Xi'an Jiaotong-Liverpool University
西交利物浦大学

Cover Sheet

Please fill in all the blanks below for your assignment:

Full Name	Wehao Wang Zixi Wang Mingnan Wei Shaoying Wu Xiangyu Xiao Ningning Xu Yilin Xue	Student ID Number	2032083 2032139 1613489 2034860 2034863 2032686
Group Number (e.g. 7)	8	Programme	INT
Module Title	Data mining and big data analytics	Module Code	INT402
Assignment Title	Cluster different genres according to music features		
Submission Deadline	Oct 28th, 2020	Tutor's Name	Xi Yang
Final Word Count	1668		
If you agree to let the University use your work anonymously for teaching and learning			yes

I certify that I have read and understood the University's Policy for dealing with Plagiarism, Collusion and the Fabrication of Data (available on ICE: http://ice.xjtlu.edu.cn/mod/resource/view.php?id=7178). With reference to this policy, I certify that:

- My work does not contain any instances of plagiarism and/or collusion.
- My work does not contain any fabricated data.

By uploading my assignment onto ICE, I formally declare that all the above information is true to the best of my knowledge and belief.

Scoring – For Tutor Use									
Stage of Marking	**Initials**	**TF Features Achieved**		**LU Features Achieved**			**TF Score**	**LU Score**	**Final Score**
		(A-D)	**(E-G)**	**(A-D)**	**(E-G)**	**(H-J)**			
1st - red pen									
Moderation – green pen		The original mark has been accepted by the moderator (please circle as appropriate):							Y / N
		Data entry and score calculation have been checked by another tutor (please circle):							Y

2nd (if required) - green pen									

For Academic Office Use			**Possible Academic Dishonesty (please circle as appropriate)**		
Date Received	Days Late	Late Penalty	Plagiarism: - Minor - Major	Collusion (major)	Data: - Embellishment (minor) - Fabrication (major)

Students: Please start your assignment on the next page.

Xi'an Jiaotong-Liverpool University

西交利物浦大学

INT 402 Peoject Proposal - Cluster different genres according to music features

Student Name:	Wenhao Wang, Zixi Wang, Mingnan Wei
	Shaoying Wu, Xiangyu Xiao, Ningning Xu, Yilin Xue
Student ID:	2032083 2032139 1613489
	2034860 2034863 2032472 2032686
Lab Group #:	8
Date of Deadline:	28th / Oct / 2020

Contents

1 Introduction

Music, as a carrier of emotion transmission, has become an inseparable part of people's daily life. Under different occasions and emotions, we tend to choose different types of music, in that case, it's essential to distinguish dissimilar features of songs then label them in diverse clusters. Every good online music service platform has no shortage of recommendations for different genres of music.

Although musical genre classification is a difficult and problematic task, we believe it is worth pursuing and worth to be improved. In recent years, the popularity of music streaming services like NetEase's cloud music and Spotify,etc. is increasing rapidly, with more and more numbers of users. Our research result would help consumers and music streaming service companies as well as artists. For consumers, they can get customized and personalized playlists. For companies, they can use music catalogues, either as recommendations to customers or as products alone to make a profit. As for artists, attracting potential subscribers to their music will become easier and easier.

In the follow section 'The Problem' we put forward the problems we will encounter and solve. In 'The Data' section, we list the data sets we will apply and analyze. The Methodology' paragraph states an overview of the methods used in this project, mainly divided into two parts, data preprocessing and select model, besides we also simply mention how far we have prepared by now. After that, we focus on the results, significance and possible limitations and challenges of this project in 'The discussion' part. Finally, we put a Gantt Chart,which breaks down the project by phase and task, task start and end date, and percent completed.

2 The Problems

For human beings, it is not very difficult to classifying music into different genres, but let the computer to understand and differentiate the musical genre is not that simple. Previous researches try to use different kinds of machine learning methods to deal with music data in order to differentiate music genres. What they attempt to use including decision tree, regression, SVM, Naïve Bayes, and neural network. However, most of these studies utilized limited and abstract music dataset with low-level features, such as timbre.

