MARTIN-LUTHER
UNIVERSITÄT
HALLE-WITTENBERG

# Analyzing the Million Song Dataset

Berg, Dmitrij

dmitrij.berg@student.uni-halle.de

Reimer, Jan Heinrich

jan.reimer@student.uni-halle.de

July 9, 2018

Martin-Luther-Universität Halle-Wittenberg
Institute of Computer Science
Big Data Analytics

| | |
|---|---|
| **Lecturer:** | Prof. Dr. Matthias Hagen |
| **Semester:** | Summer 2018 |
| **Students:** | Dmitrij Berg, Jan Heinrich Reimer |
| **Matriculation numbers:** | 216226012, 216204166 |
| **Version:** | 1.0.0 |
| **Date:** | July 9, 2018 |

## Statutory declaration

The authors of this paper, Dmitrij Berg, born on 16.03.1998, and Jan Heinrich Reimer, born on 25.01.1998, declare that they have authored this thesis independently, that they have not used other than the declared sources/resources, and that they have explicitly marked all material which has been quoted either literally or by content from the used sources.

Halle (Saale), July 9, 2018

Date

Signatures

# Contents

# 1. Introduction

## 1.1. Motivation

Out of the many applications of big data analytics, from news headlines to video streaming metrics to the human DNS, the subject of Music Information Retrieval (MIR) is probably one of the most interesting.

Everyone can easily identify with the relevance of the MIR as music accompanies us in everyday life. May it be learning an instrument or mixing music. Even just listening to music leads to a natural personal interest to learn what features distinguish songs we like from those we don't like.

# 2. Data

"The The Million Song Dataset (MSD) is a freely-available collection of audio features and meta data for a million contemporary popular music tracks."(Bertin-Mahieux 2012b) It contains songs dated from 1922 to 2010 and was collected 2010 in an effort to make a dataset of close to commercial size available to researchers (Bertin-Mahieux et al. 2011, p. 591).

That comprehensive collection of titles, artists, publishing year etc. — with direct connections to further datasets such as lyrics or genres — seems to be the optimal data source for music-based investigation.

## 2.1. Structure

The dataset contains 1 000 000 track files. Each one is stored in the HDF5 format (The HDF Group 2018) and consists of the fields as in Bertin-Mahieux (2012a). Along structural information on the song's beats, sections etc. the MSD also provides song meta information fetched from The Echo Nest as described by Bertin-Mahieux et al. (2011, p. 592).

## 2.2. Additional datasets

One major advantage of the MSD is that it provides easy access to additional third party data sets. Many projects already exist, that build upon the MSD and provide additional data.

### 2.2.1. The musiXmatch lyrics dataset

While songs lyrics seem to be the most relevant information source to interpret the artist's intentions and feelings that data is generally rarely open to the public and protected by copyright restrictions.

Musixmatch, formerly called musiXmatch, "is the world's largest lyrics platform"(Musixmatch 2018). According to Bertin-Mahieux (2012d) the musiXmatch lyrics dataset (MXM) provides bag-of-words lyrics for 237,662 tracks of the MSD. Specifically when performing summarizing/reducing tasks on a big dataset like the MSD that bag-of-words structure turns out to be equally useful as the full lyrics.

### 2.2.2. The tagtraum genre annotations dataset

The MSD itself "does not contain readily accessible genre labels. Therefore, multiple attempts have been made to add song-level genre annotations"(Schreiber 2015a, p. 241). The tagtraum genre annotation dataset (TTG) for the Million Song Dataset benefits from combining multiple data sources, namely the Last.fm dataset (Bertin-Mahieux 2012c), the Top-MAGD dataset (Schindler, Mayer, and Rauber 2012) and the beaTunes Genre Dataset (Schreiber 2015a, p. 241) as described by Schreiber (2015b, p. 244), to provide a reasonable good accuracy.

# 3. Technical implementation

For the purpose of analyzing the MSD the Hadoop map-reduce framework (The Apache Software Foundation 2018), running on a single-node cluster on the Java Virtual Machine, was chosen. It scales up to big datasets very well and is very flexible for doing different map-reduce tasks. The mappers, reducers, data- and other classes were written in the Kotlin language (Jetbrains 2018).

## 3.1. HDF5 song file input format

A custom Hadoop input format and a serializable `Song` data class had to be written to support parsing respectively storing the HDF5 song files in a format usable in Hadoop. Moreover that `Song` class has the advantage of conveniently accessing the fields of the HDF5 files, especially the arrays, in the Hadoop mappers and reducers without having to access the HDF5 file directly.

