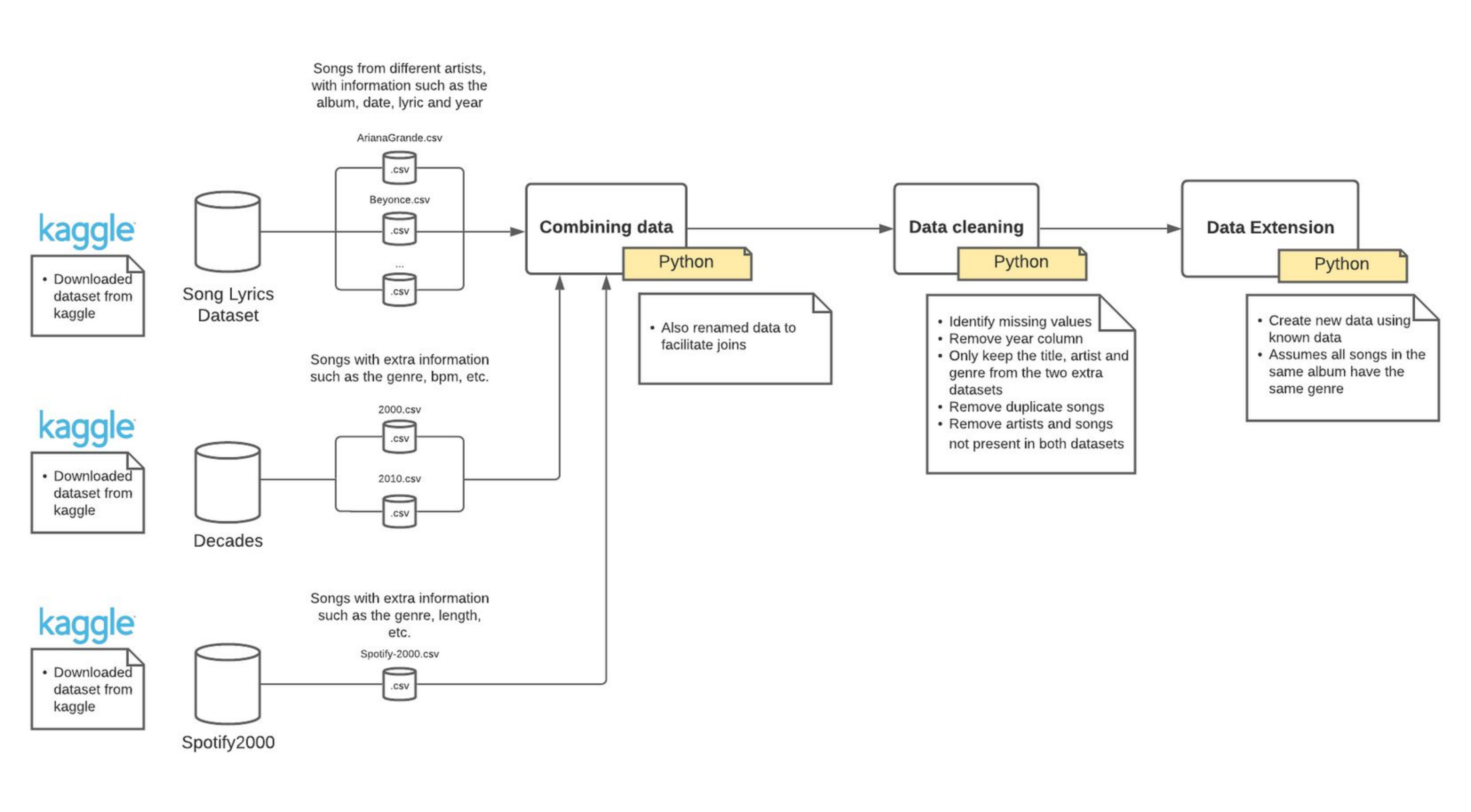
SONG INFORMATION AND LYRICS

PRI - 1ST MILESTONE

Afonso Caiado, up201806789
Diogo Martins, up201806280
José Maçães, up201806622

PIPELINE



kaggle

- Downloaded dataset from kaggle

Song Lyrics Dataset

Songs from different artists, with information such as the album, date, lyric and year

ArianaGrande.csv

.csv

Beyonce.csv

.csv

...

