2018's Best Hit Recipe

Jean-Thomas Furrer jean-thomas.furrer@epfl.ch

Benjamin Girard benjamin.girard@epfl.ch

Emily Hentgen emily.hentgen@epfl.ch

Abstract

As an aspiring music composer writing a new song, how much effort should be put into music theory and how much into more matter-of-facts aspects, to maximise the chances it becomes popular? Do some aspects beside the audio features themselves make a song popular? We explore this question in this project, with an analysis based on the Million Song Dataset.

1 Introduction

The Million Song Dataset (MSD) is a collaboration between the LabROSA and The Echo Nest, released in 2011 and freely available. With about 280 GB, it compiles the audio features and metadata for a million popular contemporary tracks.

The goal of this project is to determine whether certain aspects of a song tend to make it popular. To this end, we analyse a few high level features such as the song duration, the artist popularity and country of origin, as well as the music genre. Using the related musiXmatch dataset further allowed us to perform an analysis of the lyrics.

We briefly present some related work in section 2, and explain how the data was collected in section 3. Section 4 gives an overview of the datasets, and the results of the analysis are presented in sections 5. Finally, we conclude in section 6.

2 Related work

Several other projects have been conducted on the MSD, many of them being machine learning-oriented ones, with the task of predicting a next song popularity.

3 Data collection

The analysis is based on three different datasets: the Million Song Dataset, containing records for 1 000 000 contemporary songs, and the musiXmatch dataset, which contains bag-of-words lyrics (due to copyright issues) for 237 662 out of 1 000 000 songs, and the Taste Profile dataset.

We retrieved the data from the whole dataset, made available on the ADA cluster, as single json file. Beforehand, we drop entries we decided to not analyse. This significantly reduces the size of the data, so that the subsequent analysis could be performed locally.

4 Dataset description

4.1 The Million Song Dataset

Each record of the MSD contains 55 fields. For our analysis, we retrieved only a subset of these fields, namely `song id`, `track id`, `title`, `year`, `duration`, `song hotttnesss`, `danceability`, `energy` (song-related features) `artist id`, `artist name`, `artist terms`, `terms weight`, `artist latitude`, `artist longitude`, `artist location` and `artist hotttnesss` (artist-related features) (it turned out however that none of the songs had their `energy` or `danceability` fields defined, i.e. not set to 0.0). What is more, we dropped all entries whose `song hotttnesss` was `NaN`, which means that in the end, we worked on a subset of 553 068 songs.

4.1.1 Release year

The songs in the dataset range over almost a century: the oldest one dates back to 1922, and the most recent one is of 2011. Most songs however were sampled between 1995 and 2010.

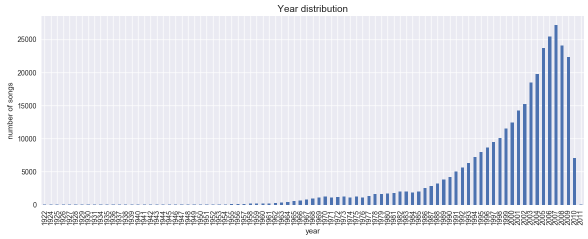

Figure 1: Release year distribution

4.1.2 Song hotttnesss

Our metric for measuring the popularity of a song based on the `song hotttnesss` field of a song, value between 0.0 and 1.0, according to the Echo Nest in December 2010, when the dataset was compiled. Since the `song hotttnesss` fields can be `NaN`, it remains unclear whether a value of 0.0 corresponds to a nonexistent `song hotttnesss`, or if the feature has not been evaluated for this song. In Figure 2, we plot the `song hotttnesss` distribution rounded to

2 decimals, excluding 0 values. The shape of this distribution is rather unusual: the right part of the curve is smooth, while the left one is completely erratic.

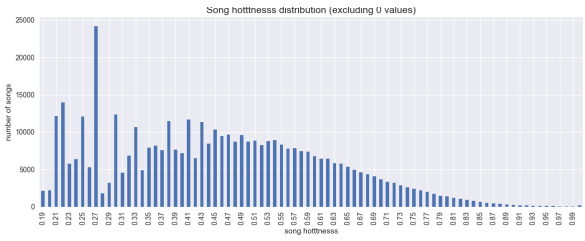

Figure 2: Song hotttnesss distribution

4.2 The musiXmatch dataset

The musiXmatch dataset is a collection of four different files: `full_word_list.txt` (7 MB) containing the complete list of distinct words in the collected lyrics, `mxm_779k_matches.txt` (85 MB) containing the `track id`, `artist name` and `title` for both MSD and the musiXmatch dataset, `mxm_dataset_train.txt` (100 MB) and `mxm_dataset_test.txt` (13 MB), for which each record contains an MSD `track id`, a musiXmatch `track id` and a dictionary mapping word ids of the top-5000 most frequent words to the number of occurrences for this song. We use the train set and test set as single dataset, since our analysis is no prediction task per se.