## 3.2. Heatmap generation

For the purpose of creating the heatmaps in Figure 3 and Figure 4 a flexible Kotlin program was written that could read in the Tab-separated values (TSV) output by Hadoop and draw the heatmap on a predefined background.

## 3.3. Challenges

### 3.3.1. HDF5 native library

During setting up the Hadoop cluster, including the HDF5 data input format, internally accessing the native HDF5 library, turned out to be difficult.

As normally each Hadoop data node is running on a separate physical machine, the mappers running on the data nodes can not access files on the name node. Thus one needs to configure the native library on each data node separately. Writing custom wrappers around the Hadoop analyze tool and using Hadoop file system's shared cache it was possible to share the HDF5 native library across the data nodes.

### 3.3.2. Single-node cluster

Setting up Hadoop to run on a single-node cluster of course is not the supposed usage of the Hadoop architecture. Though, as there was no multi-node cluster available to the authors at the time of writing, this project was limited to a single-node cluster running on a standard consumer laptop.

For each analyzed input file Hadoop creates a mapper. Although normally each mapper can process multiple splits (regions) of a file, the custom input format made for HDF5 songs doesn't support splits — neither does the underlying HDF5 library. Consequently Hadoop would have to create 1 000 000 mappers (10 000 for the subset) which of course overflows the RAM on most consumer hardware. Therefore only parts of the subset — the `B` sub-folder containing a total of 2380 song files — were analyzed.

Speaking of hardware limitations, apart from memory usage the used scripts don't use up all system resources such as CPU. HDD access rates seem to be the main speed-limiting factor for the mappers themselves.

Table 1: The three most common stemmed English nouns per genre, without normalization, taken from Table 5.

| Genre | Most common words |
|---|---|
| Blues | alley, love/place/ride/road/train |
| Country | babi, love, jingl |
| Electronic | adventur, disgrac/ritual |
| Folk | heart/hope, citi |
| International | call/ride, countri/eye |
| Jazz | doctor, time, life/love |
| Latin | danc, time, faith |
| Pop / Rock | love, time, way |
| Rap | light/shit, music |
| Reggae | road, pictur, man |
| R'n'B | call, stori, babi |

# 4. Studies

## 4.1. What are the main topics of different genres?

When hearing music of some specific genres one sometimes could get the impression that all songs are about the same topics, mostly using the same keywords. In the following two different approaches on interpreting raw word counts output by the Hadoop map-reduce tasks are being discussed.

### 4.1.1. Most common nouns

To get a good understanding of what topics are relevant in a given set of words a good estimation is to look at the nouns contained in the set. Table 5 shows the most common English nouns for each genre's lyrics. Based on that list Table 1 shows a summarized view of the three most commonly used words in song lyrics for each genre. Here words separated by slashes (/) occurred equally often.

Interestingly some stereotypes could therefore be disproved while some others were confirmed by the data: Love is a consistent topic across many of the examined genres, including Blues, Country, Folk ("heart"), Jazz, Pop/Rock and R'n'B ("babi"). And this strongly conflicts the stereotype that Blues would be all about sadness, that is also supported by the fact that travelling ("place", "ride", "road",

"train") is another frequent topic in Blues music. While other genres' topics are as expected, like Latin music being about dancing ("danc") and religion ("faith") — most of Latin America is Christian religious (Pew Research Center 2014) — or International music being about countries ("counti") and communication ("call"), some other genres are dominated by words one would not expect. For instance, Table 1 shows that rituals ("ritual") are a frequent topic in Electronic music.

#### 4.1.2. Word count normalization

Just looking for nouns certainly is not sufficient enough for being able to determine the main topic as it excludes words that might have a far more relevant meaning as only the nouns, like verbs or adjectives. Also nouns that could likewise be verbs or adjectives are overweighted in the results.

However that is not a trivial task as filling words, foreign words, names or punctuation clutter the word lists for each genre. For instance the lyrics of Latin music contain a lot of Spanish words.

To counteract for that filling words, foreign words, names and punctuation were filtered out the list:

> &, a, all, am, an, and, aqui, are, be, been, but, ca, con, could, de, dem, denk, do, e, el, ella, en, es, esta, estou, for, get, got, had, have, he, her, hey, i, ich, if, ihr, in, is, it, just, la, let, me, mi, michael, mil, muy, my, nicht, no, not, não, o, of, oh, on, por, que, s, se, será, she, sie, so, su, t, te, that, the, they, this, to, tu, un, und, was, we, whi, will, with, would, y, you, your

After filtering these words genres were filtered out if they had no word left that occurred 30 times or more often:

> International, Latin, Jazz, Folk, Blues

Note that the filtered words as well as the minimal word count required for the further interpretation were chosen somewhat randomly by the authors. That said, changing these parameters may have a great effect on the results. Nonetheless this approach was considered good enough as there's not much algorithmic assistance available to adequately filter out filling words.

