Computing Lab and Data Warehousing and BI Project – Milestone II

Niti Mishra, Miquel Torrens and Bálint Ván

Analytics view

We plan to cross-examine relationships between most relevant variables of the *One Million Song Dataset* in the analytics part of our project. The dashboard will be partitioned in different tabs according to these variables, which at the same time will also be interrelated. We use different subsets of the data in each tab to try to tackle some related questions, such as:

1. **Genre Differentiation:** How different are genres in their musical components?
2. **Music Evolution:** How have songs evolved over decades in their components?
3. **Lyrical Significance:** What are the characteristics of the lyrics of the most popular songs? Does the occurrence of certain words in a song imply its musical characteristics?
4. **Popularity:** Are there some key built-in features that contribute to a song's popularity?
5. **Origin:** How do music produced in different places differ?
6. **Recommender System:** We will try to find closest neighbors to songs or bands to suggest likely matches to users' selections.

We will explore in each case the visualization tools (charts, graphs, interactive info-graphics, maps…) best suited to understand the data on hand.

Data view

In the data view of the dashboard, we will display the characteristics of the interest variables from our analytics view. For technological limitations, we are currently working on a random subset of 10,000 songs from the entire dataset. This data is collected from different online music service websites such as the Echo Nest, musiXmatch, Musicbrainz and Last.fm.

A brief description of each variable will be provided in the dropdown lists of our data view. At any given time, we will display only those variables that are required in the respective analysis. Below is the description of the main groups with few example variables that we wish to use:

1. **Musical summary variables**: These are the technical characteristics of the songs calculated out of the sound files. They are exhaustive at song level and complete for all the songs. Duration, energy, key, mode, loudness, tempo, danceability, etc. are some examples of the most relevant variables.
2. **Metadata:** These are obtained from various sources and include information such as artist name, year of release, location of the artist, genre of the song, artist tags, related artists, song "hotness" and artist "hotness", among others. Many observations have missing values on some of the variables. We will either have to find a way to infer these missing values or work with smaller samples.
3. **User data:** Complete play counts available on 39% of the songs broken down by more than one million anonymous users of the Echo Nest.
4. **Lyrics:** Lyrics are represented in a bag of words format. The 5,000 most important words were selected when creating the dataset. We know how many times these words occurred in each song.

We are aware that we may not be able to include all of these variables due to time and resource constraints. Also, despite our interest, some relevant variables may have to be dropped on some analyses depending on the significance of their missing values and integrity of the data source.