.csv

kaggle

- Downloaded dataset from kaggle

Decades

Songs with extra information such as the genre, bpm, etc.

2000.csv

.csv

2010.csv

.csv

kaggle

- Downloaded dataset from kaggle

Spotify2000

Songs with extra information such as the genre, length, etc.

Spotify-2000.csv

.csv

Combining data

Python

- Also renamed data to facilitate joins

Data cleaning

Python

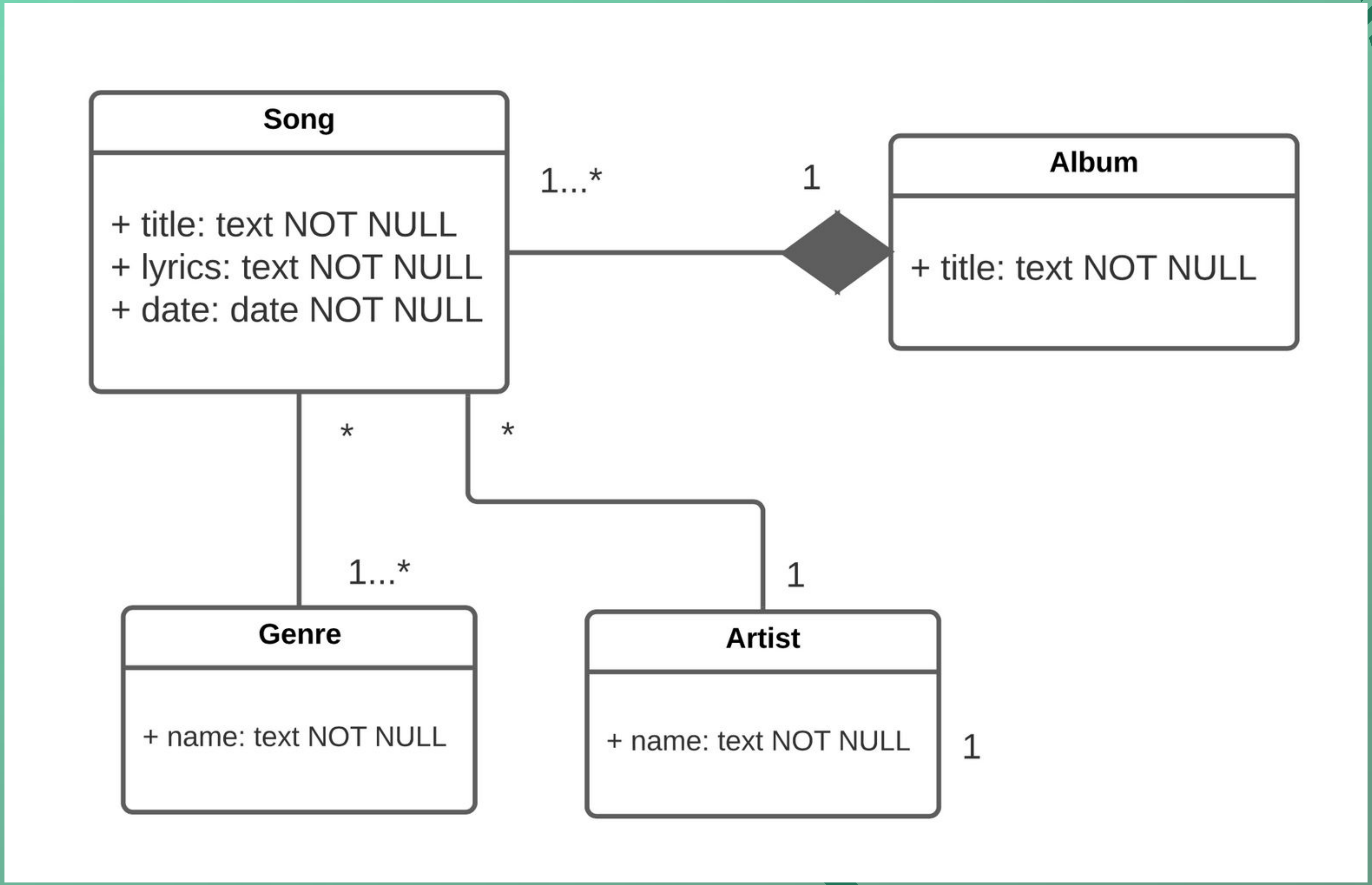
- Identify missing values
- Remove year column
- Only keep the title, artist and genre from the two extra datasets
- Remove duplicate songs
- Remove artists and songs not present in both datasets

