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

1 Introduction

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

In this project, we analysed 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

3 Data collection

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

3.1 The Million Song Dataset

Initially, we gained insight into the dataset by working with the official MSD subset which covers 10 000 songs, freely available for download at `https://labrosa.ee.columbia.edu/millionsong/pages/getting-dataset`. We then moved to the full dataset of 1 000 000 songs; however, the analysis only uses a subset of it, because, because we drop entries with undefined value for the `song hotttnesss` (`NaN`). In the end, we worked on a subset of XXX XXX songs.

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.

3.2 The musiXmatch dataset

The musiXmatch dataset can directly retrieved from `https://labrosa.ee.columbia.edu/millionsong/musixmatch` There are 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 154 663 tracks with the `mxm_779k_matches` dataset (779 056 songs in total), but in the end, out of 159 430 songs, only 60 146 songs could be matched with the concatenation of the `mxm_dataset_test` and the `mxm_dataset_train` datasets.

4 Dataset description

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`.

4.1 Release year

The MSD was put together in 2010 The oldest songs in the dataset dates back to 1925, but most songs were sampled between 1995 and 2010.

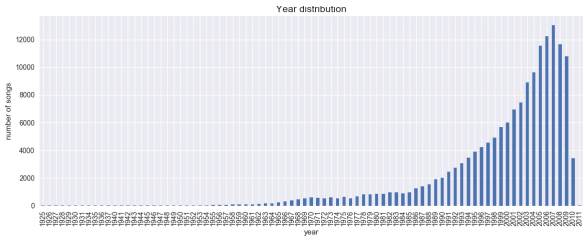

Figure 1: Release years distribution

4.2 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.

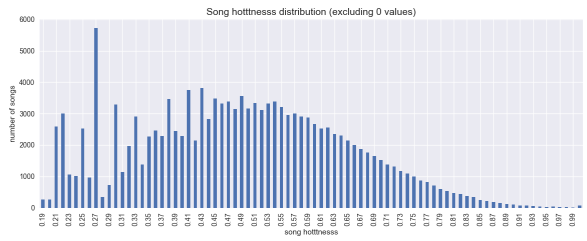

Figure 2: Song hotttnesss distribution

4.3 Artist's country of origin

Matched coordinates to country using reverse geolocation (`pygeocoder` and `pygeolib` packages) (35 106 songs with valid coordinates)

Among those songs, there is an unequal repartition between the different countries. An overwhelming majority of songs comes from the United States (20 276), followed by the United Kingdom (4 831), but the number goes down to a single song for Afghanistan, China and Bolivia.

5 Results and findings

5.1 Artist popularity

There is 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.

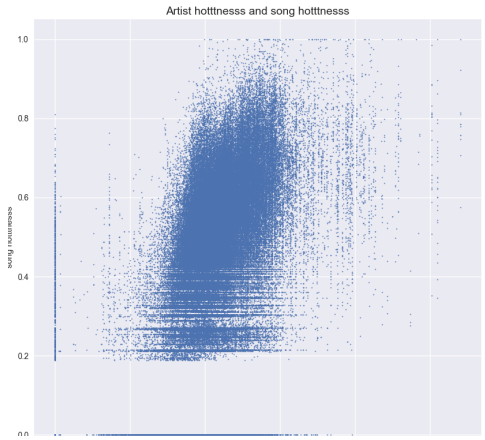

Figure 3: Artist hotttnesss compared to the average song hotttnesss

5.2 Song duration

Despite some outliers - the shortest and longest "songs" are 0.67 seconds respectively 3007 seconds (50min 7s) long - close to 75% (73.47%) of the songs are 2 to 5 minutes long, a very common duration for contemporary music tracks.

Figure 5 displays the average song duration for different threshold intervals from 1925 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:

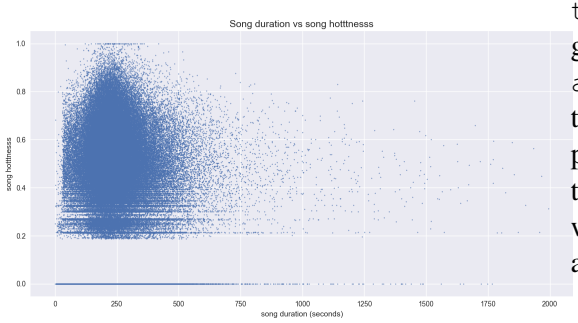

Figure 4: Song duration against song hotttnesss

1. The top songs have a duration in a tight interval
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.

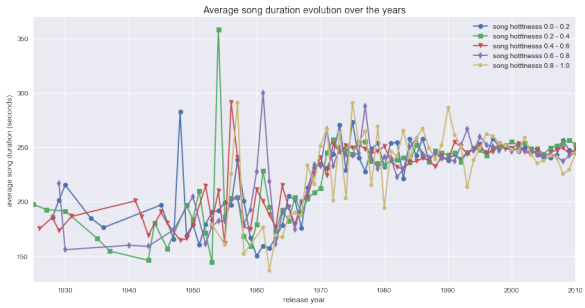

Figure 5: Average song duration evoluation over the years

5.3 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”. Actually, the most recurrent `artist terms` do correspond to music genres, while the least common ones, used for a single song for instance, are more like miscellaneous tags. The dataset is largely dominated by the rock, pop and electronic genre.

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 0.9 by their corresponding `artist term weights`. We then normalized the result, to obtain the ”proportion” of the most popular music genres. For a lower threshold, the ranking is essentially very similar: only the weights decrease, since more songs are included, and hence more `artist terms` as well.

`artist_terms`	weight (%)	occurrences
rock	4.617761	426
pop	3.738941	409
alternative	2.493392	264
indie	2.221263	256
alternative rock	2.011225	188
guitar	1.987991	308
indie rock	1.904490	163
acoustic	1.577093	231
electronic	1.488118	182
soundtrack	1.482128	172
soul	1.328866	180
classic rock	1.181799	116
punk	1.166998	149

Table 1: Top-12 `artist_terms` (weighted) for a `song_hotttnesss` threshold of 0.9 (500 songs), and a minimum number of occurrences of 0

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 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.

In Figure 6, 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.

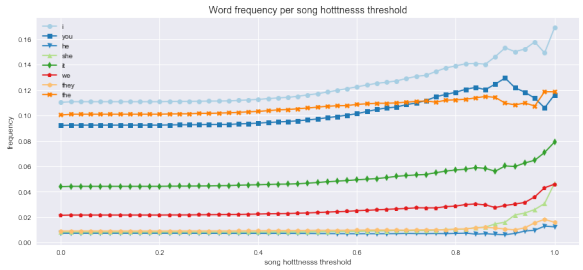

Figure 6: Pronouns frequency for different `song hotttnesss` thresholds

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 the freely-available lists of stopwords at `https://www.ranks.nl/stopwords` and `http://www.lextek.com/manuals/onix/stopwords1.html`. When ignoring the stopwords in the ranking, we can point out three characteristics that we tend to encounter in 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' (although the frequency of 'oh' drops above a threshold of 0.9)
3. The use of informal speech like 'gonna', strengthened by the fact that it is often found together with 'yeah'

While 'yeah', oh' or 'gonna' could actually also be considered as stopwords, it is interesting to see they are the second "type" of words to rank top, after the stopwords.

5.4.3 Language and country of origin

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. The country of origin of the artist indeed does not necessarily match 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. A few of the top songs have Spanish lyrics, but these songs are still a minority though.

6 Conclusions

MSD: unequal distribution of songs regarding countries musiXmatch limitations: bag-of-words (due to copyright issues)

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`.