NTDS - Peer Review

Sacha Leblanc, Etienne Caquot, Grégoire Mayrhefer, Alexis Mermet

October 2018

NTDS PROJECT: THE RELATIONSHIPS IN THE MUSIC INDUSTRY

1 Introduction of the project

This project has for goal to study the relationships between the multiple actors of the music industry. But we will not only care about well known artist but also to the people working in the shadows (songwriters, producers, etc...).

Our dataset that we will quickly overview later on, could allow us to answer multiple questions about the music industry as:

- Which songwriters/producers are the most relevant in the industry? (Who wrote/produced songs for the most famous artists, who wrote/produced the most popular hits, etc...)
- Does writers/producers always work with the same people? Can we find clusters in working relationships (can we find the label, can friendship be detected, etc...)?
- Does songwriters/producers always work on the same type of songs (can we find genre related clusters in the dataset)?

2 The dataset

We created the dataset using both Spotify and Genius APIs. First we obtained an already created song dataset on Kagle to have a basis of songs to study. But the Spotify API doesn't allow us to obtain informations about the producers/lyricists/songwriters of a song. Thus we had to use the Genius API to request these informations. Right now, our dataset is still pretty raw since it is really slow to pull song informations from genius. But we have already found ways to scrap the data in the way we want to obtain a dataset of the following form:

- track name
- artist name
- genre
- popularity on spotify
- producers
- songwriters
- label (if we want it to compare afterwards to our clusters but this is not really needed in general)

index	track_name	artist_name	genre	popularity
0	Thriller	Michael Jackson	pop	70

songwriters	label	producers
Bruce Swedien	Epic Records	Bruce Swedien

In the end we would like to have approximately 200 000 entries in the dataset for it to be relevant but this will be hard since only the most popular songs (or at least well-known) are well represented in the HTML requests. Indeed the parsing for less known songs is far harder because the HTML data is inconsistent.

3 Data exploration and exploitation

We will create our graph representation in the following way: Our nodes will be the persons working in the industry. The edges will link people that worked together on a song. An edge between two persons will be weighted by the number of songs they have done together and the popularity of the said songs.
We could also create a graph with songs where two songs will be linked if they have a common writer, a graph with only interprets to see which interpret wrote songs for another, a graph considering only popular songs, etc...

In the end , we hope to obtain a small world graph in which we would be able to apply clustering algorithms.