Data Analysis and Machine Learning on Spotify Data

Katranitsiotis Panagiotis

Abstract. This report investigates Spotify, which is the most common music platform. More specifically, a dataset regarding numerous songs' characteristics uploaded to Spotify, will be used. Spotify contains an enormous variety of songs which categorize the chosen data into the Big Data ones. Several machine learning algorithms will be performed in order to predict several variables that are efficient for Spotify and its users. Both supervised and unsupervised techniques will be used for this accomplishment. More specifically, multiple linear regression, logistic regression and regression trees will be performed for accurate predictions. The purpose of this paper is to use these algorithms in order to provide efficient predictions for songs characteristics, both for Spotify users and music industries such as the prediction of a track's popularity, valence of a song etc. Finally, hierarchical clustering will be performed in order to produce an efficient recommendation song algorithm.

Keywords: Spotify; Big Data; Data Science; Regression; Decision Trees; Supervised learning; Unsupervised learning; Hierarchical clustering;

Contents

1 Introduction

Spotify, nowadays, has become the most common music streaming platform. It was founded in 2006 by Daniel Ek and Martin Lorenzon. The purpose of its creation was to familiarize people with listening to every type of music whenever and wherever they prefer [1]. More specifically, Spotify led to a huge breakthrough in music industry, as the majority of songs were only available on CDs, at that time. If someone wanted to listen to some music, he should either have the appropriate CD or via radio. It is highly considerable, that at that time was also difficult for listening to music during a daily routine such as work, walk, gymnastics, etc. Consequently, Spotify became a highly significant tool for music.

In 2021, Spotify has more than 70 million tracks and 356 million users [2]. Nowadays, almost every song can be found on Spotify on digital form, leading people to the ability of listening to a variety of songs of several types. Moreover, even artists and music industries release their albums directly on Spotify.

Consequently, several data science algorithms can be used for a significant scientific research of this popular streaming platform. More specifically, these tools will be highly efficient for music industries to predict the success of their upcoming albums and generally which song types are more preferable to users.

The purpose of this paper is the usage of several machine learning algorithms on Spotify data, in order to identify the most common songs' characteristics such as tempo, valence etc. This research is highly important as it can be very helpful not only to the musicians but also to Spotify users. More specifically, considering these characteristics, Spotify can produce personalized playlists and recommendation of new songs for users related to their previous preferences. The models of this report will also try to predict the popularity of a song considering several variables such as accousticness, danceability, instrumentalness, etc. Furthermore, efficient models will be produced for predicting the existence of explicit content on a song's lyrics. Finally, a model for predicting the valence of a track will also be produced. Each algorithm for these research questions will be explained extensively in the next chapters.

2 Machine Learning Algorithms

From technical perspective, the dataset used for this report's purpose, contains Spotify data until 2021. This dataset was retrieved from Kaggle and it consists of 170,653 observations of 19 variables [3]. More specifically, it contains audio features from a large amount of songs released from 1922 to 2021. Several variables are provided such as accousticness, key scale, duration, existence of explicit content, tempo, etc. All the variables are explained extensively on Table 1 (*Appendix A*).

In order to produce helpful and benefit prediction models, several machine learning algorithms will be used and more specifically:

- Multiple Linear Regression
- Logistic Regression
- Regression Trees

- Hierarchical Clustering

The necessity of each technique and its purpose for answering each question is presented extensively in the upcoming chapters.

2.1 Pre-processed steps

In this section, all the methods described above, will be performed in order to produce accurate and efficient prediction models. For this accomplishment, a proper training of each model is essential. More specifically, the dataset will be divided into two datasets. The training set, which will contain the 70% of the data, and the test set consisting of the remaining 30%.

From technical perspective, each model will be trained on the training set and will use the test set as the evaluation data. As a result, we will be able to calculate the accuracy of the model and try to produce the most accurate one. More specifically, for the evaluation method, the cross-validation will be used, which is the most common and proper method for choosing the least statistical error model [4].

2.2 Multiple Linear Regression

This method will be used in order to create a model for predicting the popularity of a song. It is highly important for music industries to have the ability to predict the popularity of their upcoming song accurately, both for financial and popularity reasons.