#### 4.1.3. Most common words

After filtering Table 6 shows the most common English words for each genre's lyrics. Table 2 summarizes the three most commonly used words for each of these genres. Again slashes (/) denote words that occurred equally often.

Table 2: The three most common stemmed English nouns per genre, after normalization, taken from Table 6.

| Genre | Most common words |
|---|---|
| Country | babi, like, love |
| Electronic | there, go, now |
| Pop / Rock | know, never, love |
| Rap | die, like, up |
| Reggae | road, pictur, like/man/while |
| R'n'B | call, feel, stori |

This view on the data shows a much more diverse result as verbs and presumably important words support the meaning of main topics and especially interpreting it.

The topic of love in contrast to Table 1 doesn't outweigh other equally important topics. Rather uncommon words like "disgrac" or "ritual" don't dominate lyrics of Electronic music anymore, instead Table 2 shows that in this genre things are not as important as activities ("go", "now"). In Pop/Rock music the word "never" comes to attention. This may denote desires or unfulfilled dreams, maybe related to "love". The most commonly used words in the genres of Country and R'n'B remain almost unchanged, main topics being feelings ("feel", "like") and love ("babi", "love", "call").

## 4.2. Where do the most familiar artists live?

In the analyzed subset there were 1719 artists with known familiarity, of which 639 have a location linked to their name in the MSD.

An artist's familiarity score can be understood as the likelihood that a randomly selected person recognizes them. On the map an artist's location resembles their last known residence.

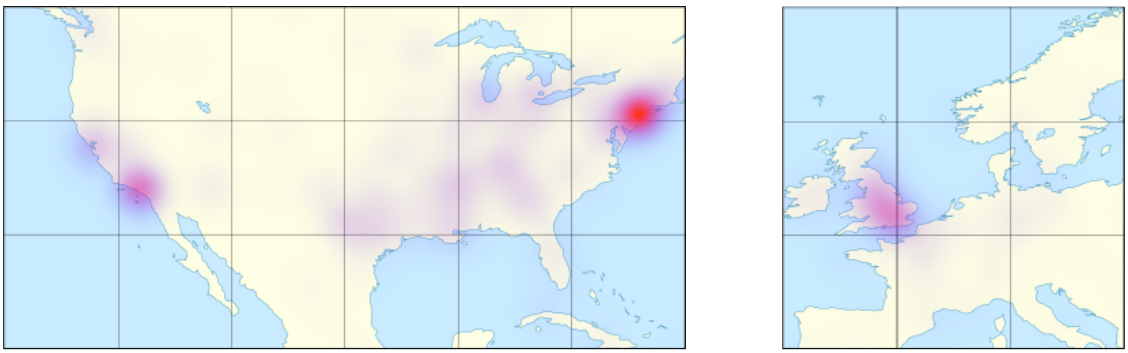
A lot of successful artists are known to travel around the globe in order to create their music, however assuming that most of an artist's impact on the music industry is bound to where they live, it would be interesting to see if there are regions of the world where exceptionally many artists with a high familiarity score pile up.

Figure 3 shows a world map on which for each coordinate the familiarity score

Table 3: Artists and locations with the highest average familiarity. See Table 7.

| Rank | Artist name | Average familiarity | Location |
|---|---|---|---|
| 2. | Lil Wayne | 0.990 | New Orleans, LA, USA |
| 3. | Britney Spears | 0.947 | Los Angeles, CA, USA |
| 4. | Avril Lavigne | 0.942 | Belleville, ON, Canada |
| 7. | Alicia Keys | 0.934 | New York City, NY, USA |
| 8. | Muse | 0.929 | UK |
| 9. | Slipknot | 0.929 | Des Moines, IA, USA |
| 16. | Taking Back Sunday | 0.912 | Long Island, NY, USA |
| 20. | Daddy Yankee / Bounty Killer | 0.909 | Puerto Rico |
| 22. | Natasha Bedingfield | 0.901 | London, UK |
| 26. | The Smashing Pumpkins | 0.888 | Chicago, IL, USA |
| 27. | Killswitch Engage | 0.888 | Boston, MA, USA |

Figure 1: Artist's familiarity for selected regions. See Figure 3.



