|  |  |
| --- | --- |
|  | February, 2024  **IA101**  **Algorithmic Information**  **& Artificial Intelligence**  Micro-study  [teaching.simplicitytheory.science](http://teaching.simplicitytheory.science/) |

Name: Aymane Hamdaoui & Charles Kayssieh

# Lyricognizer

## Abstract

The main purpose of this micro-study is to develop a method that associates some lyrics to their author using compression and distances (NCD) between the analyzed lyrics and a database of song texts from predetermined artists which have been processed to clean data.

If more details on the project of a better overview of the results are needed, you can check the project’s Github page: <https://github.com/sharle4/lyricognizer/>

## Problem

Is it possible, for a little code on a computer and given a database containing different songs from different artists, to find the one who is most likely the writer of some lyrics using only NCD?

## Method

We started by building the database. To get a relevant database, we chose 10 various artists. 7 French rappers form the 'base' of our artists. We've added an English rapper, whose lyrics are in English only, and whose prediction results we expect to be very good (vocabulary very different from the others). Two different French artists to see if the difference in rap/pop genre had an impact on the results. We used the Genius API (a website of lyrics) and grabbed the lyrics of the 100 most popular songs for each artist.

After that we processed all the lyrics with three separate methods whose effectiveness we want to test. the first is simply to 'clean up' the text: we juxtapose all the words in lower case, without punctuation, simply separated by a space. The second method is to carry out the first and then remove the stop words (words that are very common and not very specific to the identification of each artist). The third method is to perform the second and lemmatize the remaining words (replace each word with the representative of its lexical field). For the last two methods, since stop words and lexical fields depend on the language of the text, we had to implement a language detection option for the music of Charles Aznavour, Soolking and Mylène Farmer, who write mainly in French, but also in English, Italian, Spanish, Arabic, German,...

Finally, we shuffled our database and randomly allocated 80% of the music to our training data and 20% to the test data. For each text processing method (simple processing, stop word deletion, lemmatization), each of the 200 test songs is compared to all the 800 songs in the training data, by calculating the NCD using three different compression methods (zlib, nzma, bz2). The purpose of using three different compression methods is to see whether this choice has an impact on the accuracy of our results.

## Results

The results are stored in a very large python dictionary (600,000 characters). The form of the results is as follows: for each piece of music in the test data, a dictionary is created whose keys are the artists and whose values are a list associating the artist with its average distance and its rank in the proximity ranking with the music. We then average the results for the 200 tracks, grouping them by artist, so that each artist has a dictionary of the same form as the one explained above.

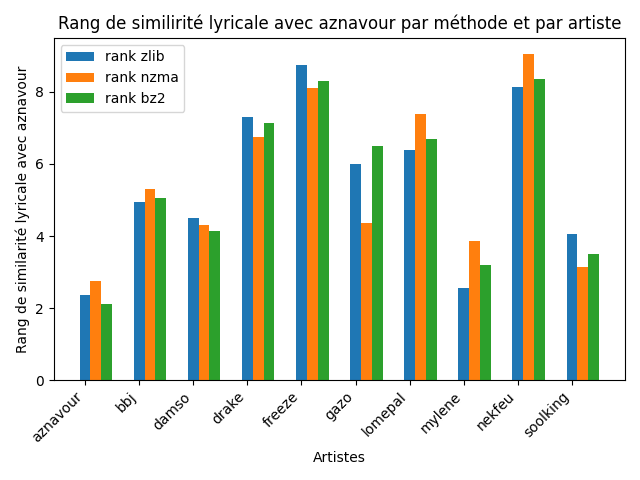
## The graphs here correspond to the average results per artist, using the stop-word suppression method for text processing. Each artist A is given two graphs. One associates the 10 artists on the x-axis with an average closeness between 0 and 1 to the music of artist A on the y-axis. (for a test song: closeness = 1-distance; the higher the closer)

## The other associates the 10 artists on the x-axis with an average rank representing the closest proximity to the music of artist A on the y-axis. The explanation is a little complicated, see the example below for a better understanding.

Moreover, in each graph three are condensed: the three colours correspond to the three compression methods, see graph legend.

**Charles Aznavour:**

**Une image contenant texte, capture d’écran, diagramme, Tracé

Description générée automatiquement**

Explanation: the first graph shows the average proximity of each artist to Charles Aznavour's lyrics. The values don't really matter on this graph, it's the comparison between the different artists that counts. Here, for each compression method (zlib=blue, nzma=orange, bz2=green), Charles Aznavour has the greatest average proximity to the 20 Aznavour texts analysed, closely followed by Mylène Farmer. Freeze Corleone has the lowest average proximity, which means that on average out of the 20 Charles Aznavour songs analysed, it's Freeze Corleone's 80 training texts that give the greatest distance (NCD). The second graph shows the average proximity rank to Charles Aznavour texts. This graph is directly related to the first but is normalized for the three compression methods. The results are the same as in the first graph, but more explicit.