From statistical perspective, in order to create a linear model, the linearity assumption of the data must occur [5]. However, due to the huge amount of data, a pairwised scatterplot will not provide essential insights. For this reason, Pearson correlation coefficient will be calculated for popularity with each numeric variable. Based on figure 1 (*Appendix B*), it is shown that mostly, year, accousticness, energy and loudness are highly related to the popularity of a song. In fact, the variable year has a Pearson correlation with popularity nearly to 0.85, indicating a strong linear relationship between these two variables [6]. A multiple linear model is fitted on the training set, as shown in figure 2 (*Appendix B*), given all the variables as the predictors. It is shown that only the duration of the song is not necessary for predicting its popularity.

Given the test set as the evaluation data, it is concluded that the test error of this model is 117.1354. Although, a mean statistical error of zero indicates the ultimate accuracy, in the case of Big Data this error remains not significantly high. Considering the diagnostic plots, on figure 3 (*Appendix B*), there are some points with high leverage.

2.3 Logistic Regression

Logistic regression is a machine learning algorithm used to predict a binary variable [7]. This method will be used in this dataset to predict whether a song contains an explicit content or not. More specifically, for this purpose several variables will be used as the predictors i.e. key, mode, danceability, liveness, energy, valence, accousticness and loudness.

Performing logistic regression on the training set with these variables, as shown on figure 4 (*Appendix C*), it is observed that nearly all the variables are essential for predicting the explicit content. More specifically, the variables used for this prediction, are the ones with a non-zero coefficient on the logistic regression formula [8]. As far as the songs keys are concerned, the most important are the keys 1, 4, 5, 6, 8, 9 and 11. In addition, significant variables are the mode, danceability, liveness, energy, valence, acousticness and loudness. All these coefficients were accepted considering 95% confidence.

Furthermore, the test data will be used in this model in order to calculate the accuracy of this logistic regression. For this accomplishment, the confusion matrix of the predictions is provided in figure 5 (*Appendix C*). It is observed that this model was able to predict slightly accurately the existence of explicit content. More specifically, its accuracy is 0.92 which represents a well effective prediction model.

2.4 Regression Tree

This technique will try to predict the valence of a track considering its most correlated variables. Valence is a measure with a range from 0 to 1 describing the positiveness of a track. For instance, a track with high valence sounds more positive, i.e. cheerful, happy, enthusiastic. On the other hand, low valence tracks express more negative feelings such as sadness, depression, etc [9].

Firstly, a tree model is fitted on the training set for observing which variables are more useful for the prediction of valence. From the produced tree, as shown in figure 6 (*Appendix D*) and its summary on figure 7 (*Appendix D*), it is observed that only danceability, energy and year were used as the predictors. Furthermore, testing this model on the evaluation data, it is concluded that the mean square of this tree is 0.041.

However, a 10-cross validation will be used to calculate which tree size is the most preferable and consequently the most accurate one. Based on figure 8 (*Appendix D*), 10-cross validation shows that the least statistical error tree, is the one with size 10. Eventually, no pruning of the original tree is needed, indicating that this is the most accurate model.

As a result, an accurate prediction model was created based on the tree methods. More specifically, this model can predict the valence of a track considering only danceability, energy and year as the input variables. Consequently, this method provided a highly accurate prediction model of valence, as its mean square error is extremely low.

2.5 Hierarchical Clustering

Hierarchical Clustering is one of the most common unsupervised machine learning algorithms [10]. This method will be used on Spotify data in order to produce a recommendation song algorithm. More specifically, it will consider the tracks' characteristics such as Accousticness, Danceability, Energy, Instrumentalness and Liveness to create song clusters. From technical perspective, the data of Table 1 (*Appendix A*) needed to be restructured. For simplicity, the above characteristics were transformed into binary variables. For instance, each track with Energy higher than 0.5 was transformed into 1,

meaning that this song is energetic. As a result, each variable will have two levels, 0 and 1, indicating low or high existence of this characteristic in the song. Each variable of this new data is described extensively on Table 2 (*Appendix A*).

