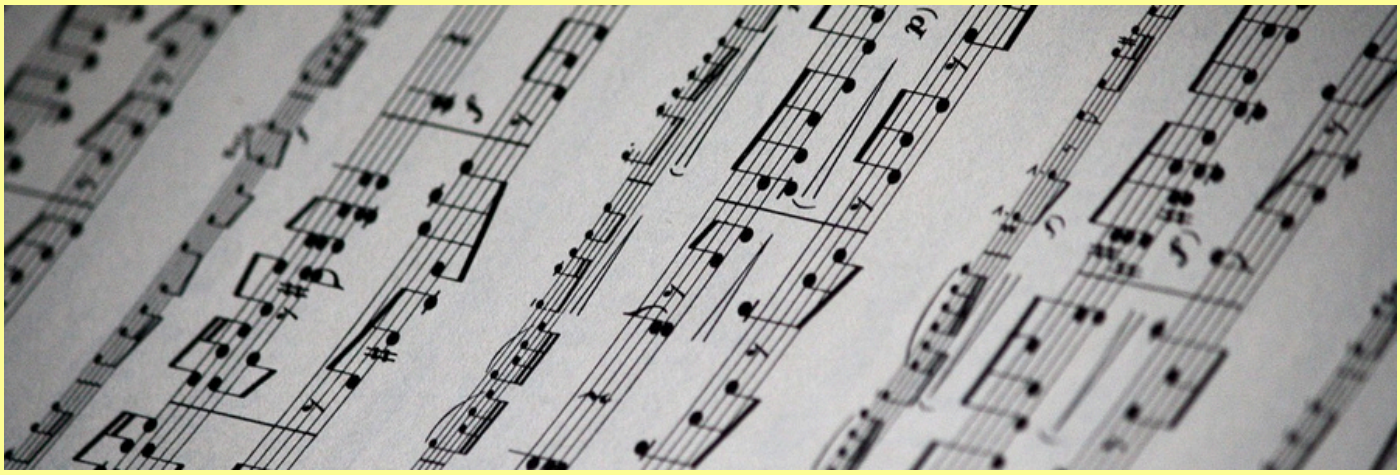
DATA PROJECT

FRENCH SONGS



EDA AND PREDICTIVE MODEL

ALEX
DAUCOURT

dauczer.github.io/portfolio-app/

INTRODUCTION / EDA

This project is based on two datasets, which I found on kaggle :
https://www.kaggle.com/datasets/quentinlelan/french-rap-lyrics-several-dataset-union
https://www.kaggle.com/datasets/carlosgdcj/genius-song-lyrics-with-language-information

The goal of this project is to create a model that can predict wether a song is a rap or a pop one. First, I'll do some basic descriptive analysis of the first dataset (which is focused on rap musics), and then I'll create my model.

The first dataset is composed of ~50 000 french rap songs. It only contains artists, title, year and lyrics of each song.

After handling the missing values, I create a tokenizer function that can isolate words (to create lists of words from the lyrics column.)

With that function, we can now produce some graphs with, for example, number of words used by an artist, differents words in total etc.

We'll start with two graphs that represent the number of words used by each artists in average. For this data, we'll represent the bottom 20 artists and the top 20.

We can observe that there is a lot of differences among these artists, from only a few words to more than 1200. However, rappers that are in the extreme parts only have a few songs.

