Project Description – Computing Lab and Data Warehousing & BI

By: Niti Mishra, Miquel Torrens and Bálint Ván.

Data Set

The data set chosen for the project is the Million Song Dataset, developed for research purposes by the University of Columbia. The data is available in the following link:

- http://labrosa.ee.columbia.edu/millionsong/

This set contains detailed information on the structure of one million songs, technical and otherwise, as well as metadata related to these songs.

There are also related additional data sets, which we will study if they would be beneficial and/or feasible to use in this project, namely:

- Cover songs: http://labrosa.ee.columbia.edu/millionsong/secondhand
- Lyrics: http://labrosa.ee.columbia.edu/millionsong/musixmatch
- Song-level tags and similarity: http://labrosa.ee.columbia.edu/millionsong/lastfm
- User listening data: http://labrosa.ee.columbia.edu/millionsong/tasteprofile

We may work with a subset of the data of about 10,000 songs for logistic reasons, given the amount of information of the full data set, which is about 500 GB.

Objective

We have two main questions that we would like to address with this data set.

The first one is to understand how different kinds of music differ from each other, based on how they are built internally. Are different genres of music really different? How has music evolved over time? How distant is music across generations? Do the topics in the songs make a difference across time and genres? We will try to measure their distance in multidimensional spaces using the detailed data that this dataset can provide to measure all these gaps.

The second one is to spot similarities between songs and bands using a preliminary version of a recommender system. This system is built in a way where those songs and bands that we can spot as very close can be clustered to suggest near neighbours as possible similar candidates for a song or band that the user selects.