in the external environment, which can increase our model the ability to recognize changes in the external environment and its robustness.

Nevertheless, our model is excellent enough. Using our model correctly, you can get the influence of music. Thank you for your trust in our team and hope our work will satisfy you.

Your sincerely
Team 2125116

6 References

[1] James Bergstra, Dumitru Erhan, Aggregate features and ADABOOST for music classification, Springer, 2006

[2] Zhouyu Fu; Guojun Lu; Kai Ming Ting; Dengsheng Zhang, A Survey of Audio-Based Music Classification and Annotation, IEEE Transactions on Multimedia, 2011

[3] Mandel, Michael I.; Ellis, Daniel P. W. Song-Level Features and Support Vector Machines for Music Classification, Columbia University Libraries, 2005

[4] Changsheng Xu; N.C. Maddage; Xi Shao, Automatic music classification and summarization, IEEE Transactions on Speech and Audio Processing. 2005

[5] Keunwoo Choi; György Fazekas; Mark Sandler; Kyunghyun Cho, Convolutional recurrent neural networks for music classification, 2017

[6] Keunwoo Choi, György Fazekas, Mark Sandler, Kyunghyun Cho, Transfer learning for music classification and regression tasks, CVPR 2017

[7] Joseph Zakzeski, Pieter C. A. Bruijnincx, Anna L. Jongerius, The Catalytic Valorization of Lignin for the Production of Renewable Chemicals, ACS, 2010

[8] https://blog.csdn.net/guanyuqiu/article/details/85109441

[9] https://blog.csdn.net/dlhlsc/article/details/107300514

[10] https://blog.csdn.net/zm_1900/article/details/89106643

[11] http://blog.sina.com.cn/s/blog_915bf3ed01017v0h.html

[12] https://blog.csdn.net/github_39261590/article/details/75009069

[13] https://blog.csdn.net/qq_39514033/article/details/88931639

7 Appendix

In the four-day competition, the two programmers wrote a total of more than 3000 lines of code. Due to space limitations, some core codes were posted for reference.

Get Data From Excel by Python Code

```
def GetArtistData():
    df = pd.read_csv('data_by_artist.csv')  # 返回一个 DataFrame 的对象，这个是 pandas 的一个数据结构
    df.columns = ["Col1", "Col2", "Col3", "Col4", "Col5", "Col6", "Col7", "Col8", "Col9", "Col10", "Col11", "Col12",
                  "Col13", "Col14", "Col15", "Col16"]

    X = df[["Col3", "Col4", "Col5", "Col6", "Col7", "Col8", "Col9"]]
    X = np.array(X)
    X = (X - np.min(X, axis=0)) / (np.max(X, axis=0) - np.min(X, axis=0))
    ArtistId = df[["Col2"]]
    ArtistId = np.array(ArtistId)
    return X,ArtistId

def CreateMap():
    genre = ["Vocal", "Stage & Screen", "Religious", "Reggae", "R&B;",
             "Pop/Rock", "New Age", "Latin", "Jazz", "International", "Unknown",
             "Folk", "Electronic", "Easy Listening", "Country", "Comedy/Spoken", "Classical",
             "Children's", "Blues", "Avant-Garde"]
    for i in range(len(genre)):
        DIC[genre[i]] = i

def GetDistance(ArtistDataArr):
    numlist = [[0]*(len(DIC)) for i in range((len(DIC)))]
    Distance = [[0]*(len(ArtistDataArr)) for i in range((len(ArtistDataArr)))]
    GenreDis = [[0]*(len(DIC)) for i in range((len(DIC)))]
    for i in range(len(ArtistDataArr)):
        for j in range(len(ArtistDataArr)):
            if MusicionType[i] in DIC.keys() and MusicionType[j] in DIC.keys():
                numlist[DIC[MusicionType[i]]][DIC[MusicionType[j]]]+= 1
                Distance[i][j] = np.sqrt(np.sum(np.square(ArtistDataArr[i]-ArtistDataArr[j])))
                GenreDis[DIC[MusicionType[i]]][DIC[MusicionType[j]]] += Distance[i][j]
    GenreDis = np.array(GenreDis)
    numlist = np.array(numlist)
    GenreDis = GenreDis/numlist
    return Distance,GenreDis

def AllMusicionType():
    df = pd.read_csv('influence_data.csv')  # 返回一个 DataFrame 的对象，这个是 pandas 的一个数据结构
    df.columns = ["Col1", "Col2", "Col3", "Col4", "Col5", "Col6", "Col7", "Col8"]

    list1 = df[["Col1", "Col3"]]  # 抽取前八列作为训练数据的各属性值
    list2 = df[["Col5", "Col7"]]

    temp1 = np.array(list1)
    temp2 = np.array(list2)
    temp = np.concatenate((temp1,temp2),axis=0)

    MusicionType = np.array(list(set([tuple(t) for t in temp])))
    return MusicionType

def GetMusicionType(UnorderType,ArtistId,section):
    MusicionType = [None for i in range(len(ArtistId))]
    for i in range(len(ArtistId)):
        tempId = str(ArtistId[i][0])
        if tempId in section:
            ind = section.index(tempId)
            MusicionType[i] = UnorderType[ind][1]
    return MusicionType
```