(a) for the USA

(b) for the UK

of artists living there is summed up using a lookup table that maps artist ID from the MSD to their last known residence's coordinates. While most of the world is empty, the USA and UK make up the most part of the familiarity score. For better visibility Figure 3 has been cropped to only show the USA (Figure 1(a)) respectively the UK (Figure 1(b)).

It turns out that big cities attract big personalities. New York City stands out as a glowing red spot, due to the vast number of instantly recognizable musicians originating in "the big apple".[1] Artists like Jay-Z, Lady Gaga, 50 Cent, Alicia Keys and KISS — just to name a few — live there (Ranker, Inc. 2018d). As shown in Table 7, Alicia Keys (7., 0.934) and 50 Cent (29., 0.883) are among the top 30, though only Alicia Keys' location is known in the MSD (see Table 3). And with a place that is known for attracting massive amounts of people by its tourist attractions and therefore is well known (NYC & Company, Inc. 2018), the high familiarity in this region is quite obvious.

Los Angeles is another hot spot on the map. Artists like P!NK, Miley Cyrus, Metallica, Guns'n'Roses (58., 0.844) or Red Hot Chili Peppers (78., 0.834) (Ranker, Inc. 2018b) might be essential to this bump in familiarity. Table 3 includes the Pop icon Britney Spears (3., 0.947) as a familiar representative for Los Angeles, contributing to the clearly visible spot on the map (Figure 1(a)).

The city of Nashville is frequently called "music city USA". It achieved this nickname due to its vivid western and Country music scene (Harper, Cotton, and Benefield 2013). Naturally the "music city" should be home to popular artists. Some of them like Kesha and Paramore are well known (Ranker, Inc. 2018c). The comparatively low overall familiarity of Nashville can be explained by the steadily decreasing appeal of Country music worldwide.

Outside the US there is another colored spot on the world map, located in the UK, centered on London. That city features popular artists and bands like The Rolling Stones (102., 0.815), Led Zeppelin (139., 0.787) and Amy Winehouse (Ranker, Inc. 2018a). Table 3 includes Natasha Bedingfield (22., 0.901) as a London-based very familiar musician.

## 4.3. Where do the artists with the hotttessst songs live?

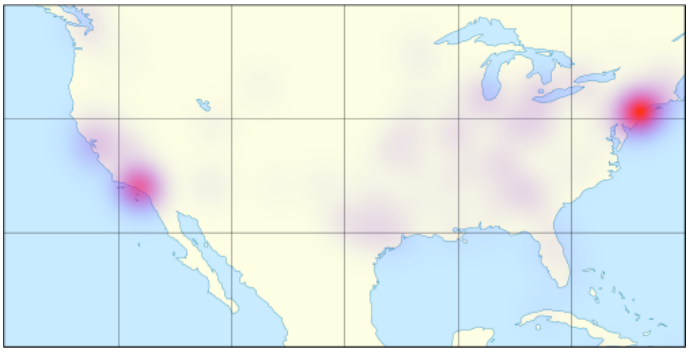
A song's hotttnesss can be understood as the amount of online traffic, news headlines and listeners it recently (2010, when the MSD was created) generated. Natu-

[1] That nickname was frequently used by Jazz musicians in the 1930s.

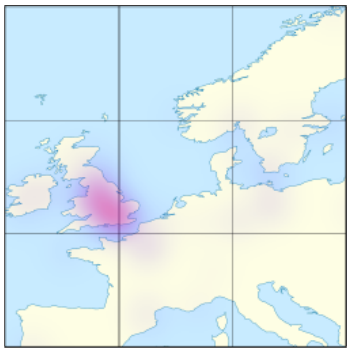
Table 4: Artists and locations with the highest average song hotttnesss. Only the artists with known location with the highest 10 scores are shown. See Table 8.

| Rank | Artist name | Average song hotttnesss | Location |
|---|---|---|---|
| 2. | The White Stripes | 0.972 | Detroit, MI, USA |
| 4. | Nickelback | 0.910 | Vancouver, BC, Canada |
| 6. | Public Image Ltd | 0.874 | London, UK |
| 10. | Britney Spears | 0.840 | Los Angeles, CA, USA |
| 11. | Lighthouse Family | 0.821 | Newcastle upon Thyne, UK |
| 21. | The Radio Dept | 0.788 | Lund, Sweden |
| 24. | Motograter | 0.771 | Santa Barbara, CA, USA |
| 26. | Lupe Fiasco | 0.764 | Chicago, IL, USA |
| 29. | Strata | 0.755 | San José, CA, USA |
| 30. | Arsonists Get All The Girls | 0.755 | Santa Cruz, CA, USA |

Figure 2: Artist's average song hotttnesss for selected regions. Only the artists with known location with the highest 10 scores are shown. See Figure 4.