50K

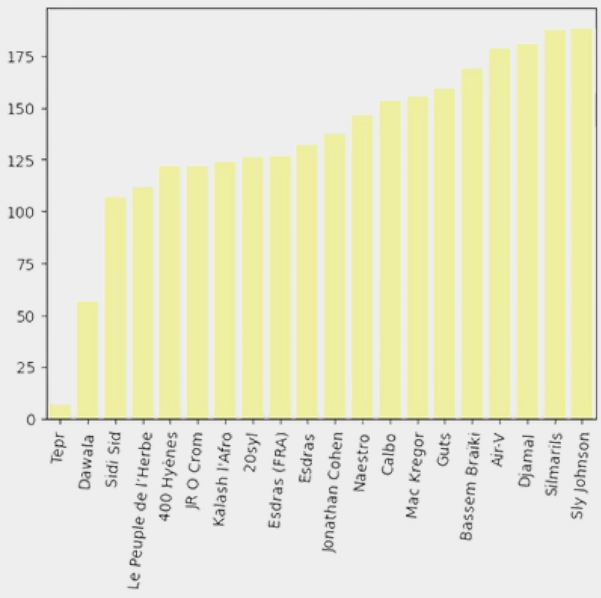
French songs in the dataset

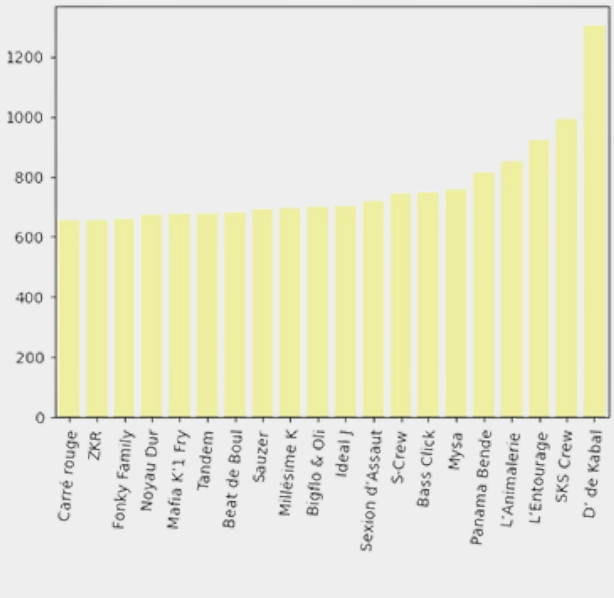
724

Artists in the dataset

395

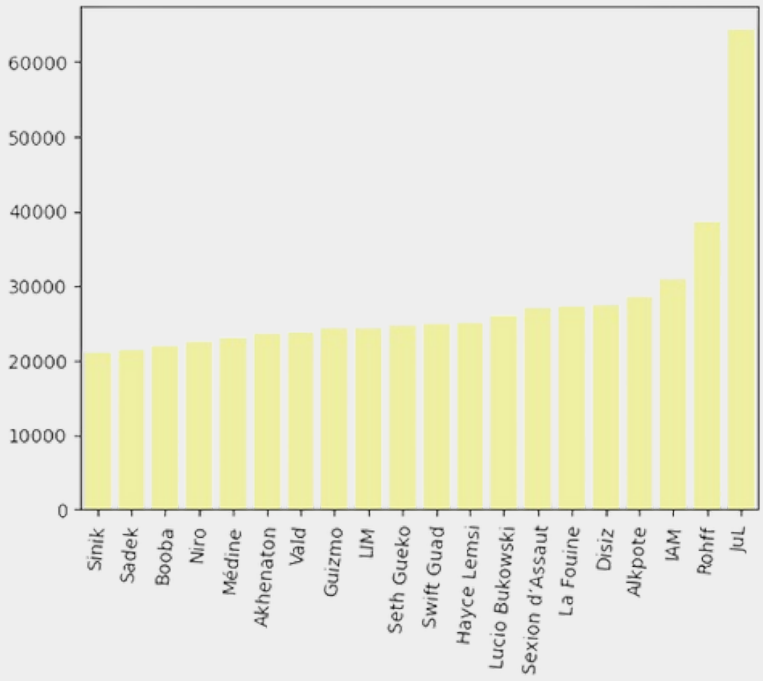
Average differents words per song





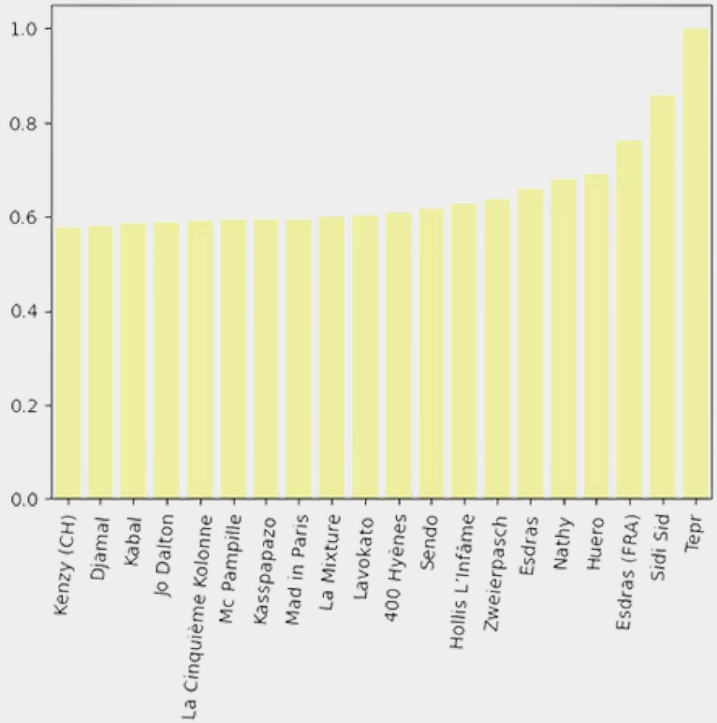
INTRODUCTION / EDA

We can also see the numbers of total words per artist on this graph :



However, this data is not really representative, because it depends on the number of songs of the artist. For example, Jul has now more than 20 albums, and isn't really famous for his lyricist skills.