The correspondence between the MSD and the musiXmatch dataset is made through the `track ids` of a song. We could match 467 366 tracks with the `mxm_779k_matches` dataset (779 056 songs in total), but in the end, out of 467 366 songs, only 162 125 songs could be matched with the concatenation of the `mxm_dataset_test` and the `mxm_dataset_train` datasets. The dataset can directly be downloaded `https://labrosa.ee.columbia.edu/millionsong/musixmatch`.

4.3 The Taste Profile dataset

This dataset contains real triplets such a form : (user, song, play counts) collected from undisclosed partners of the Labrosa laboratory. All songs are matching with those in our main dataset. The collection is about 1'019'318 users, having listened to 384'546 songs, all of this in 48'373'586 triplets. It consists of one single `train_triplets.txt` file (2.79 GB).

5 Results and findings

5.1 Music genre

This analysis is based on the `artist terms` of a song, which can be seen as music genre tags given by the Echo Nest. There are at least 6 482 distinct `artist terms` used in the MSD, including all possible variations such as "rap rock", "funk rock", funk rap", "jazz funk", "jazz rap" or "jazz rock". The most recurrent `artist terms` (e.g. 'rock', 'pop', ...) actually do correspond to music genres, while the least common ones with a low `artist term weights`, used for very few songs, are more like miscellaneous tags (e.g. 'taubertal festival 08', 'guitar hero', 'witty', ...).

In order to determine which music genres are the most popular, we weighted all the `artist terms` of songs with a `song hotttnesss` greater or equal to a certain threshold by their corresponding `artist term weights`. We then normalized the result, to obtain the "proportion" of the most popular music genres. It turns out the dataset is largely dominated by the rock, pop and electronic genre, with rock and pop ranking top, followed by all possible variations.

5.2 Artist popularity

Since rock and pop oriented music tends to showcase the band or the lead singer, it seems natural that the artist must be somehow linked to the "hotttnesss" of the song itself. And indeed, we can observe a correlation between the `song hotttnesss` and the corresponding artist's `artist hotttnesss`. Hot songs tend to be associated to hot artist, but it turns out the hotttessst artists do not make the hotttessst songs. Notice that the gap between 0.0 and 0.2: as mentioned before, almost no song has a `song hotttnesss` below 0.2.

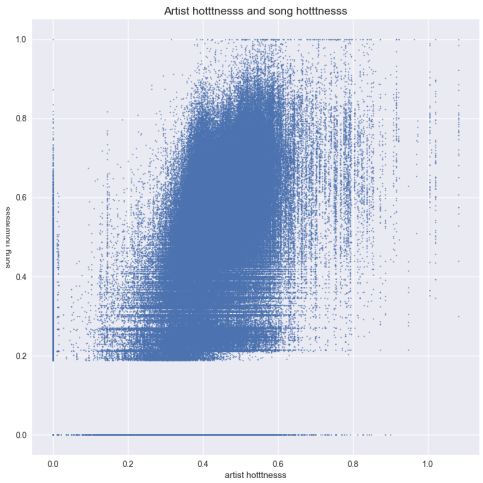

Figure 3: Artist hotttnesss compared to the average song hotttnesss

5.3 Song duration

Despite some outliers - the shortest and longest "songs" are 0.52 seconds respectively 3032 seconds (50min 32s) long - close to 75% (72.64%) of the songs are 2 to 5 minutes long, a very common duration for contemporary music tracks. Figure 4 displays the average song duration for different threshold intervals from 1922 to 2010. Unfortunately, as can be seen by the very irregular values before 1960, there is most likely not enough data before 1960 to make any reasonable conclusion regarding that period. We can nonetheless make the

following observations:

1. The top songs have a duration in a relatively tight interval, between 120 s and 656 s, and an average of 247 s.
2. Over the years, the average song duration of the top songs displays less and less variations
3. In the last years, the average top songs duration dropped below the other average song durations

This suggests that in order to be rated "hot", songs tend to comply more and more to a standard duration.