(a) for the USA



(b) for the UK

rally it is likely that popular and well-recognized artists will generate awareness on their songs easier than newer, unknown artists as they already have more followers. Their songs will create more headlines, have more listeners and result in a higher hotttnesss score.

A quick comparison of Figure 4 and Figure 3 shows little to no difference. Presumably dominant spots on the world resemble "popularity-centers" which unite 2010's most important songs.

Lamere (2009), the Director of Developer Platform at The Echo Nest, the company associated with creation of the MSD (Bertin-Mahieux et al. 2011, p. 591), wrote about plotting each artist's familiarity against their average song hotttnesss to detect upcoming popular artists, hence when their song's hotttnesss exceeds its author's familiarity. The resulting plot shows that an artist's song hotttnesss is nearly proportional to their familiarity, with only few exceptions. These become less likely to show up, the less data is shown.

Again the world map in Figure 4 has been cropped to only show the USA (Figure 2(a)) respectively the UK (Figure 2(b)) for better visibility.

An example for a place that has less hottt songs than familiar artists is Nashville, Tennessee, USA. It doesn't show up in Figure 2(a), meaning its popular musicians are losing listeners.

## 5. Conclusions

Analyzing just a small set of 2380 songs contained in the B sub-folder of the MSD subset, some results may not be as precise because there's a greater chance of

statistical outliers to cause false interpretations.

## 5.1. Analyzing the whole dataset

A significant increase in size of the input data should slightly change the appearance of both Figure 3 and Figure 4 as well as improving meaningfulness of the lyrics analysis. There would remain less “empty” space on the maps and potentially other regions of the world could form hot spots on the map. The differences between familiarity and song hotttnesss could also become more clear. Less normalization would be required to get satisfactory results when analyzing lyrics.

With the availability of a performant cluster analyzing the whole MSD would be possible without much modification of the scripts implemented for this thesis. Given the MSD’s size of 1 000 000 and the analyzed subset’s size of 2380 song files a cluster of around 400 customer laptops would suffice to answer the above questions for the whole MSD. Of course that number of data nodes could be lowered by using hardware with more memory capacity.

# Acronyms

**MIR** Music Information Retrieval. 3

**MSD** The Million Song Dataset. 3, 4, 8–11

**MXM** musiXmatch lyrics dataset. 4

**TSV** Tab-separated values. 4

**TTG** tagtraum genre annotation dataset. 4

# List of Tables

# List of Figures

## References

Bertin-Mahieux, Thierry (2012a). Field list - Million Song Dataset. Accessed on 2018-07-06. URL: https://labrosa.ee.columbia.edu/millionsong/pages/field-list.

– (2012b). Million Song Dataset, official website. Accessed on 2018-07-06. URL: https://labrosa.ee.columbia.edu/millionsong/.

– (2012c). The Last.fm Dataset - Million Song Dataset. Accessed on 2018-07-08. URL: https://labrosa.ee.columbia.edu/millionsong/lastfm.

– (2012d). The musiXmatch Dataset - Million Song Dataset. Accessed on 2018-07-08. URL: https://labrosa.ee.columbia.edu/millionsong/musixmatch.

Bertin-Mahieux, Thierry et al. (2011). "The Million Song Dataset". In: Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR 2011), pp. 591–596.

Gaba, Eric (2008). Map of the world in an equirectangular projection with Tissot's Indicatrix of deformation. Ed. by Wikimedia Commons, the free media repository. Accessed on 2018-07-07. URL: https://commons.wikimedia.org/w/index.php?title=File:Tissot_indicatrix_world_map_equirectangular_proj.svg&oldid=203837838.

Harper, Garrett, Chris Cotton, and Zandra Benefield (2013). Nashville Music Industry. Impact, Contribution and Cluster Analysis.

Jetbrains (2018). Kotlin Programming Language. Accessed on 2018-07-08. URL: https://kotlinlang.org/.

Lamere, Paul (2009). Hottt or Nottt? Accessed on 2018-07-08. URL: https://musicmachinery.com/2009/12/09/a-rising-star-or/.

Musixmatch (2018). About. Accessed on 2018-07-08. URL: https://about.musixmatch.com/.

NYC & Company, Inc. (2018). Mayor de Blasio and NYC & Company Announce New York City Welcomed Record 62.8 Million Visitors in 2017. Accessed on 2018-07-08. URL: https://www1.nyc.gov/office-of-the-mayor/news/146-18/mayor-de-blasio-nyc-company-new-york-city-welcomed-record-62-8-million-visitors-in.