**B.B. Jacques:**

**Une image contenant texte, ligne, capture d’écran, Parallèle

Description générée automatiquementUne image contenant capture d’écran, diagramme, ligne, Tracé

Description générée automatiquement**

**Damso:**

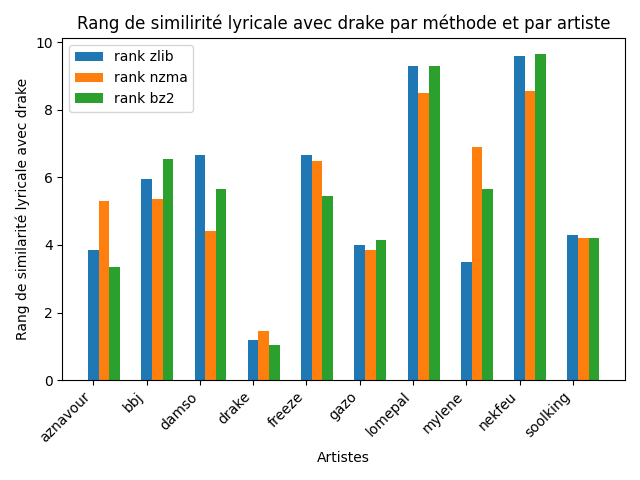
**Une image contenant texte, ligne, capture d’écran, Tracé

Description générée automatiquementUne image contenant capture d’écran, diagramme, ligne, Caractère coloré

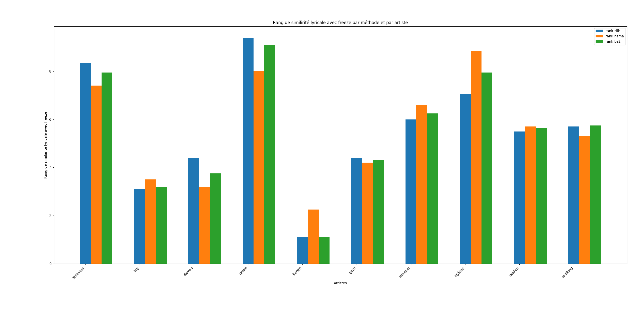
Description générée automatiquement**

**Drake:**

**Une image contenant texte, capture d’écran, diagramme, Tracé

Description générée automatiquement**

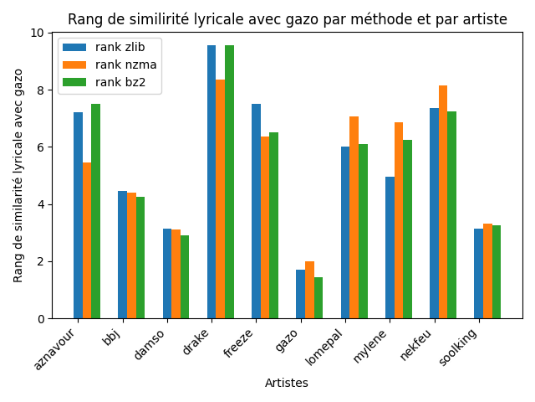
**Freeze Corleone:**

**Une image contenant texte, ligne, Parallèle, Tracé

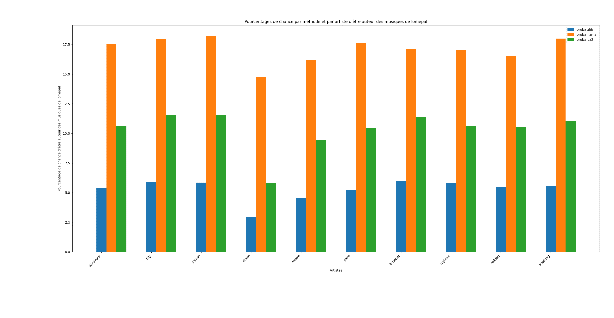
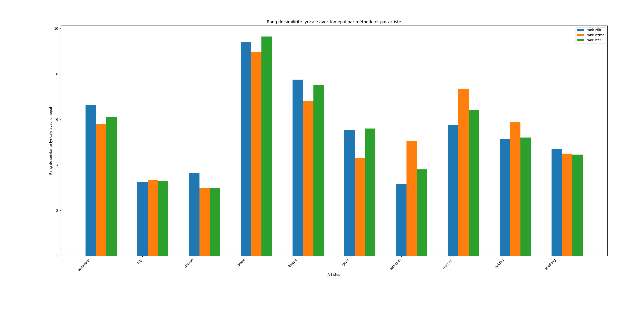
Description générée automatiquement**