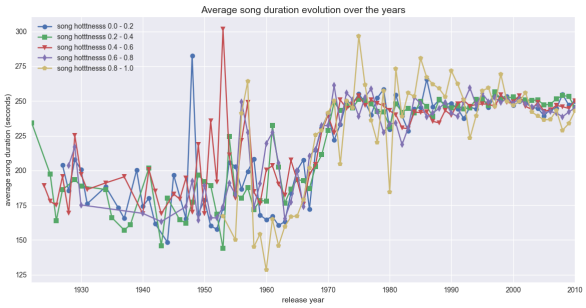

Figure 4: Average song duration evoluation over the years

5.4 Lyrics

Initially, we analysed the lyrics by considering all the words in the top-5000 list of the musiXmatch dataset. In a second step, we exclude the stopwords from the analysis.

5.4.1 With stopwords

The main observation resulting from this analysis is that the frequency of English words in lyrics differ from the frequency of English words in prose, particularly regarding the use of pronouns. While 'the' is the number one word in term of frequency in English prose, it is "only" ranked third for a `song hotttnesss` threshold greater than 0.8, behind 'I' and 'you', which are 11th respectively 14th in English prose. The popular songs hence seem to be rather about 'I', 'you' and 'we', and less about 'they' or 'it'. Although 'it' has a stable ranking of 8th against 11th in English prose, 'they' (ranked 26st in English prose) in fact systematically scores lower than 'we' (ranked 27th in English prose) in the lyrics. One explanation is that the recurrent use of such pronouns increases the level of identification of the listener with the content of the song. In Figure 5, we show the frequency of the English pronouns in the lyrics as a function of the `song hotttnesss` threshold. We also plot the frequency of the stopword 'the', whose evolution we expect does not very much across songs, independently of their popularity, and which can then serve as a baseline. Interestingly, while the frequency of 'I' and 'we' keeps increasing, the one for 'you' suddenly drops past a `song hotttnesss` threshold of 0.9 (note that contractions of pronouns and verbs such as 'you'll', 'you've' etc were not counted).

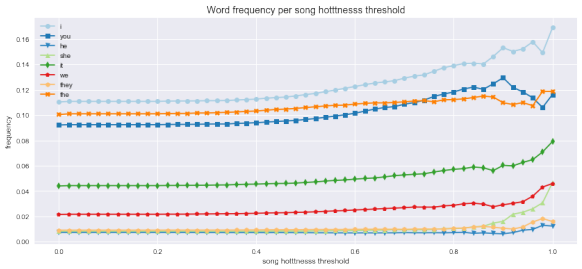

Figure 5: Pronouns frequency for different `song hotttnesss` thresholds

5.4.2 Without stopwords

It is not always obvious to determine whether a word is a stopword or not. We heuristically compiled a list of stopwords using a few freely-available lists of stopwords. When ignoring the stopwords in the ranking, we can point out three characteristics of popular songs (although these are typically not present all at the same time in hot rated songs):

1. The presence of the specific words 'love', 'time' and 'feel', which actually keep ranking among the top places for high `song hotttnesss` thresholds as well as low ones
2. The non-negligible use of the interjections 'yeah' and 'oh'
3. The use of informal speech like 'gonna', strengthened by the fact that it is often found together with 'yeah'

This analysis is nonetheless somewhat limited by the bag-of-words nature of the musiXmatch dataset. Still, while 'yeah', 'oh' or 'gonna' could actually be considered as stopwords, it is interesting to see they are the second "type" of words to rank top after the stopwords compiled in our list.

5.4.3 Language and artist's country of origin

We matched the `artist latitude` and `artist longitude` of 124 216 songs with valid coordinates to a country using reverse geolocation. Among these songs, there is an unequal repartition between the different countries. An overwhelming majority of songs indeed comes from the United States (78 129), followed by the United Kingdom (14 093) and Canada (4399), but the number goes down to a single song for Martinique. This means that more than 60% of the songs come from countries of Anglo-Saxon culture (United States, United Kingdom, Canada, Australia, ...). As a consequence, the vast majority of lyrics are written in English, so that words of different languages tend to be eclipsed in the analysis. Even after removing the English stopwords, the top ranked words are still English words.

However, we also suggest this does not simply reflect the fact that much more songs were sampled from these countries, but also a general trend which can be observed when it comes to popular songs. In fact, there

tends to be a mismatch between the country of origin of the artist and the language in which the lyrics were written. By checking the top songs from non Anglo-Saxon countries such as Sweden, Denmark or Greece, it turns out the lyrics are actually written in English, not in Swedish, Danish or Greek.

