Billboard Songs Dataset

ISLA CASSAMO, FEUP, Mozambique
MARIA BAIA, FEUP, Portugal
RICARDO NUNES, FEUP, Portugal

Music has been part of the day to day life of humanity for centuries but in order to obtain more than the standard information on them, one still has to scour the internet for small bits of information and piece it together from the various sources. Our project aims to gather information on all the songs published on their weekly top 100 to the last 48 years and compile them to a single unanimous dataset. This will be followed by the refinement and analysis of the data so to allow future searches on the dataset.

Additional Key Words and Phrases: datasets, data gathering, data cleaning, data analysis

ACM Reference Format:
Isla Cassamo, Maria Baia, and Ricardo Nunes. 2021. Billboard Songs Dataset. *ACM Trans. Graph.* 1, 1, Article 1 (November 2021), 5 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Music is the art of arranging sounds in time through the element of melody, harmony, rhythm and timber, and has been with humanity since the begging of times.[Epp21] With the popularization of the internet since 1998, music stopped being distributed through the typical physical means, such as CD and cassettes, and assumed its well-deserved place on the internet, allowing the easy globalization of songs. For having numerous stands, themes, origins and consequently being targeted by diverging audiences, institutions started collecting information on anything about music, such as lyrics, artist articles, writers notes, samples and placing it on diverse platforms across the web, to provide easy access to such desirable information. As each platform gathered its own sources through private, and sometimes unverified means, and by there is so much information available on anything about music and also about any song in particular, information that is scattered tends to become inconsistent and somewhat incomplete. Our project then aims to centralize this information by collecting data already existent on the internet ,with particular focus on the most popular songs (those recognized by billboard top 100),so as to allow those that procure information about a song to do so in just one click in a platform with data that was refined and analysed thus proposing never before seen searches.

Authors' addresses: Isla Cassamo, FEUP, Maputo, Mozambique; Maria Baia, FEUP, Porto, Portugal, up201704951@g.uporto.pt; Ricardo Nunes, FEUP, Porto, Portugal, up201706860@g.uporto.pt.

0730-0301/2021/11-ART1 $15.00
https://doi.org/10.1145/1122445.1122456

2 DATA PREPARATION

As music has been present in the life of humanity for so long, there is a lot of data available on the topic. Therefore, before finding the appropriate datasets, it was necessary to choose a feasible scope that would represent the industry well, while being relevant to the society at hand and with enough documentation to still be aligned with the project's objectives. Find musical sources did not prove to be a difficult task and to obtain a simple representative set of song names the choice came to the Billboard dataset [Tau18]. This dataset proved adequate as for a reasonable size of data it portrayed the most representative sample that aligned with the reality of the industry. The second step was to merge this primary data, that contained no more than the song name, it's leading artist, the release date and Billboard ratings calculated using internal and not reliable data, with other sources of information available on the internet. In this project, this other sources came from the Genius API [Mil21] and its Python extension lyricsGenius, which allow access to data from a reputable company that collect and creates data of the musical scope. Despite not allowing direct access to its datasets, one could use Python and crawling methods in order to scour the API and retrieve information about the primary data and arranging it in CSV files one by one.

2.1 Data storage

The process of obtaining data from the Billboard dataset put together a great quantity of primary data that were compiled into a single CSV file named songs_1970_2018.csv. Using Python, the next step was to crawl the web using the genius API and the existing songs on the songs_1970_2018.csv file, which required careful calibration in order to create the artist.csv, album.csv and lyrics.csv files, that were complete in such a manner that they would still be relevant in relation to the primary data used. The decision to compile our data in CSV files was due to the necessity to store data in a structural manner, that would be supported by Python and that could be easily navigated through.