**Gazo:**

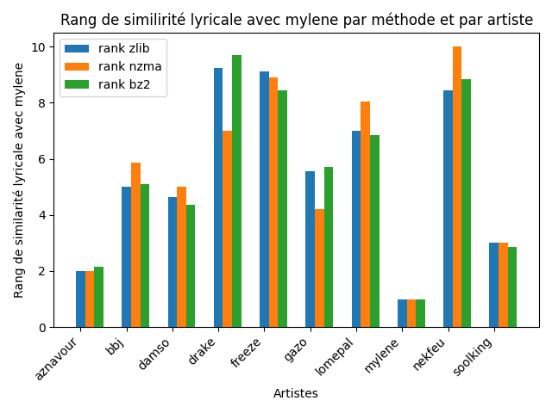
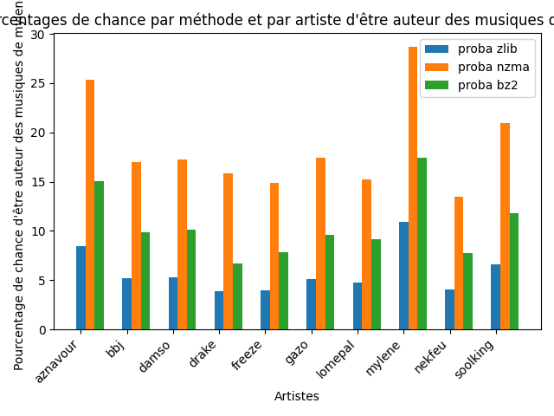
**Une image contenant texte, capture d’écran, Tracé, diagramme

Description générée automatiquement**

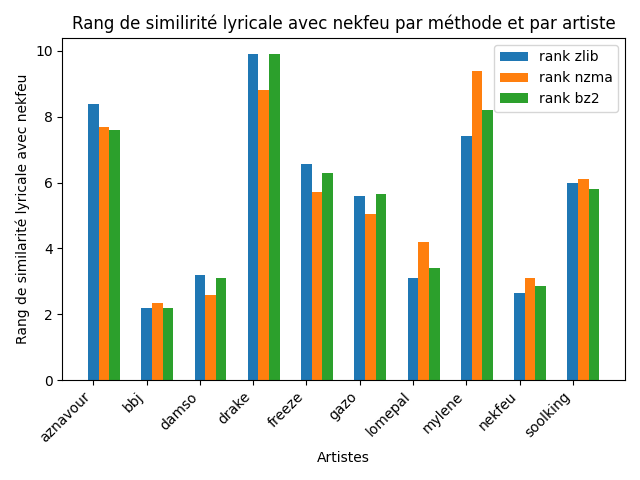
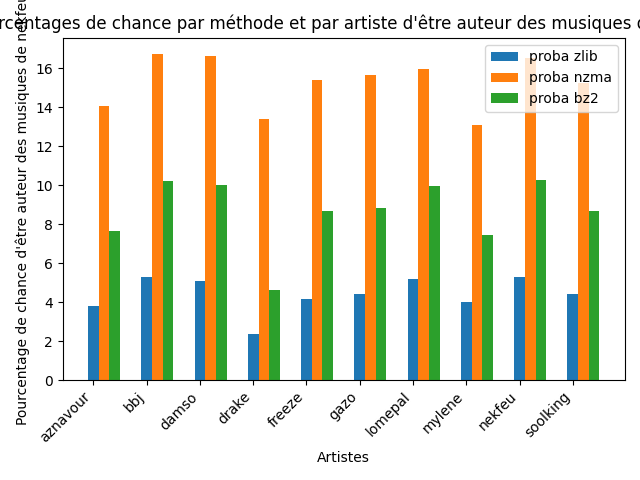
**Lomepal:**

** **

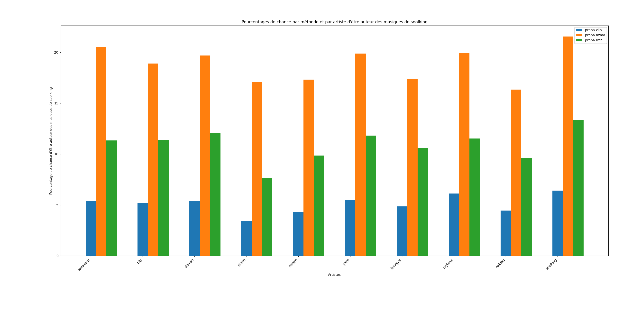
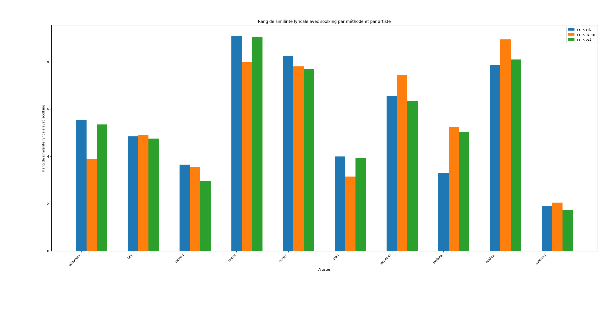
**Mylène Farmer:**