Prior to the creation of the clusters, the Euclidean distances of the data needed to be calculated [11]. As a result, these calculations lead to a massive amount of data on the RAM. Consequently, given all the dataset for training, will provide a memory error in R language. In order to avoid this problem, a random dataset of 20,000 songs was used as a training set for clustering. Then, a hierarchical tree was created using the complete method, so as to choose the most appropriate number of clusters, as shown on figure 9 (*Appendix E*) . For this accomplishment, the most appropriate height must be considered. The best height is approximately 500 which corresponds to a size 4 tree. Consequently, the songs were placed into 4 clusters.

For instance, the song Perfect by Ed Sheeran can be found on the data on row 19,024. As shown on figure 10 (*Appendix E*), this song belongs to the second cluster. This means, for providing the five most recommended songs for a user listening to "Perfect", we just need to observe five registrations from the second clusters. Based on figure 11 (*Appendix E*), it is concluded that five recommended songs, among others, are "Power Is Power", "When We Die", "Korma Sönmez", "Bir Zamanlar Bizde Millet" and "Love for Guns".

3 Conclusion

The purpose of this paper was to provide helpful and accurate prediction models considering the dataset of Spotify both for its users and music industries. Various songs characteristics were taken into account for this accomplishment such as the popularity, valence, accousticness, track's key etc.

More specifically, the created multiple linear regression model, is able to predict the popularity of a song slightly accurate. This is highly important especially for music industries due to financial reasons. In addition, the existence of explicit content, which has an immense impact on Spotify, as it is used even by young people, is able to be predicted through the provided logistic regression model. Furthermore, the valence of a song, which describes its positiveness, can also be predicted by the creation of the regression tree. Finally, the unsupervised technique of hierarchical clustering, was also performed in order to create a song recommendation algorithm for a Spotify user. More specifically, it provides five recommended songs close to the user's preference.

Taking everything into consideration, all the above machine learning algorithms were used in order to provide essential insights and accurate prediction model for Spotify. All the presented models are helpful for music industries and Spotify owners as well as for its users as they provide a more personalized interpretation with the application.

References

1. Haupt, J. “Spotify”. *Digital Media Reviews.* Music Library Association. 69(1), pp.132-138, (2012).
2. Spotify, (2021) *Company info*. [Online] Available from https://news-room.spotify.com/company-info/. [Accessed: 22nd May 2021].
3. Kaggle, (2021) *Spotify Dataset 1922-2021*. [Online] Available from https://www.kaggle.com/yamaerenay/spotify-dataset-19212020-160k-tracks. [Accessed: 22nd May 2021]
4. Berrar, D., “Cross-validation”. *Encyclopedia of Bioinformatics and Computational*. Elsevier. 1, pp.542-545, (2018).
5. Uyanik, G. K., Guler, N. “A study on multiple linear regression”. *Procedia – Social and Behavioral Sciences*. Elsevier. 106, pp.234-240, (2013).
6. Bonesty, J., Chen, J., Huang, Y., Cohen, I., “Pearson Correlation Coefficient”, *Springer Topics in Signal Processing*. Springer, pp.1-4. (2009).
7. LaValley, M. P. “Logistic Regression”. *Circulation*. 117(18), pp.2395-2399, (2008).
8. Manning, C., (2007) *Logistic Regression (with R)*. [Online]. Stanford. Available from https://nlp.stanford.edu/~manning/courses/ling289/logistic.pdf. [Accessed: 22nd May 2021]
9. Santos, J. D. (2017) *Is my Spotify music boring? An analysis of involving music, data and machine learning*. [Online]. Towards Data Science. Available from https://towardsdatascience.com/is-my-spotify-music-boring-an-analysis-involving-music-data-and-machine-learning-47550ae931de. [Accessed: 22nd May 2021]
10. Zhu, L., Lin, C., Huang, H., Chen, Y., Yille, A.. “Unsupervised structure learning: Hierarchical Recursive Composition, Suspicious Coincidence and Competitive Exclusion”. *Computer Vision – ECCV 2008.* pp.759-773
11. Bouguettaya, A., Yu, Q., Liu, X., Zhou, X, Song, A. “Efficient Agglomerative hierarchical clustering”. *Expert Systems with Applications.* Elsevier. 42(5), pp.2785-2797, (2015).

Appendix A: Description of the variables

The dataset of Table 1 consists of 170,653 records with 19 variables about Spotify.