Data Extension

Python

- Create new data using known data
- Assumes all songs in the same album have the same genre

UML



Song

+ title: text NOT NULL
+ lyrics: text NOT NULL
+ date: date NOT NULL

Album

+ title: text NOT NULL

Genre

+ name: text NOT NULL

Artist

+ name: text NOT NULL

DATA CLEANING AND COMBINING

- Renamed some column headers to facilitate joins
- Joined different lyrics files onto one single file
- Joined all genre files onto one single file
- Joined lyrics file and genres file on Artist and Title, obtaining one file with the desired format
- Removed unwanted columns such as sound attributes and year
- Removed duplicate values

DATA EXTENSION

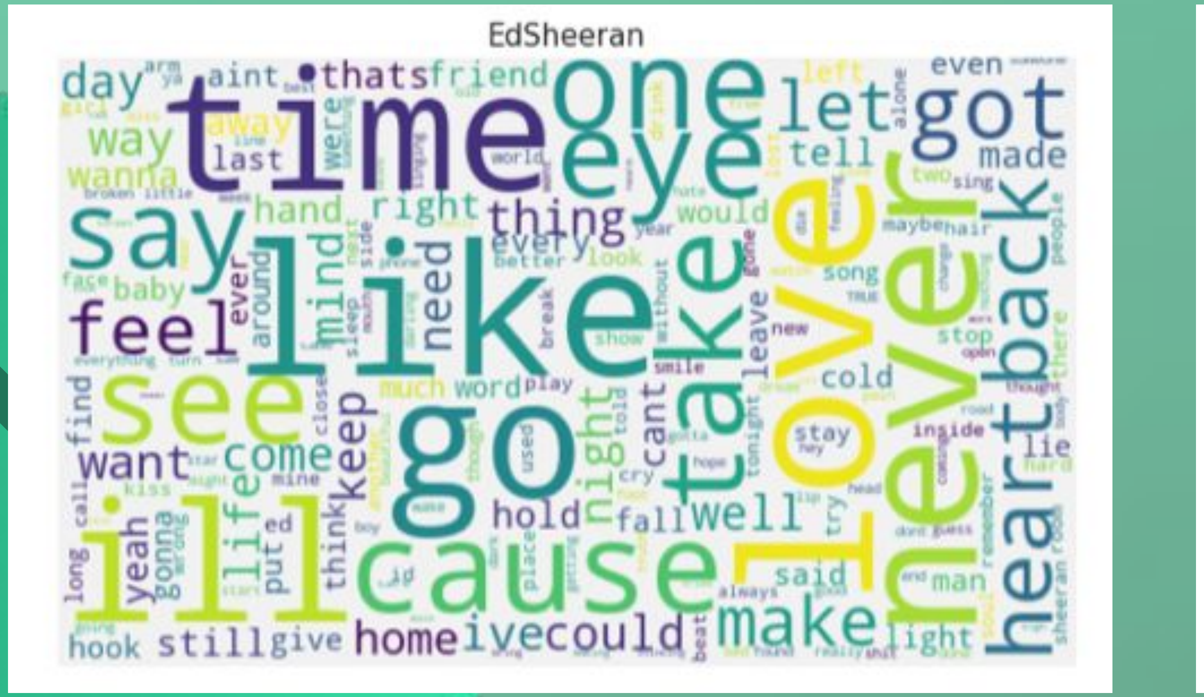
Create new data using the known data:

