PRI-66 Project Report
Spotify lyrics Data

André Santos, Edgar Torre, Miguel Amorim
Faculdade de Engenharia, Porto

ABSTRACT

Information has never been so available as nowadays. That's why the processing of acquiring information and preparing that information is as important as never before. This report intends to document the creation of a good database and an information processing in the context of 2022/2023 Project for Information Processing and Retrieval course.

Keywords: Information Processing and Retrieval, Music, Lyrics, datasets, M.EIC, statistics, music's data

1 INTRODUCTION

As part of course PRI, a project, to develop during the whole semester, was requested. In the end, this project will be an information search system, which means work on data collection and preparation, information querying and retrieval, and retrieval evaluation.

As per the chosen topic, at first we did not know what would be the best option like food, movies or music. Our decision was made based on three reasons, we wanted a dataset with a large amount of data, with at least thousands rows, we wanted different information in the dataset, with at least 10 rows. We also wanted different types of data. With all of that, we decided on a music database.

2 DATASET

2.1 Dataset choice

We start looking for a dataset with a lot of different data, like release year, artist, album, popularity, and most importantly, the lyrics. We found a dataset with more than 4 million rows with the lyrics of each one and some more data about each music. Although it was difficult to work with so a large amount of data we decided to work with this one found on Kaggle (link below).

However, this dataset had some problems, including information that we do not need (which we deleted), too many rows to work with (which we deleted and kept just a part of the rows), and lacking the different types of information. With that problem in mind, the solution we found was finding some API that could give us the data we need. We found that the Web Spotify API (link below) could give us what we were looking for, like popularity, albums, and duration of music.

In conclusion, we selected a dataset of music data collected by someone else and added to it a lot of information with the Spotify API.

Kaggle link: https://www.kaggle.com/datasets/nikhilnayak123/5-million-song-lyrics-dataset .

API link: https://developer.spotify.com/documentation/web-api/ .

2.2 Dataset Content

The chosen dataset contains the following:

- 4 million musics with attributes as artist name, release year, lyrics, title (which we used 80000 musics, picked randomly).

From the API, for each music we got the following data:

- track number, album's name, album's number of tracks, track's qualification, duration, popularity, album's release year.

2.3 Data Quality and Source

Kaggle is a well-known platform for data searches, the post author states that all the data comes from genius.com, a well-known music platform with 5.5 million followers on social media, with this in mind, we can infer that this data is reliable. About Spotify, since it's an app that is used all around the world and already receives some awards, we can also state its reliability.

About the quality of the data, we can say that met our criteria, the date we needed was there and with no unexpected values.

3 PIPELINE

Our Data Preparation Pipeline is done, mostly, on Python scripts, where pandas and matplotlib libraries were fundamental to cleaning and manipulating the data. With that, we manage to create a structured and clean SQL database system.

3.1 Pipeline Scheme

After all the analysis of the problem, the final version of the pipeline process was the following:

As we can see, the initial data come from the Kaggle dataset ”ds2.csv” and we refined this data, eliminating the duplicated ones and eliminating the columns that we did not need (year, features, id, views). Furthermore, we picked a smaller population of data from the bigger one we had on our hands and added more data that came from the API (duration, track's number, popularity, release year, track qualification, album's number of tracks, and album's name).

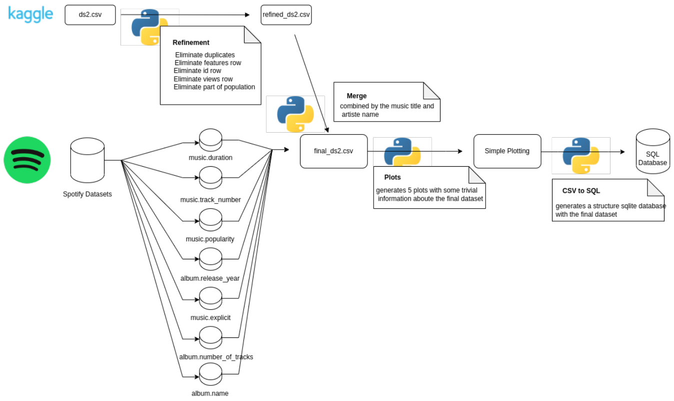

Figure 1: Pipeline Scheme

The join was made based on the title of the music and the name of the artist. After that, the process consists of plotting some graphs and turning the final version of the dataset on a structured SQL database system.