Table 1. Description of Variables

	Variable Name	Type	Description
1.	Valence	Numeric	Describes the positivity of each track (From range 0 to 1)
2.	Year	Integer	The year of each registration
3.	Accousticness	Numeric	The accousticness of each song from range 0 to 1
4.	Artists	Character	The artist of the registered song
5.	Danceability	Numeric	Describes how suitable the song is for dancing (from range 0 to 1)
6.	Duration_ms	Numeric	The duration of each song in milli-seconds
7.	Energy	Numeric	Describes how energetic the track is (from range 0 to 1)
8.	Explicit	Categorical	Whether the song contains explicit content (1) or not (0).
9.	Id	Character	The id of each track created by Spotify
10	Instrumentalness	Numeric	The ration of instrumental sounds
11.	Key	Categorical	The key of each song, described as categories from 1 to 11
12.	Liveness	Numeric	Describes the probability of the song to be recorded live (from range 0 to 1)
13.	Loudness	Numeric	The loudness of the song in dB ranged from -60 to 0
14.	Mode	Categorical	The scale of the track. 0: Minor 1: Major
15.	Name	Character	The name of each song
16.	Popularity	Integer	The popularity of each track.
17.	Release date	Integer	The released year of the album containing the specific track.
18.	Speechiness	Numeric	The ration of spoken words
19.	Tempo	Numeric	The tempo of each song in BPM

The dataset of Table 2 is the restructured data of Table 1.

Table 2. Description of Variables for Unsupervised Learning

	Variable Name	Type	Description
1.	Name	Character	The name of each track
2.	Accousticness	Integer	Whether the song is acoustic (1) or not (0).
3.	Danceability	Integer	Whether the song is danceable (1) or not (0).
4.	Energy	Integer	Whether the song is energetic (1) or not (0).
5.	Instrumentalness	Integer	Whether the song is instrumental (1) or not (0).
6.	Liveness	Integer	Whether the song recorded live (1) or not (0).

Appendix B: Multiple Linear Regression

```
    valence        year acousticness danceability duration_ms      energy instrumentalness    liveness    loudness  popularity
 0.01420043  0.86244201  -0.57316177   0.19960617  0.05959667  0.48500504      -0.29675025 -0.07646407  0.45705062  1.00000000
speechiness       tempo
-0.17197872  0.13331015
```

Fig. 1 Pearson Correlation only for population

```
Call:
lm(formula = popularity ~ valence + year + acousticness + danceability +
    duration_ms + energy + instrumentalness + liveness + loudness +
    speechiness, data = train.set)

Residuals:
    Min      1Q  Median      3Q     Max
-63.946  -7.129  -1.384   5.686  69.631

Coefficients:
                  Estimate Std. Error  t value Pr(>|t|)
(Intercept)     -1.287e+03  3.430e+00 -375.248  < 2e-16 ***
valence          4.842e-01  1.668e-01    2.903  0.00369 **
year             6.688e-01  1.715e-03  389.977  < 2e-16 ***
acousticness    -4.242e+00  1.448e-01  -29.294  < 2e-16 ***
danceability     2.800e+00  2.442e-01   11.467  < 2e-16 ***
duration_ms     -4.569e-07  2.584e-07   -1.769  0.07697 .
energy          -1.615e+00  2.602e-01   -6.204 5.51e-10 ***
instrumentalness -4.135e+00  1.162e-01  -35.581  < 2e-16 ***
liveness        -2.961e+00  1.883e-01  -15.729  < 2e-16 ***
loudness         2.500e-02  9.820e-03    2.546  0.01091 *
speechiness     -6.922e+00  2.177e-01  -31.802  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.84 on 119446 degrees of freedom
Multiple R-squared:  0.7538,    Adjusted R-squared:  0.7537
F-statistic: 3.656e+04 on 10 and 119446 DF,  p-value: < 2.2e-16
```