Pew Research Center (2014). "Global Religious Diversity. Half of the Most Religiously Diverse Countries are in Asia-Pacific Region". In: URL: https://commons.wikimedia.org/w/index.php?title=File:Tissot_indicatrix_world_map_equirectangular_proj.svg&oldid=203837838.

Ranker, Inc. (2018a). List of Famous Bands from London. Accessed on 2018-07-08. URL: https://www.ranker.com/list/london-bands-and-musical-artists-from-here/reference.

– (2018b). List of Famous Bands from Los Angeles. Accessed on 2018-07-08. URL: https://www.ranker.com/list/los-angeles-bands-and-musical-artists-from-here/reference.

– (2018c). List of Famous Bands from Nashville. Accessed on 2018-07-08. URL: https://www.ranker.com/list/nashville-bands-and-musical-artists-from-here/reference.

– (2018d). List of Famous Bands from New York. Accessed on 2018-07-08. URL: https://www.ranker.com/list/new-york-bands-and-musical-artists-from-here/reference.

Schindler, Alexander, Rudolf Mayer, and Andreas Rauber (2012). "Facilitating Comprehensive Benchmarking Experiments on the Million Song Dataset." In: Proceedings of the 13th International Society for Music Information Retrieval Conference (ISMIR 2012), pp. 469–474.

Schreiber, Hendrik (2015a). "Improving Genre Annotations for the Million Song Dataset." In: Proceedings of the 16th International Society for Music Information Retrieval Conference (ISMIR 2015), pp. 241–247.

– (2015b). tagtraum genre annotations for the Million Song Dataset. Accessed on 2018-07-06. URL: http://www.tagtraum.com/msd_genre_datasets.html.

The Apache Software Foundation (2018). Apache™ Hadoop®. Accessed on 2018-07-08. URL: https://hadoop.apache.org/.

The HDF Group (2018). HDF5. Accessed on 2018-07-06. URL: https://support.hdfgroup.org/HDF5/.

# A. Appendix

## A.1. Processed output data

### A.1.1. Most common nouns per genre

Table 5: Most common stemmed English nouns per genre, without normalization. Words that could be both noun or verb are counted too. Only the highest three counts are shown.

| Genre | Word | Count | Proportion |
|---|---|---|---|
| Blues | alley | 6 | 1,058% |
| | love, place, ride, road, train | 4 | 0,705% |
| | babi, cruel, day, gin, people, salt, truth, wine, wound | 3 | 0,529% |
| Country | babi | 35 | 3,359% |
| | love | 17 | 0,845% |
| | jingl | 16 | 0,796% |
| Electronic | adventur | 11 | 0,651% |
| | disgrac, ritual | 10 | 0,592% |
| | feet, look | 9 | 0,533% |
| Folk | heart, hope | 7 | 0,552% |
| | citi | 6 | 0,474% |
| | eye, face, ride | 5 | 0,395% |
| International | call, ride | 5 | 1,285% |
| | countri, eye | 3 | 0,771% |
| | babe, daughter, day, father, fear, gentlemen, light, sheep, week | 2 | 0,514% |
| Jazz | doctor | 5 | 1,160% |
| | time | 4 | 0,928% |
| | life, love | 2 | 0,464% |
| Latin | danc | 11 | 0,596% |
| | time | 8 | 0,433% |
| | faith | 6 | 0,325% |
| Pop / Rock | love | 162 | 0,621% |
| | time | 103 | 0,395% |
| | way | 88 | 0,338% |
| Rap | light, shit | 25 | 0,387% |
| | music | 21 | 0,325% |
| | peopl | 17 | 0,263% |
| Reggae | road | 44 | 3,839% |
| | pictur | 15 | 1,309% |
| | man | 11 | 0,960% |

| Genre | Word | Count | Proportion |
|---|---|---|---|
| R'n'B | call | 41 | 1,593% |
| | stori | 21 | 0,816% |
| | babi | 20 | 0,777% |

### A.1.2. Most common words per genre

Table 6: Most common stemmed English words per genre, after normalization, explained in section 4.1.2. Only the highest three counts are shown.