5.5 User-song plays

A first step to do with the Taste Profile dataset is to compare the number of times a song has been listened to its song hotttnesss. To this purpose, we collected, for each song, the number of time it was heard over all users. As we said, there is 384 546 such record, going from only one play to 726 885 plays (max). The mean is only at 360 plays, with a standard deviation of $\approx$3256. As we can see in Figure 6, it follows kind of a long tail distribution.

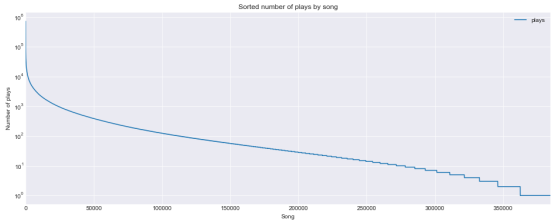

Figure 6: Users connected by the songs they heard in common

Then we normalized those values to get something comparable to the song hotttnesss attribute and 'inner' merged them to the original data.

Hopefully, over the 384 546 songs in the Taste Profile dataset, only 27% could have been merged. After removing 3 data points which can be guessed as outliers, we obtain a result not so convincing. Figure 7 suggest a small correlation. Actually we obtain a correlation of 0.18, which is not significant. Yet since $\approx$ 73% of the data lacks, we can expect a much significant correlation with the whole dataset. Hence, we can not conclude without further analysis here.

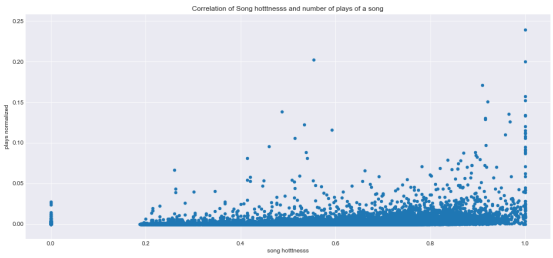

Figure 7: Users connected by the songs they heard in common

5.5.1 Clustering in users' taste

Are musical tastes clustered or not ? This is the question we wanted to address in this section. First of all, we had a look at the connectivity for a small subset. For this, we took a set of 30 users (and all their plays). We see in Figure 8 how the users are linked together through the songs they played. Even though the graph seems connected, a few ones (7 over 30) are not connected. For further analysis, we simplified this graph in the following way : two users are connected by an edge if they listened to the same song at least once. This can also be viewed as having an edge between any pair of users (red dots) whose shortest path is of length two.

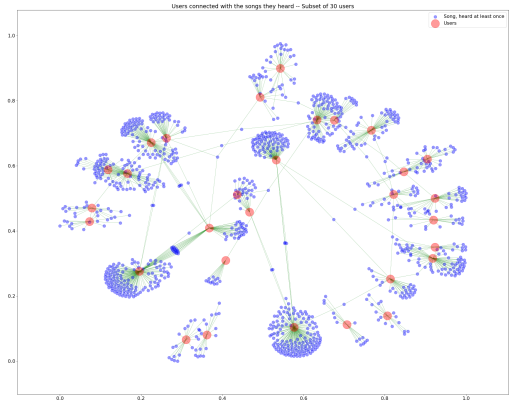

Figure 8: Users connected by the songs they heard in common

Of course when applying this to 1 019 318 users, the result is not the same. We only obtain one giant connected component. So we tried to apply the Louvain method, which is used to unfold communities

6 Conclusions

In this project, we explored several song features using the Million Song Dataset. The aim was to determine to which extent they were linked to the popularity of a song. Regarding the music genre, focusing on rock or pop seems to have the advantage of reaching a large audience. Moreover, since the most popular songs are not symphony-like compositions, the song duration should be from 2 to 7 minutes, preferably around 4 minutes (after all, fitting on the radio is a plus for promotion). Concerning the artist popularity, not being "hot" will not significantly impact the chances to produce "hot" song: not-so-hot artists have as much chances to produce a "hot" song that the "hotttessst" ones, and the "hotttessst" songs are not necessarily produced by the "hotttessst" artists. Similarly to the music genre, English should be the language of choice when it comes down to the lyrics, as it will reach a wider audience. What is more, making it about 'I & you', or 'we' and raises the chances of winning the hearts of your future listeners.

References

[MSD official documentation] https://labrosa.ee.columbia.edu/millionsong

[List of stopwords] https://www.ranks.nl/stopwords http://www.lextek.com/manuals/onix/stopwords1.html.