Fig.2 Fitting of the Linear Model

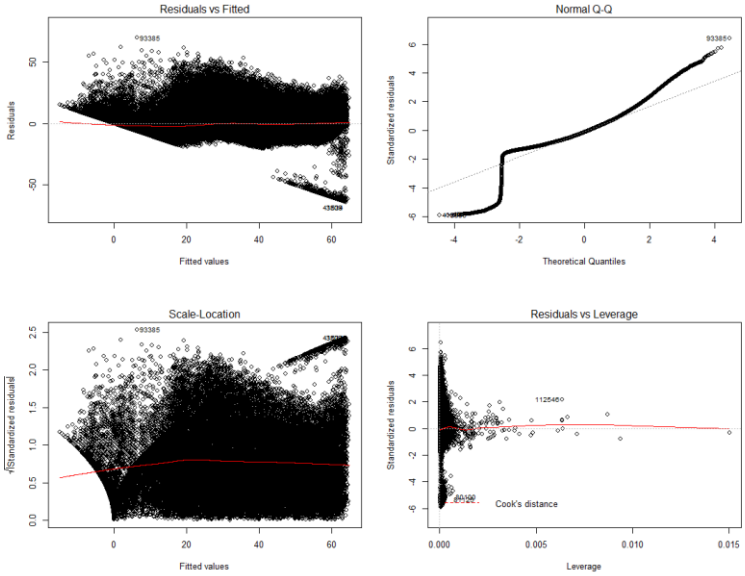

Fig. 3 Diagnostic Plots

Appendix C: Logistic Regression

```
Call:
glm(formula = explicit ~ key + mode + danceability + liveness +
    energy + valence + acousticness + loudness, family = binomial,
    data = train.set)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6070  -0.3851  -0.2129  -0.1144   3.6924

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.173111   0.119005 -26.664  < 2e-16 ***
key1          0.730378   0.047645  15.330  < 2e-16 ***
key2         -0.087172   0.052054  -1.675  0.09400 .
key3         -0.159166   0.085700  -1.857  0.06328 .
key4         -0.159693   0.057875  -2.759  0.00579 **
key5         -0.162382   0.056844  -2.857  0.00428 **
key6          0.399801   0.055155   7.249 4.21e-13 ***
key7         -0.013437   0.049579  -0.271  0.78638
key8          0.312953   0.056884   5.502 3.76e-08 ***
key9         -0.297132   0.054263  -5.476 4.36e-08 ***
key10         0.052569   0.057777   0.910  0.36290
key11         0.390666   0.051919   7.525 5.29e-14 ***
mode1        -0.353730   0.025425 -13.913  < 2e-16 ***
danceability  6.876307   0.093550  73.504  < 2e-16 ***
liveness      1.740816   0.062661  27.782  < 2e-16 ***
energy       -1.206494   0.094558 -12.759  < 2e-16 ***
valence      -2.771273   0.057358 -48.315  < 2e-16 ***
acousticness -2.814527   0.054692 -51.461  < 2e-16 ***
loudness      0.049373   0.004031  12.247  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 69661  on 119456  degrees of freedom
Residual deviance: 51104  on 119438  degrees of freedom
AIC: 51142

Number of Fisher Scoring iterations: 7
```

Fig. 4 Logistic Regression for explicit

```
temp1     0     1
    0 46464  3506
    1   491   735
```