2.2 Search Topic

The data search engine will consult a set of songs, originated between the years 1970 and 2018, their lyrics, albums they belong to, and the artists who produced them. For a more detailed search, you'll be able to filter the results based on year, artist, among other filter options. The engine will have the capacity to:

- Search for songs
- Search for artists
- Search for albums
- Search for quote of a letter

2.3 Data limitations

Although the information collected is clean and complete on it's own, it proved difficult to merge data between sources only trough the basis of the song title and or the artist name. This is due to the fact that these attributes are not always coherent amongst sources, as no string keys generally are, leading to restrictions during merge and crawl of data from the API. Due to these discrepancies, that can be as small as Billboard calling an artist "Beatles" and the API calling the same band "The Beatles" there is a diring need to normalize the data between the sources and perform a few corrections to the primary data where one was most noticeably identified. Even with normalization, due to these discrepancies being hard to detect and the lack of support by the sources to re-evaluate static data, the merge and crawl are bound to cause the appearance of null instances where the crawler could not identify the entity and it became vital to study the relevance of the data with incomplete data.

2.4 Data cleaning and refinement

For this segment of the project to be done successfully, it was imperative to consider that the search on the dataset was limited to the following sections: simple full-text search, search refinement and organization of the results obtained. With this in mind, after obtaining the billboard data in CSV format and associating its artists, lyrics and albums information from the crawl, one needed to deal with the fact that some data would be incomplete, inconsistent and duplicated due to the limitation highlighted above. To perform this task, Python functions present in the Pandas library were used. To name a few, the elimination of duplicate data from the dataset was performed using the Panda function drop_duplicates(), which when used with the flag "keep-first", removed all occurrences of equal data, keeping the only the first occurrence. The rename function was used to normalize the column names so to eliminate certain overlaps created during the merge of the tables. Finally, to remove the null values that could not be fixed during normalization, we use the dropna() function.

3 CONCEPTUAL MODEL

The base unity of the system used in the domain of the application is the song. Song has the attributes title, artistId, peakPos, lastPos, weeks, rank and lyrics. This model also makes use of the notion of artist. An artist created a song and its attributes are as follow: name and bio. Another concept that this conceptual model posses is the album. An album represents an agglomeration of songs and is created by a single artist. An album possesses the attributes name and release-date. This way, and artist can have many albums and songs to associated to itself, a song must have one leading artist and may have an album associated to it. Lastly, an album may have numerous songs and artist associated to itself. The conceptual model follows as shows in Figure 1.

4 DATA ANALYSIS

In this section, we will take a look at some key characteristics of our dataset. First, we search for the top 10 musics, artists, and albums of the Billboard based on the number of weeks they were on it (Table 1, Table 2 and Table 3, respectively) The average length of all lyrics is around *5875* characters. The word **love** is present in *8450* music lyrics, while the word **hate** is only present in *1763*!

We wanted to see the evolution of number of songs in-and-out of the Billboard through the years, so we constructed the line plot in Figure 2. As we can we see, there aren't many songs of 1970, probably because the data collection started mid year. Furthermore, there is a spike in the year 2000, where the number of songs nearly doubled.

5 CONCLUSION

In this milestone the goal required one to procure appropriate datasets, determining the tools needed to explore them and make use of those tools to the desired extent. After analysis on the primal database, it proved imperial to understand how to cross data from other sources in order to enrich the initial dataset chosen. As the merge were to be done by song title and artist name it proved itself a difficult task due to the vast presence of inconsistencies often present in string keys. One can although conclude that the final result table included a large dataset of music with complete and coherent data throughout, which can easily be queried to obtain sets of information needed to produce a useful and effective engine for the future platform.

REFERENCES

[Tau18] Michael Tauberg. *Billboard Hot-100 songs from 1970-2017*. 2018. URL: https://data.world/typhon/billboard-hot-100-songs-from-1970-2017.

[Epp21] Gordon Epperson. *music*. 2021. URL: https://www.britannica.com/art/music.

[Mil21] John W. Miller. *LyricsGenius: a Python client for the Genius.com API*. 2021. URL: https://pypi.org/project/lyricsgenius/.