Create Graph by Python Code

```
def create_g(arr):
    global dic
    global G
    global node_max
    global indegree

    for i in range(len(arr)):
        influencer_id           = arr[i][0]
        influencer_name         = arr[i][1]
        influencer_main_genre   = arr[i][2]
        influencer_active_start = arr[i][3]
        follower_id             = arr[i][4]
        follower_name           = arr[i][5]
        follower_main_genre     = arr[i][6]
        follower_active_start   = arr[i][7]

        if influencer_id not in dic:
            node_max += 1
            dic[influencer_id] = node_max
            DIC[node_max] = arr[i][0:4]

        if follower_id not in dic:
            node_max += 1
            dic[follower_id] = node_max
            DIC[node_max] = arr[i][4:8]

        G[dic[influencer_id]][dic[follower_id]] = 1
        indegree[dic[follower_id]] += 1
        outdegree[dic[influencer_id]] += 1
```

Get Distance by Python Code

```
def GetDistance(ArtistDataArr):
    numlist = [[0]*(len(DIC)) for i in range((len(DIC)))]
    Distance = [[0]*(len(ArtistDataArr)) for i in range((len(ArtistDataArr)))]
    GenreDis = [[0]*(len(DIC)) for i in range((len(DIC)))]
    for i in range(len(ArtistDataArr)):
        for j in range(len(ArtistDataArr)):
            if MusicionType[i] in DIC.keys() and MusicionType[j] in DIC.keys():
                numlist[DIC[MusicionType[i]]][DIC[MusicionType[j]]]+= 1
                Distance[i][j] = np.sqrt(np.sum(np.square(ArtistDataArr[i]-ArtistDataArr[j])))
                GenreDis[DIC[MusicionType[i]]][DIC[MusicionType[j]]] += Distance[i][j]
    GenreDis = np.array(GenreDis)
    numlist = np.array(numlist)
    GenreDis = GenreDis/numlist
    return Distance,GenreDis
```

PCA by Python Code

```
    CreateMap()
    ID_GEN = AllMusicionType()
    X, y = gd.GetAllData()
    X = X[0:30000]
    y = y[0:30000]
    l = [_ for _ in range(7)]
    print(l)
    tmpX = select_dim(X, l)
    # X, r_id = GetArtistData() # X 0 ~ 10
    pca = PCA(n_components=7)
    pca.fit(X)
    show_vector(pca.explained_variance_ratio_)
    show_vector(pca.explained_variance_)
```