Fig. 5 Confusion matrix

Appendix D: Regression Tree

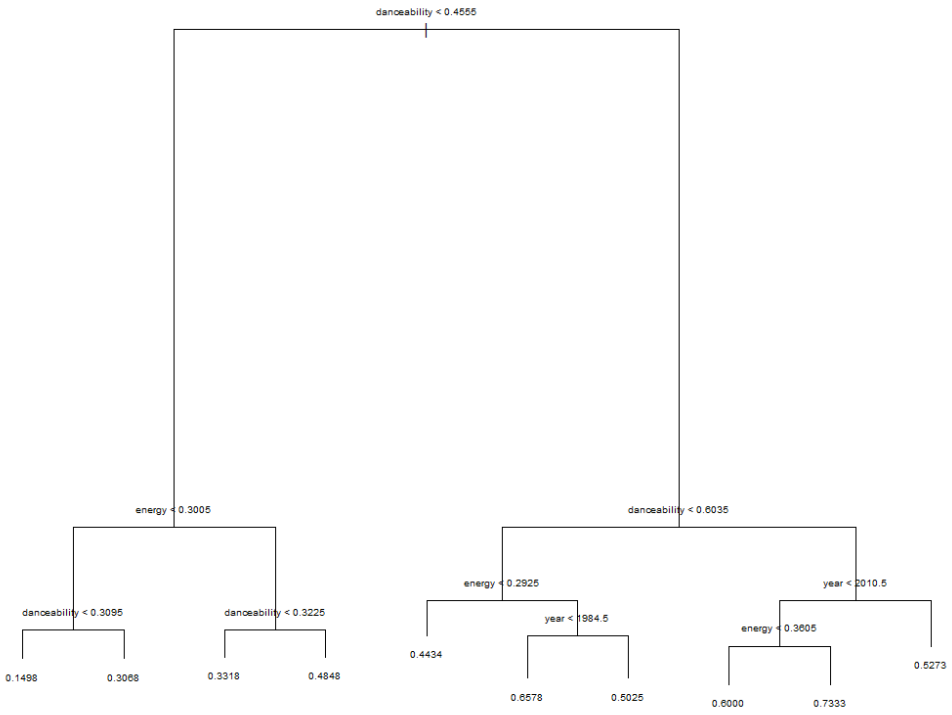

Fig. 6 Regression Tree for Valence

```
Regression tree:
tree(formula = valence ~ ., data = train.set)
Variables actually used in tree construction:
[1] "danceability" "energy"       "year"
Number of terminal nodes:  10
Residual mean deviance:  0.04093 = 4889 / 119400
Distribution of residuals:
       Min.     1st Qu.      Median        Mean     3rd Qu.        Max.
-0.7157000 -0.1375000  0.0002437  0.0000000  0.1497000  0.8192000
```

Fig. 7 Summary of figure 6 tree

```
$size
[1] 10  9  8  7  5  4  3  2  1

$dev
[1] 4904.627 5097.105 5108.966 5542.287 5542.287 5715.045 5995.704 6383.354 8265.990

$k
[1]      -Inf  101.0237  107.9447  144.9520  146.5462  171.6571  279.3059  390.3547 1888.7889

$method
[1] "deviance"

attr(,"class")
[1] "prune"         "tree.sequence"
```

Fig. 8 Cross-validation

Appendix E: Hierarchical Clustering

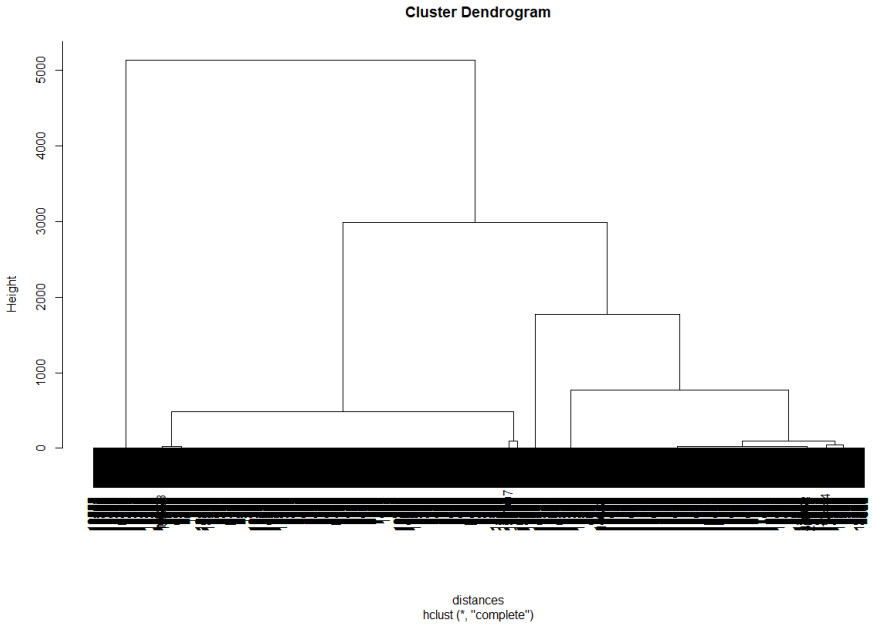

Fig. 9 Hierarchical Tree

```
> clusterGroups[19024]
19024
    2
```

Fig. 10 Cluster of the song "Perfect" by Ed Sheeran

```
> cluster2=subset(new.df,clusterGroups==2)
> cluster2$name[1:5]
[1] "Power Is Power"          "When We Die"             "Korkma Sönmez"           "Bir Zamanlar Bizde Millet" "Love for Guns"
```

Fig. 11 Recommended songs for "Perfect"