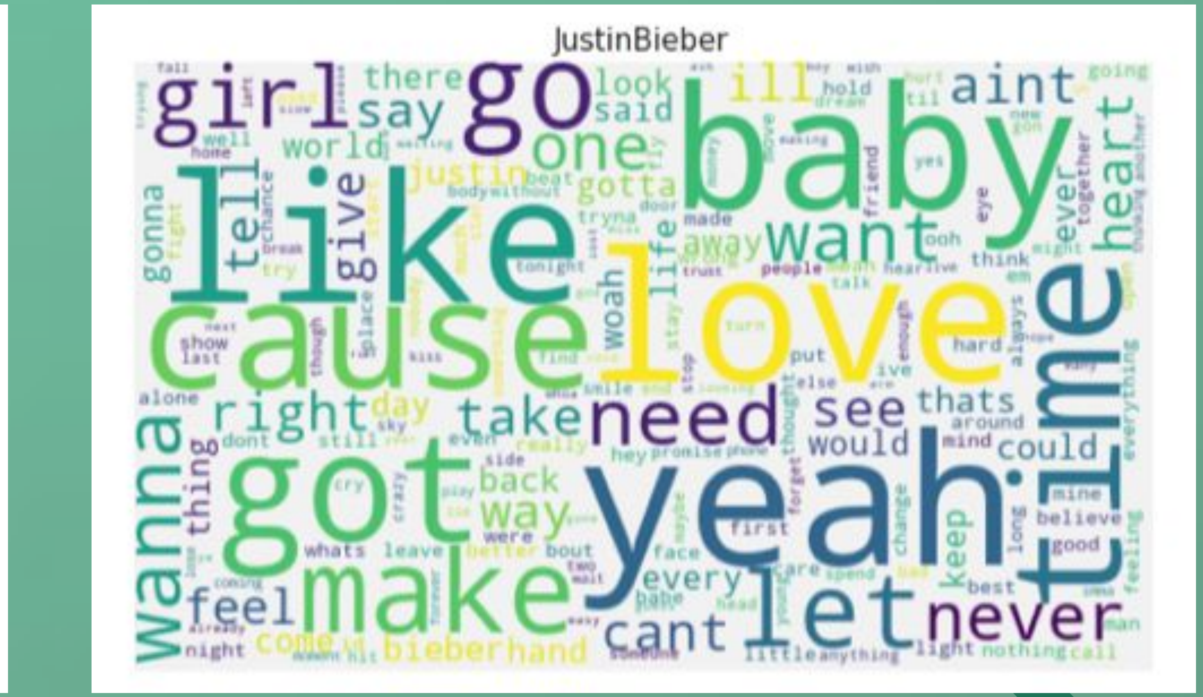
- From the known genres of songs, we searched for songs of the same album and assumed the same genre for all the songs in the álbum.