To do this research, it will be an important part of our project to use a dataset with various features that can help us to do our research. The objective of our research is to, using the entire Spotify dataset with a different approach – clustering analysis to differentiate music genres. There will be three main questions for us to explore. Firstly, how do genres look like? Secondly, we are going to use K-nearest Neighbours and Cosine Similarity Algorithm to figure out can we predict genres. Thirdly, we want to know can we cluster genres into bigger genre groups. In the end, we deeply hope our research could contribute positively to the music industry.

3 Relevant Work

Music style classification provided by music streaming service is a very meaningful direction in the field of music analysis, because it can facilitate users to choose their own sound conveniently and efficiently. For accurate classification, iMusic system is introduced [1]. It classifies classical music through feature point analysis and matching. It uses a k-means clustering algorithm for deep learning, and the iMusic based on this algorithm framework can achieve an accuracy of + 88%. There are many clustering algorithms to solve this problem [2]. The common clustering methods can be divided into partition based, hierarchy based, density based and network-based algorithms. The k-means algorithm mentioned just now is a clustering method based on partition. It is characterized by large amount of calculation, which is very suitable for small and medium-sized databases. In order to improve the efficiency of music matching [3], one-sided continuous matching clustering algorithm can be used to automatically group data sets. This method can not only improve the efficiency, but also ensure the accuracy We want to solve the problem can be understood as a statistical analysis method to classify and study the same feature. Our goal is to classify objects with greater similarity into the same category, so as to maximize the efficiency when users choose their own music classification.

4 Data Collection Report

This section is a report on the dataset information we will analyze. We got this dataset from the official website of Kaggle. This dataset is about the audio features of 160k+ songs released in between 1921 and 2020, and the provenance of this dataset is Spotify Web API. The collection methodology is using the search query of Spotify for data collection. More specifically, the authoir searched for tracks with respect to the year given (e.g. search?q=year:2018 for songs released in 2018) Given the features of each song, he could build the first part of dataset. Using the id key, the audio features of each track id were retrieved . After further data cleaning process such as one-hot encoding, the data was ready.

The dataset is made of one files with a size of 28.3MB. There are 19 features of the data set in this file, and the sample size is 169,910. The features that used to describe the dataset on numeric are acousticness, danceability, duration_ms, instrumentalness, valence, popularity, tempo, liveness, loudness, speechiness and year; Some dummy features in the dataset are mode and explicit; Features key, artists, release_date and name are used to describe the dataset categorically.

Looking at the entire data set, we will find that some songs have some missing items or some outliers (the value of some features cannot be negative), such as duration_ms. For this situation, we decided to temporarily fill them with NA. After analyzing the relationship between each feature, they need to be normalized.

For features with strong relationships, we use the mean value filling method to solve them. For weak relationship features, it may be more accurate to use K-means to obtain data.

Based on the overall situation, we will carefully clean the data according to the unique nature of each feature. For example, release_date and year can basically be regarded as duplicates. Since there are mostly outliers in release_date (items with no exact date), this feature can be deleted and data analysis can be performed by using feature 'year'.

5 Methodology

In the following, we will explain the methodology that we will use in this project. It should be noted that the current plan is just the beginning of this project. We have carefully planned the project, but if there are some uncertain factors in the future, we will still make some adjustments to the project. The general idea will not change.

5.1 Data Preprocessing

In the data preprocessing part, because our data is not mined by ourselves, some features need to be adjusted accordingly. For example, the features of singers. In some examples, a song is completed by two singers, but in the sample, they are put into a value. Therefore, when processing this dataset, we need to split them into two datasets. For another example, some features are related to making music, such as tone, key, etc. We also need to learn the corresponding background knowledge.