Draw photo by Python Code

```
def GetSimpleMatrix(Num_Matrix):
    Minor = []
    SimpleMatrix = []
    SimpleGenre = []
    for i in range(len(Num_Matrix)):
        if max(Num_Matrix[i]) <= 650:
            Minor.append(Num_Matrix[i])
        else:
            SimpleMatrix.append(Num_Matrix[i])
            SimpleGenre.append(genre[i])
    SimpleMatrix.append(np.mean(Minor,axis=0))
    SimpleGenre.append("Minority music")
    SimplePercentArr = SimpleMatrix/(np.sum(SimpleMatrix,axis=0))*100
    return SimplePercentArr,SimpleGenre

def CreateLineChart(SimplePercentArr,SimpleGenre):
    x = year
    for i in range(len(SimplePercentArr)):
        y = SimplePercentArr[i]
        plt.plot(x,y,marker='.')
    plt.legend(SimpleGenre, loc='lower left', bbox_to_anchor=(0.77, 0.2))
    plt.title("The way genres change over time.")
    plt.xlabel("The Year")
    plt.ylabel("The Percentage")
    plt.show()
```

```
def ShowChart():
    MusicArr = GetReadyData()

    all_musicion_type = AllMusicionType().tolist()
    section = [i[0] for i in all_musicion_type]
    MusicionType, MusicArr = GetMusicionType(all_musicion_type,
MusicArr, section, MusicArr)

    Num_Matrix = [[0] * (len(year)) for i in range((len(genre)))
]
    for i in range(len(MusicArr)):
        Num_Matrix[MusicionType[i]][YEAR_DIC[str(MusicArr[i][1])
]] += 1

    SimplePercentArr, SimpleGenre = GetSimpleMatrix(Num_Matrix)
    CreateLineChart(SimplePercentArr, SimpleGenre)
```

DFS by Python Code

```
Nodelist = []
def dfsGetMaxIndex(index):
    visit1[index] = 1
    global Nodelist
    Nodelist.append(index)
    for i in range(len(templist[index])):
        if (visit1[templist[index][i]] == 0):
            dfsGetMaxIndex(templist[index][i])

SimilarNum = 0
AllNum = 0
#对节点数最多的树进行 dfs，求得所有边条数和所连节点类型相同的边条数
def dfs(index):
    global SimilarNum
    global AllNum
    visit2[index]=1
    for i in range(len(templist[index])):
        AllNum += 1
        if DIC[index][2] == DIC[templist[index][i]][2]:
            SimilarNum += 1
        if(visit2[templist[index][i]]==0):
            dfs(templist[index][i])

MaxNum = 0
MaxNumIndex = -1
#得到各孤立树中树的节点数最多的树的根节点
def GetMaxSimilarity():
    global MaxNum
    for i in range(1,node_max + 1):
        if visit1[i] == 0 and indegree[i]==0:
            visit1[i] = 1
            dfsGetMaxIndex(i)
            if(len(Nodelist)>MaxNum):
                MaxNum=len(Nodelist)
                MaxNumIndex=Nodelist[0]
                Nodelist.clear()
    dfs(MaxNumIndex)
    return SimilarNum/AllNum
```

SVM by Python Code

```
def train(X, y, X_test, y_test):
    # svm_clf = svm.LinearSVC(C = 30, max_iter=3000)
    # svm_clf = svm.SVC(C=1, kernel='rbf')
    svm_clf = svm.SVC(C=1, kernel='linear')
    svm_clf.fit( X, y )
    res = svm_clf.predict(X_test)

    cnt = 0
    for i in range(len(X_test)):
        if res[i] == y_test[i]:
            cnt += 1
    print(cnt/len(X_test))

def get_weight(X, y):
    print("start svm ...")
    classifier = SVC(kernel='linear', C=1)
    svm = classifier.fit(X, y)
    b = classifier.intercept_
    w = classifier.coef_

    for i in range(len(X)):
        if y[i] == 1:
            print(str(np.dot(np.array(w), np.array(X[i]) + b)) +
 " " + str(y[i]))

    return w, b
```