That is why it would be more interesting to see the ratio of differents words on the total words, to see the diversity of the rapper vocabulary. This is what this graph is about :



We observe that Tepr is here again, but he only had 20 words in average on the first graph ! It is explained because Tepr is a producer, who indeed almost doesn't talk at all in his musics. We also notice Esdras is there twice, which shows data cleaning could have been better.

Let's now analyse a specific vocabulary of a rapper I affectionate a lot : Booba.

FOCUS ON 1 RAPPER

Now, I'd like to focus on a specific rapper. To do so, I'll count the most common words used by an artist. Then, I will have to delete the stopwords (basics words in a language), and also delete the most 20 words and the short words (less than 4 letters).

Then, we can see the most common words used for a specific rapper. For instance, here are the 20 most common words for **Booba**, and their translation in english :

- 'rien' : nothing
- 'veux' : want
- 'trop' : too much
- 'jamais' : never
- 'toujours' : always
- 'négro' : n-word
- 'bien' : good
- 'gros' : big
- 'ouais' : yeah
- 'mère' : mother
- 'deux' : two
- 'fuck' : explicit
- 'millions' : millions
- 'monde' : world
- 'coeur' : heart
- 'noir' : black
- 'temps' : times
- 'sais' : know
- 'cash' : money
- 'négros' : n-words

After that, we can create for each artist a **word cloud** that synthetises the most used words. We can clearly notice difference, and it is even more impressive when we are used to listen to these artists. I will show three word cloud of artists I like.

The **Booba** word cloud, where we can recognize words from our previous list.

The word cloud at the right is the one from **Kekra**.

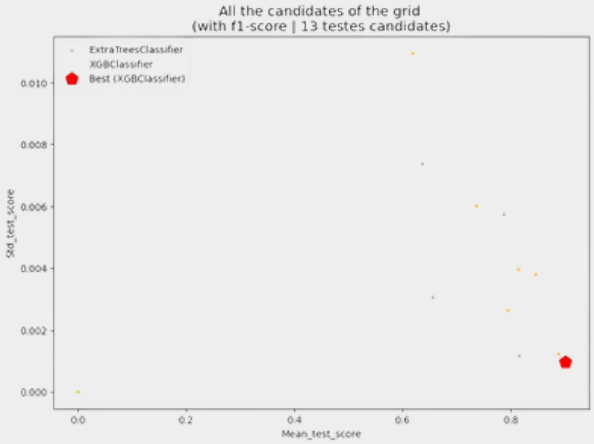
And finally, the **PNL** word cloud.

PREDICTIVE MODEL

Finally, I imported the second dataset. This one is really big, with 5 millions rows, and I only needed french pop songs. After that filter, I only had ~60 000 songs, which is relatively close to our 50 000 rap musics.
To create features for my model, I used the TfidVectorizer with only unigrams (for performance issues).

After creating two samples, a training one and a testing one, I trained my dataset on few models. Sadly, because of performance issues, I had to use only two differents models : the **XGBoost** and the **ExtraTree**. Moreover, I had to chose only few choices on my cross validation grid and a 3-fold cross validation, instead of 5 or 10.

Then, I plotted variance and f1 score of all models, and I could chose the best one out of them. Here are the results :



Indeed, the best model was a XGB one with a **learning_rate** of **0.1** and **n_estimators** of **100**.
When I tested the model on my test sample, I had this confusion matrix :

Actual / Predicted	Negative	Positive
Negative	17 221	744
Positive	2 029	13 219

Finally, I calculated some important metrics and I found, for the testing sample, a f1 score of **0.91**, an accuracy score of **0.92**, a brier score loss of **0.07** and a roc auc score of **0.97**, which are really satisfying metrics.

We can conclude that our model is very efficient, and that there is a lot of differences in the french pop and french rap vocabulary.

Thanks for reading !

French songs analysis project

by Alex Daucourt.