| Genre | Word | Count | Proportion |
|---|---|---|---|
| Country | babi | 35 | 3,359% |
| | like | 24 | 2,303% |
| | love | 17 | 1,631% |
| Electronic | there | 30 | 3,421% |
| | go | 23 | 2,623% |
| | now | 17 | 1,938% |
| Pop / Rock | know | 176 | 1,353% |
| | never | 171 | 1,314% |
| | love | 162 | 1,245% |
| Rap | die | 51 | 1,401% |
| | like | 35 | 0,962% |
| | up | 32 | 0,879% |
| Reggae | road | 44 | 7,626% |
| | pictur | 15 | 2,600% |
| | like/man/while | 11 | 1,906% |
| R'n'B | call | 41 | 3,071% |
| | feel | 26 | 1,948% |
| | stori | 21 | 1,573% |

### A.1.3. Artists with the highest familiarity score

Table 7: Artists with the highest average familiarity, rounded to three decimal places. Only the artists with the highest 30 scores are shown.

| Rank | Artist name | Average familiarity |
|---|---|---|
| 1. | Akon / Kardinal Offishall | 1.000 |
| 2. | Lil Wayne | 0.990 |
| 3. | Britney Spears | 0.947 |
| 4. | Avril Lavigne | 0.942 |
| 5. | Fall Out Boy | 0.938 |

| Rank | Artist name | Average familiarity |
|---|---|---|
| 6. | Mariah Carey | 0.935 |
| 7. | Alicia Keys | 0.934 |
| 8. | Muse | 0.929 |
| 9. | Slipknot | 0.929 |
| 10. | The Game | 0.928 |
| 11. | Evanescence | 0.921 |
| 12. | The Killers | 0.918 |
| 13. | Big Daddy Rick | 0.918 |
| 14. | Lily Allen | 0.916 |
| 15. | Madonna | 0.916 |
| 16. | Taking Back Sunday | 0.912 |
| 17. | System of a Down | 0.910 |
| 18. | 30 Seconds to Mars | 0.909 |
| 19. | AFI | 0.909 |
| 20. | Daddy Yankee/Bounty Killer | 0.909 |
| 21. | Maroon 5 | 0.905 |
| 22. | Natasha Bedingfield | 0.901 |
| 23. | Radiohead | 0.900 |
| 24. | The Maine | 0.895 |
| 25. | New Found Glory | 0.892 |
| 26. | The Smashing Pumpkins | 0.888 |
| 27. | Killswitch Engage | 0.888 |
| 28. | Beyoncé | 0.887 |
| 29. | 50 Cent | 0.883 |
| 30. | Janet Jackson | 0.882 |

### A.1.4. Artist's familiarity heatmap

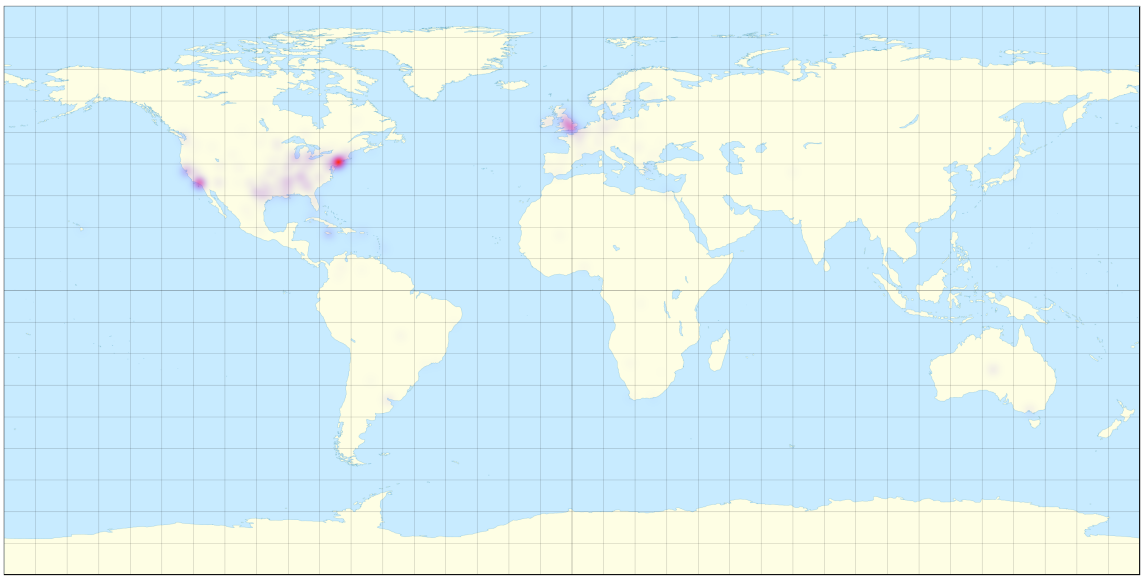


Figure 3: Artist's familiarity plotted on a map based on the artist's location (if known). Adapted from Gaba (2008).