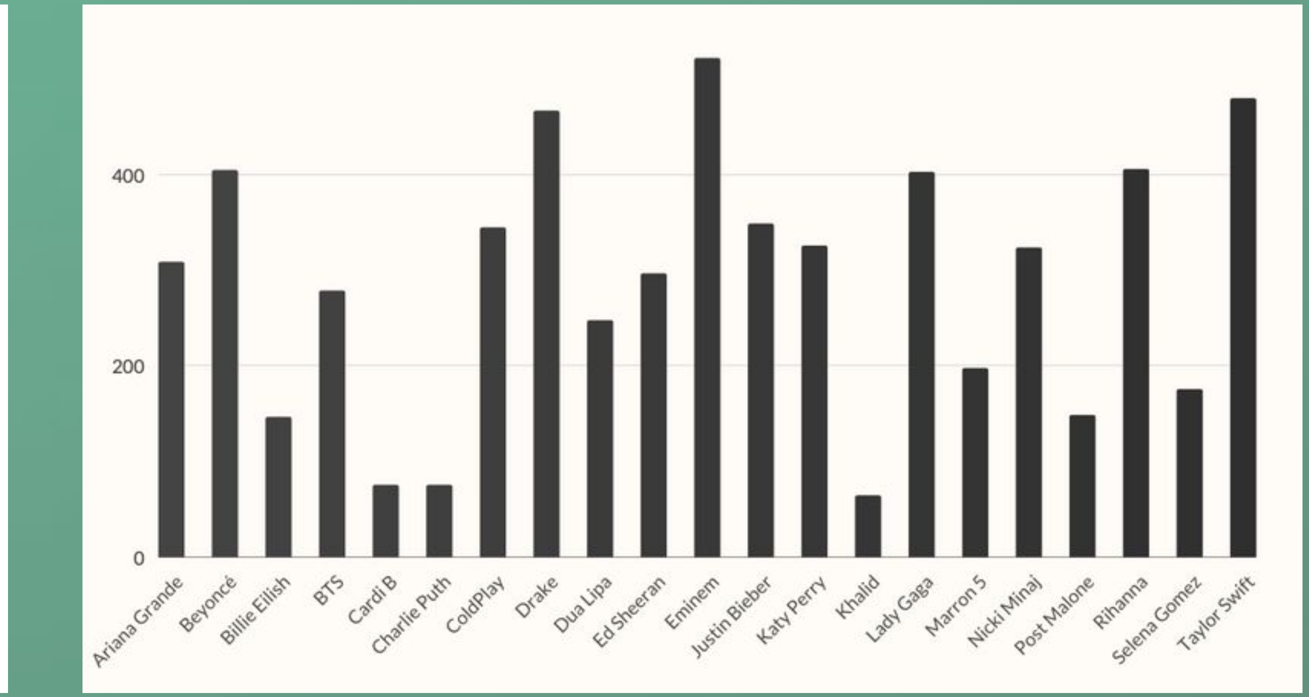
DATA CHARACTERIZATION

In the next figures, we can see the most common words in lyrics of songs by Ed Sheeran and Justin Bieber respectively. We hope that, in the future, this same information has value in a way that we can explore the data.

We also have a graph giving us the information on the number of songs in each artist’s dataset.