5.2 Model Selection

Our task is to cluster different genres according to audio features. We need to use multiple approach to adjust. In each model, hyperparameters need to be tuned. Finally, compare the accuracy of the forecast. The first thing that comes to mind is to use sklearn and its cluster module for data processing. Under the supervised learning, K-means, Naive Bayesian algorithm and Mean-Shift Clustering should also be considered.

5.3 Others

So far, we have selected a suitable model and completed relatively mature predictions. After the cluster prediction results are completed, we can compare with the author's standard data values to know whether the model is good or bad. If we are able to follow up, we plan to use a library called matplotlib pyplot to complete the data visualization process.

6 Discussion

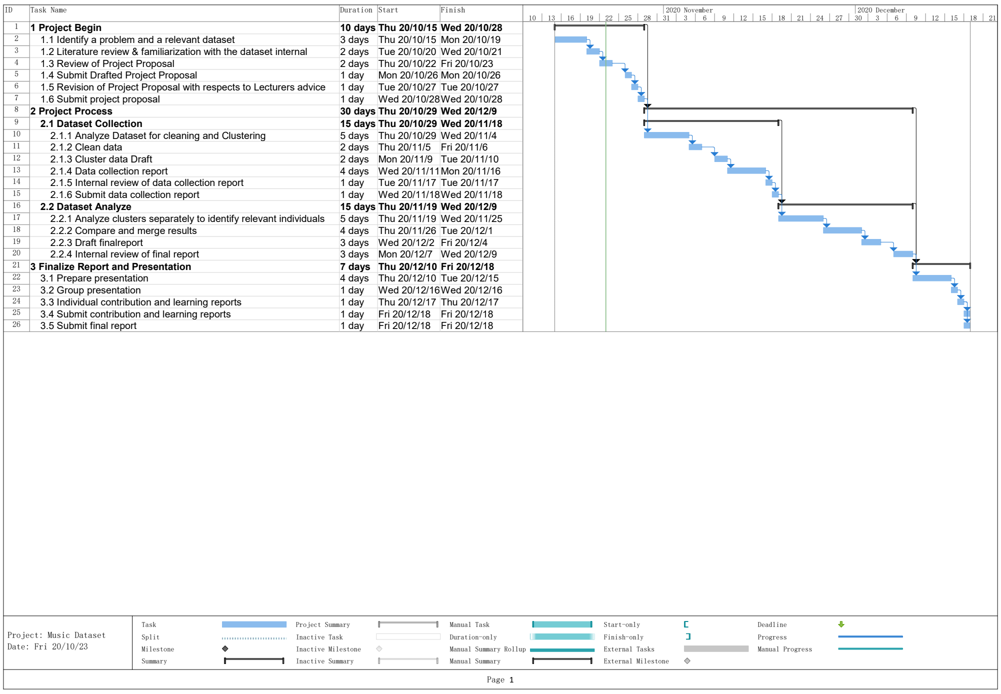
To summarize, the project will identify different genres of music, generate the cluster classifications and demonstrate the variable and valuable music trends using numerous description and visualization methods based on the music tracks dataset on Spotify during 1921 to 2020.

However, this project also has some limitations and challenges cannot be ignored. The first one is about the quality of data set. In the data collection part, the whole data applied comes from the official website Kaggle. Though the number of features for analyzing are extremely huge, the really suitable and valuable are limited. Nextly, there are also some features such as the popularity or creativity that cannot avoid the objective influence of the audiences. Especially these features also are heavily affected by the limitations of time span, because the different generation would gives different opinions or the same generation would give different opinions in different time. This means the tags of songs for users actually varies evrey moment and the fixed data set in any moment which has no ability of updating according to the taste of audiences would have a huge and serious deviations. In addition, there are some missing itemsor some outliers in the collected data. In spite of adopting the normalization to filled with these mistake items, this non-doubtly would lead to the cluster without some accuracy.

The second point is that the cluster methodology has the defects of .Moreover, the current cluster method could result in producing so massive small genres, which can not be added into bigger groups. It would heavily reduce the significance and value of song genres classification. Hence, the task of evaluation and promotion for cluster results are also vital for this project.