### A.1.5. Artists with the highest song hotttnesss score

Table 8: Artists with the highest average song hotttnesss, rounded to three decimal places. Only the artists with the highest 30 scores are shown.

| Rank | Artist name | Average song hotttnesss |
|---|---|---|
| 1. | Led Zeppelin | 1.000 |
| 2. | The White Stripes | 0.972 |
| 3. | The Mars Volta | 0.929 |
| 4. | Nickelback | 0.910 |
| 5. | Thrice | 0.876 |
| 6. | Public Image Ltd | 0.874 |
| 7. | Charlotte Gainsbourg | 0.870 |
| 8. | The Maine | 0.850 |
| 9. | Temple of the Dog | 0.849 |
| 10. | Britney Spears | 0.840 |
| 11. | Lighthouse Family | 0.821 |
| 12. | Porcupine Tree | 0.821 |
| 13. | AFI | 0.814 |
| 14. | ATB | 0.811 |
| 15. | Spooky Tooth | 0.807 |
| 16. | Salt-N-Pepa | 0.806 |
| 17. | Pink Floyd | 0.793 |

| Rank | Artist name | Average song hotttnesss |
|---|---|---|
| 18. | System of a Down | 0.792 |
| 19. | The Jeff Healey Band | 0.789 |
| 20. | 30 Seconds to Mars | 0.789 |
| 21. | The Radio Dept | 0.788 |
| 22. | Xmilk | 0.783 |
| 23. | Ayo | 0.771 |
| 24. | Motograter | 0.771 |
| 25. | Cake | 0.764 |
| 26. | Lupe Fiasco | 0.764 |
| 27. | Van Halen | 0.760 |
| 28. | Chris Rea | 0.758 |
| 29. | Strata | 0.755 |
| 30. | Arsonists Get All The Girls | 0.755 |

### A.1.6. Artist's song hotttnesss heatmap

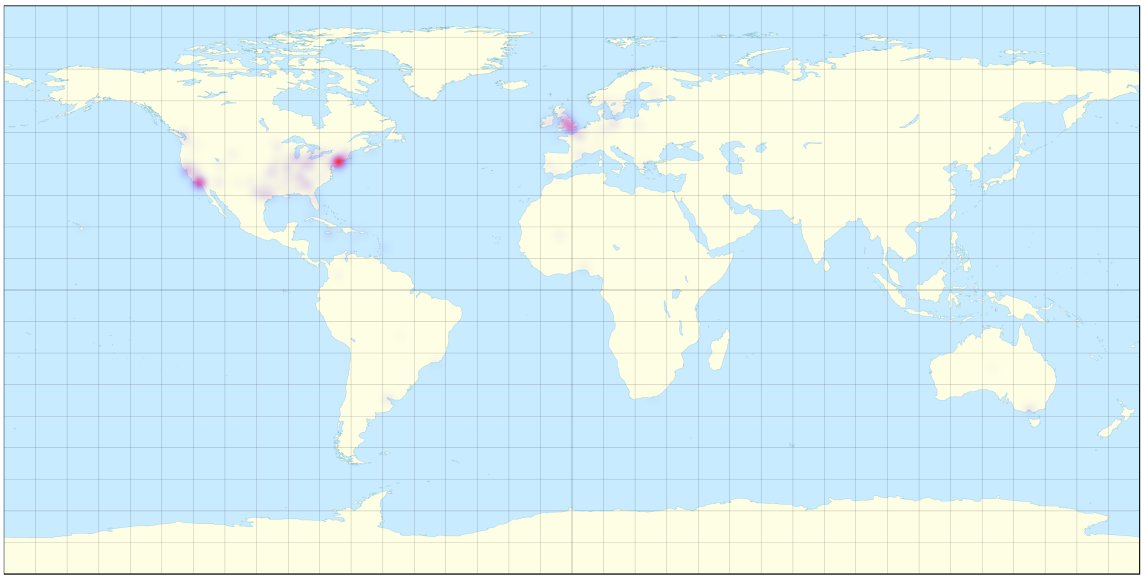


Figure 4: Artist's average song hotttnesss plotted on a map based on the song's artist's location (if known). Adapted from Gaba (2008).

## A.2. Source code and raw data

The source code and scripts for running this analysis as well as the unedited map-reduce outputs can be found publicly on the corresponding repository at `https://github.com/heinrichreimer/song-analysis`. The repository provides installation scripts for Linux and Gradle build tasks for each analysis task, including some not mentioned in this thesis.