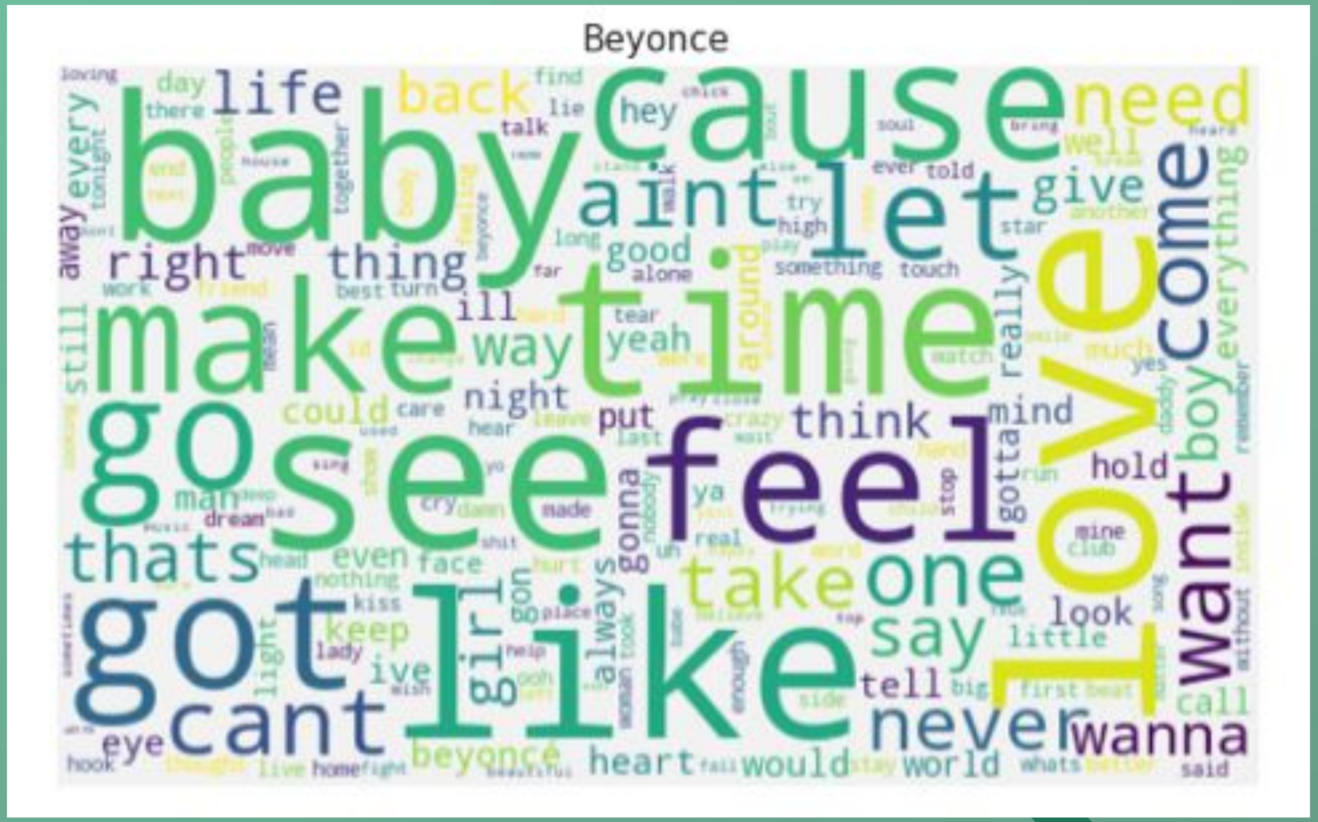






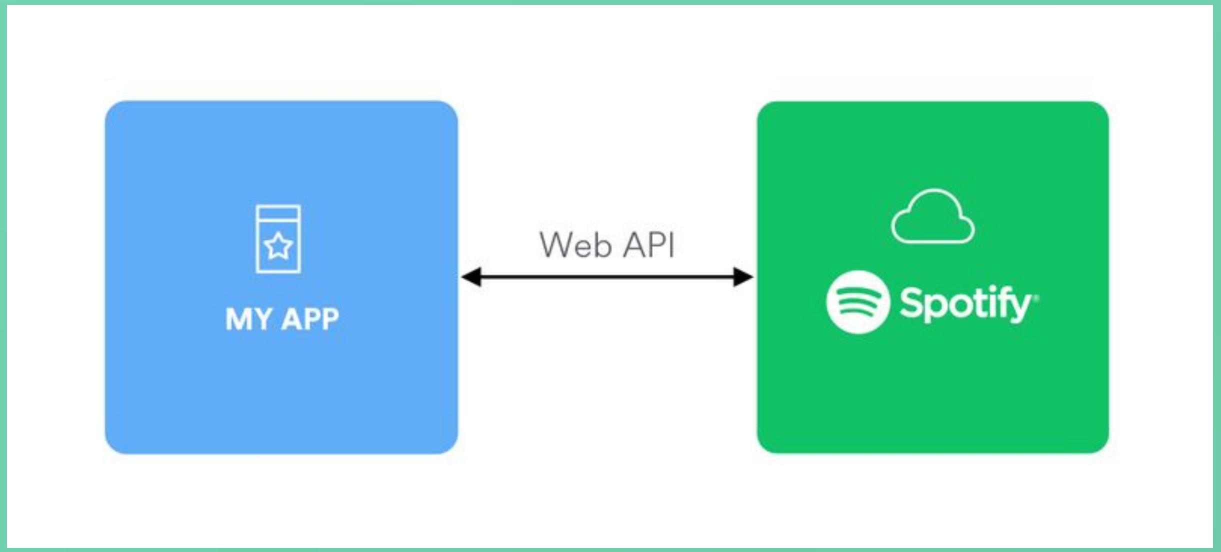
POSSIBLE QUERIES

- Search songs by artist, date, album, name
- Search songs by genre
- Search songs by words and quotes
- Search songs by the feelings and emotions expressed in it



IMPROVEMENTS

- Using the Spotify API to retrieve genre for all songs in our lyrics dataset
- Improving our Makefile



CONCLUSION

In this first phase of the project, we found and chose the data that felt relevant to our purpose. We combined it, refined it and came with an organized dataset that contains the information that we desire and that we intend to work with. The Data Model and the Pipeline were also crucial points to this project that were developed and that helped us during the elaboration of the work until now since they allowed a better organization regarding tasks and work that needed to be done. As a future work, information retrieval and the search of songs are intended to be developed by us.