**­­Computing Lab and Data Warehousing & BI Project – Milestone IV**

Niti Mishra, Miquel Torrens, Bálint Ván

We are mainly carrying out five different analyses with the One Million Song dataset. Each of them corresponds to each of the questions we initially raised.

1. **Genre differentiation**

We ran Principal Components Analysis (PCA) on selected significant technical features of each song. We had breakdowns of these features per each fraction and segment of the song, so we had to aggregate them per song in order to analyse them. Then, we classified each song to a genre based on the tags that users assign to each song. The principal components take these technical features into account and enable us to plot them in low dimensions, to see how they correlate and how each song lies in this multidimensional space. With that we can see how each genre lies in this space, whether they are grouped together or are distinctly different from other genres. We found that songs from different genres do have their distinct features that define them. However, they also intersect among songs across genres. We can see the analysis in Figure 1 in the annex.

1. **Time evolution of music**

We ran the same PCA as with genre by aggregating the data per decade this time. We plot the same principal components that rely on the technical features but we group the points according to time instead. We found no major distinct difference among music produced over time.

1. **Measuring popularity of songs**

We ran a regression analysis using the technical features of the songs to try to explain how many times each song is listened to. We analysed each of the variables and regressed the most significant technical features standardised on the logarithm of its play count. We also added the genre classification to know if belonging to a particular genre can enhance your average popularity. Some features are actually relevant, but for example genre is not.

1. **Analysing the origin of songs**

We wanted to analyse where music is being produced and whether there is concentration or sparsity of this production across the world. We had to geolocate each of the songs and then compute the country of origin with the coordinates. We used the Google Maps API for that. With these data, we aggregated the data per country and decade and computed the Herfindahl index in each decade until 2000s. We also made a prediction on the production for the current decade to observe the tendency of this concentration. During the last century the concentration went down but currently it seems to be on the rise again. We can see the table on Figure 2 in the annex.

1. **Recommender system**

The recommender system analyses standardised technical features of each song and takes Mahalanobis distances between them to suggest the closest neighbours. These songs are previously filtered using the songs from bands that are usually associated to each other band, according to usage data. Then out of these bands we pick the closest five songs if they are sufficiently close in the space.

**Annex**

Figure 1: principal components analysis on genre of the songs.

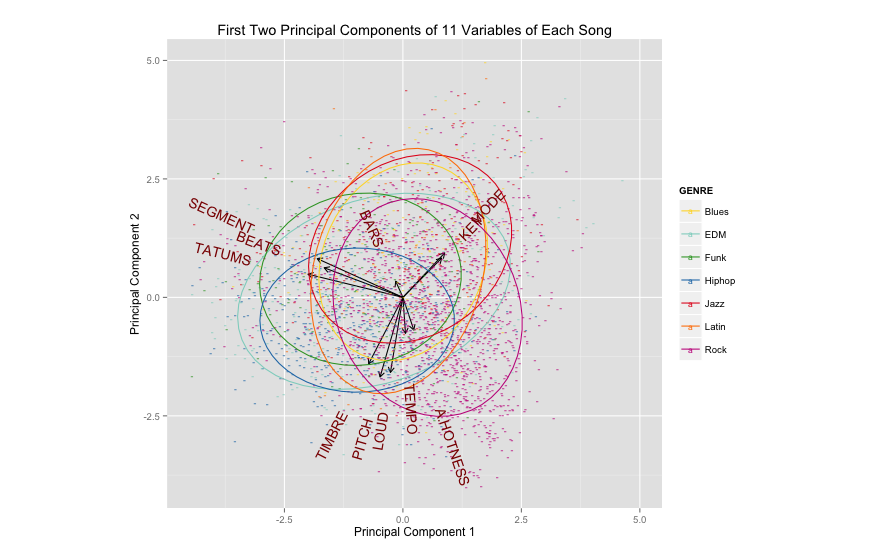


Figure 2: percentages of the top five countries in world production of music.