3.2 Domain Conceptual Model

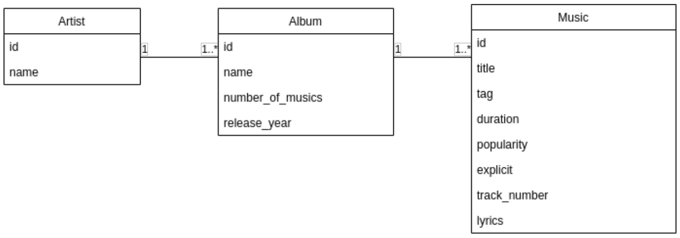

Figure 2: Domain Conceptual Model

The Database consists of three tables. The table Artist stores the name of every artist. The table Album stores the general information about the albums. And a Music table, where all the information about the music is stored, that's the main class of our database. With this in mind, and using the final version of the dataset and another Python script. we created our SQL database.

3.3 Data characterization

We made a python script to see general information and plot them, to better understand the final version of this dataset.

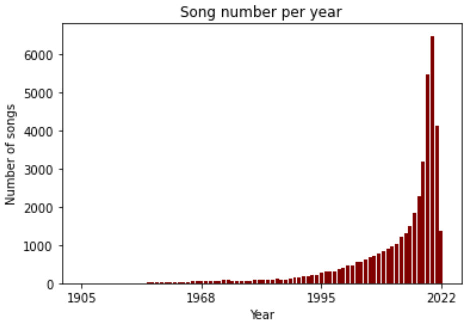

Figure 3: Musics per Year

To start, although a big part of the music of our dataset is recent, we have music from each year since before 1968, which means that the information about most recent years is much richer but we can see the evolution from years ago.

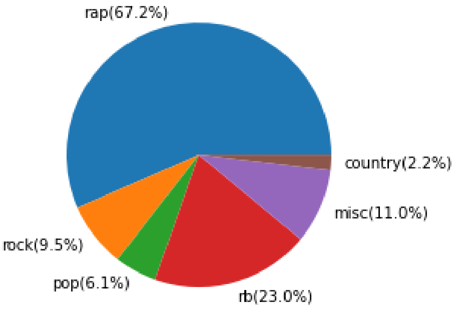

Figure 4: Musics Grouped by Genre

The second plot shows us how many genres of music we have on our dataset. We can see the variety of genres, which makes our database richer and more reliable to search.

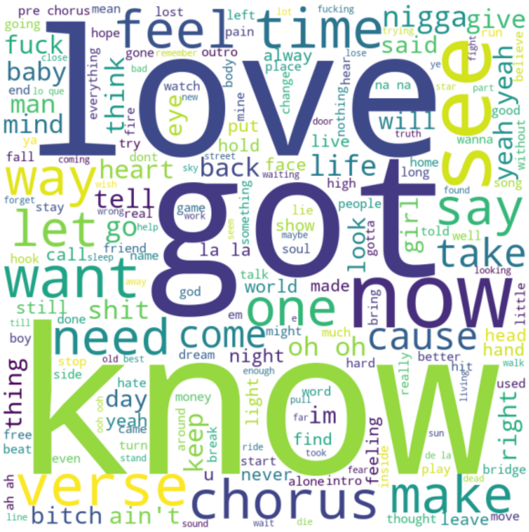

Figure 5: Lyrics Word Cloud

We also found out that would be interesting to search for the most popular words in all lyrics we had. For that, we made a word cloud (a type of graph to measure the frequency of words). We can easily see that there are words that are used much more frequently than others, like "love" or "know". This also confirmed some initial thoughts, like the high frequency of bad words.

Furthermore, we thought we could see the range of duration of the music. We concluded that most songs have mostly 2 minutes in duration. However, we have music that is in the 5 minutes range.

To finish, we also studied the average popularity of the music every year, and we can see that didn't change much over the past few years.

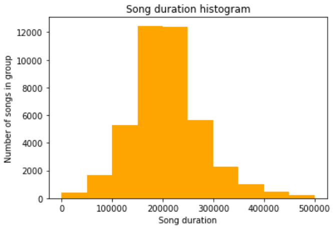

Figure 6: Music Duration Histogram

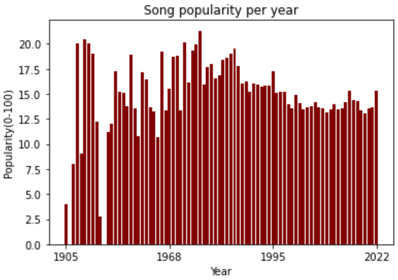

Figure 7: Music Duration Histogram

4 Conclusion

With the work we have now, we can already search for a lot of information. In the future, we want to explore this, even more, the database with the objective of creating the best retrieval tool for this specific database. We are really interested to explore, so we can also answer the following questions:

- What are the most used words by genre?
- Did the music duration change with time?
- Does the percentage of explicit rap music affect its popularity of it?
- Does the tracking number of music affect its popularity?