In the end, the application value of the project also need to test further. For example, how do we design a recommendation mechanism and share the similar songs by means of user's currently favorite genres clusters. Additionally, the project may try to test if the mechanism is able to share some excellent songs in user's non-touchable field. All these try would hope to maximize the efficiency, improve the user's experience and contribute the music industry positively.

7 Gantt Chart



ID	Task Name	Duration	Start	Finish
1	**1 Project Begin**	**10 days**	**Thu 20/10/15**	**Wed 20/10/28**
2	1.1 Identify a problem and a relevant dataset	3 days	Thu 20/10/15	Mon 20/10/19
3	1.2 Literature review & familiarization with the dataset internal	2 days	Tue 20/10/20	Wed 20/10/21
4	1.3 Review of Project Proposal	2 days	Thu 20/10/22	Fri 20/10/23
5	1.4 Submit Drafted Project Proposal	1 day	Mon 20/10/26	Mon 20/10/26
6	1.5 Revision of Project Proposal with respects to Lecturers advice	1 day	Tue 20/10/27	Tue 20/10/27
7	1.6 Submit project proposal	1 day	Wed 20/10/28	Wed 20/10/28
8	**2 Project Process**	**30 days**	**Thu 20/10/29**	**Wed 20/12/9**
9	**2.1 Dataset Collection**	**15 days**	**Thu 20/10/29**	**Wed 20/11/18**
10	2.1.1 Analyze Dataset for cleaning and Clustering	5 days	Thu 20/10/29	Wed 20/11/4
11	2.1.2 Clean data	2 days	Thu 20/11/5	Fri 20/11/6
12	2.1.3 Cluster data Draft	2 days	Mon 20/11/9	Tue 20/11/10
13	2.1.4 Data collection report	4 days	Wed 20/11/11	Mon 20/11/16
14	2.1.5 Internal review of data collection report	1 day	Tue 20/11/17	Tue 20/11/17
15	2.1.6 Submit data collection report	1 day	Wed 20/11/18	Wed 20/11/18
16	**2.2 Dataset Analyze**	**15 days**	**Thu 20/11/19**	**Wed 20/12/9**
17	2.2.1 Analyze clusters separately to identify relevant individuals	5 days	Thu 20/11/19	Wed 20/11/25
18	2.2.2 Compare and merge results	4 days	Thu 20/11/26	Tue 20/12/1
19	2.2.3 Draft finalreport	3 days	Wed 20/12/2	Fri 20/12/4
20	2.2.4 Internal review of final report	3 days	Mon 20/12/7	Wed 20/12/9
21	**3 Finalize Report and Presentation**	**7 days**	**Thu 20/12/10**	**Fri 20/12/18**
22	3.1 Prepare presentation	4 days	Thu 20/12/10	Tue 20/12/15
23	3.2 Group presentation	1 day	Wed 20/12/16	Wed 20/12/16
24	3.3 Individual contribution and learning reports	1 day	Thu 20/12/17	Thu 20/12/17
25	3.4 Submit contribution and learning reports	1 day	Fri 20/12/18	Fri 20/12/18
26	3.5 Submit final report	1 day	Fri 20/12/18	Fri 20/12/18

Project: Music Dataset
Date: Fri 20/10/23

Figure 1: Gantt Chart of Project

References

[1] S. B. M. Roy and D. De. imusic: a session-sensitive clustered classical music recommender system using contextual representation learning. *Multimedia Tools and Applications: An International Journal*, no. 79.

[2] S. Pourahmad. Does determination of initial cluster centroids improve the performance of k-means clustering algorithm comparison of three hybrid methods by genetic algorithm, minimum spanning tree, and hierarchical clustering in an applied study. *Computational and Mathematical Methods in Medicine*.

[3] T. Li. Music clustering with features from different information sources. *IEEE Transactions on Multimedia, Multimedia*, 11(3):477–485, 2020.