ANNEXES

Table 1. Top 10 Billboard musics

Song Name	Artist Name	Nº of weeks
Radioactive	Imagine Dragons	87
Sail	AWOLNATION	79
I'm Yours	Jason Mraz	76
How Do I Live	LeAnn Rimes	69
Counting Stars	OneRepublic	68
Party Rock Anthem & LMFAO Featuring Lauren Bennett	GoonRock	68
Rolling In The Deep	Adele	65
Foolish Games/You Were Meant For Me	Jewel	65
Before He Cheats	Carrie Underwood	64
Ho Hey	The Lumineers	62

Table 2. Top 10 artists on the BillBoard

Artist Name	Nº of songs
Glee Cast	183
Taylor Swift	67
Drake	63
Elton John	59
Madonna	57
Tim McGraw	50
Rod Stewart	48
Kenny Chesney	46
Chicago	44
Billy Joel	43

Table 3. Top 10 albums on the Billboard

Album Name	Nº of songs
Ulysses	46
Finnegans Wake	28
Glee: The Music, The Complete Season Two	28
Glee: The Music - The Complete Season Three	23
More Life	21
beerbongs & bentleys	17
The Big Book of Song Lists	17
Starboy	16
Views	16
Purpose (Deluxe)	15

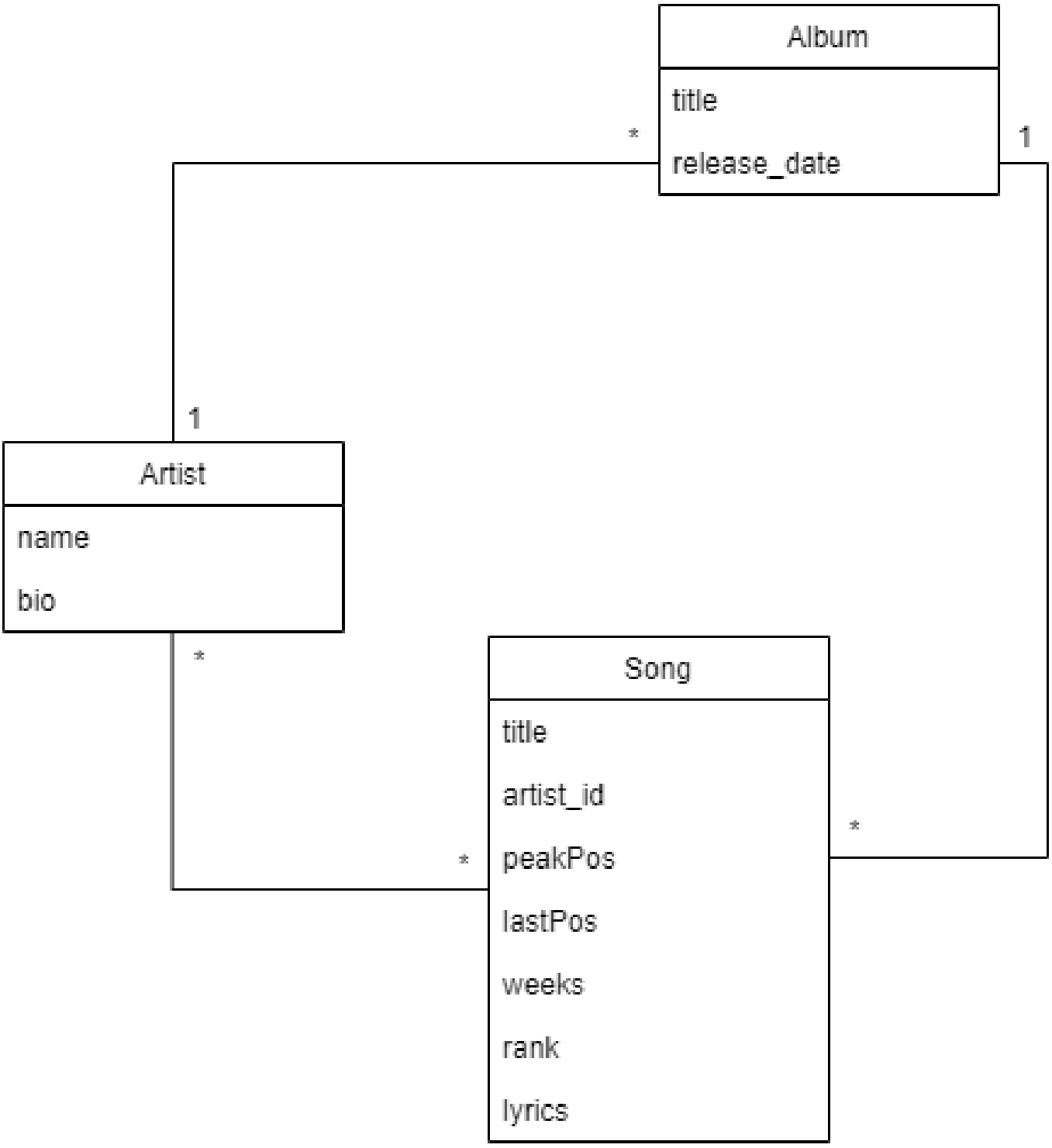

Fig. 1. Conceptual model

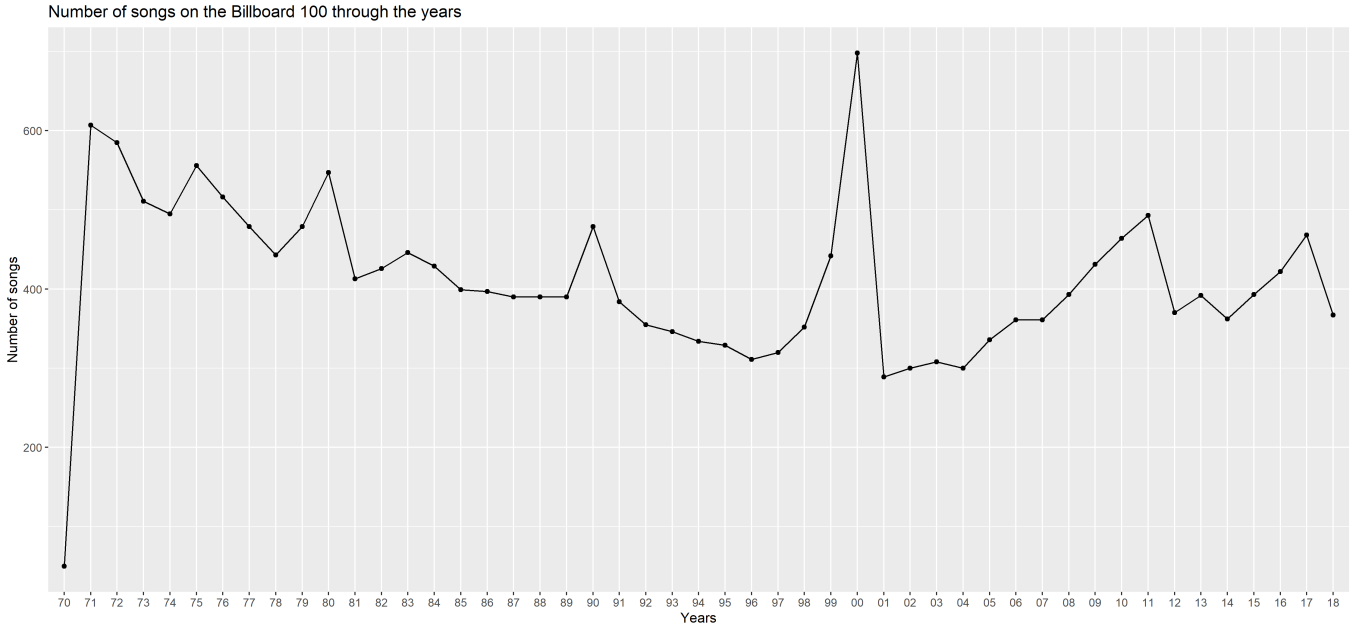

Fig. 2. Evolution of the number of musics passing through Billboard over the years