****

**Nekfeu:**

****

**Soolking:**

****

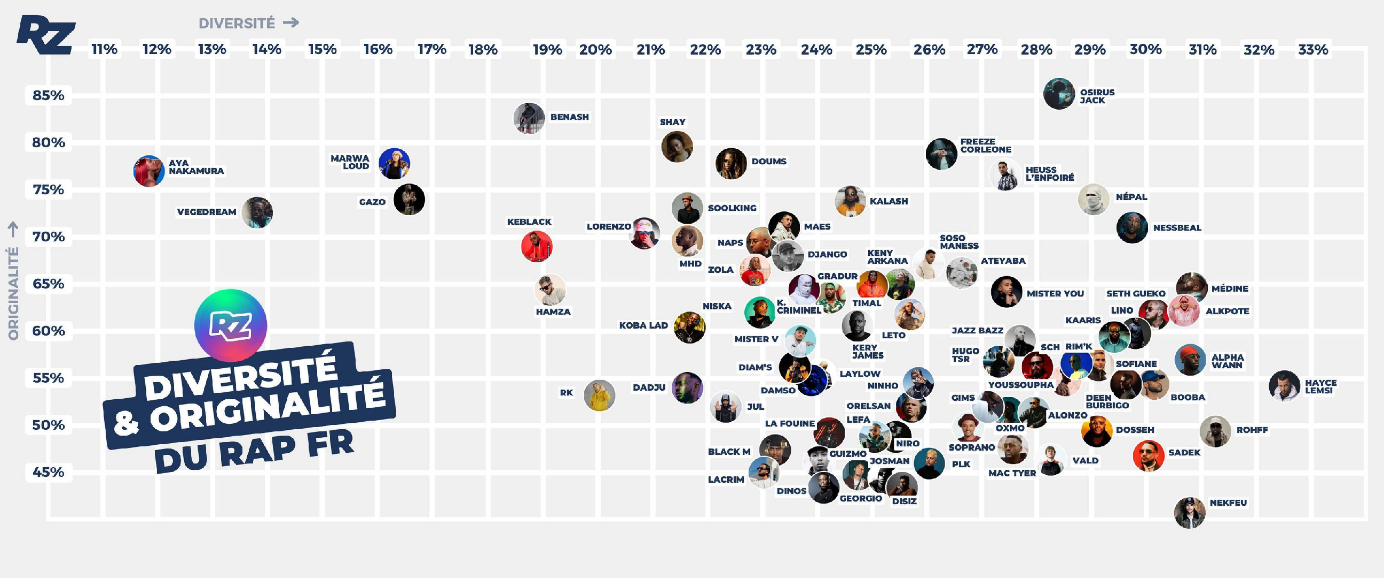
## Discussion

First, our results are quite good at first glance. On average, our program succeeded in identifying the artist behind each sound, except for Nekfeu and Lomepal. On the contrary, our tool was very good at detecting the author behind the lyrics of Charles Aznavour, Mylène Farmer and Drake.

The interpretation is as follows: for Drake's lyrics, he is almost the only one of the ten artists to write in English, which makes his vocabulary very unique. For Mylène Farmer and Charles Aznavour, their vocabulary is also very distinct from that of the other artists, as they are the only non-rappers. What's more, these two artists are very similar, and their results are reciprocal: Charles Aznavour's test texts are very close to Mylène Farmer's training data, and vice versa.

For the failure of Nekfeu and Lomepal (and the great success of Gazo and Freeze Corleone), we can rely on their diversity and originality. An artist's diversity represents the extent to which his or her lyrics vary lyrically from one to the next. An artist's originality represents the extent to which their lyrics are lyrically different from those of other artists. But for our tool to work, the artist has to be not very diverse (i.e. his lyrics have to be similar, so that the NCD between his test lyrics and his training lyrics is as low as possible) and very original (i.e. his lyrics have to be different from those of other artists, so that the NCD between his test lyrics and his training lyrics of other singers is as high as possible).

However, in the French rap landscape, Nekfeu and Orelsan are very diverse and not very original, unlike Freeze Corleone and Gazo (see previous graph). This correlation may be one explanation for the relative success of our tool.



*Diversity and originality of artists in the French rap landcape [1]*

For the three compression methods for calculating the NCD, nzma seems to give slightly less relevant results than the other two, based on the results explained in the graph above, but also using the number of songs correctly attributed to their artist-author. We have no explanation based on how the methods work for this difference.

We have identified several areas for improvement in our system. The first is to improve results. We had in fact planned to implement NGD in addition to NCD. NCD gives an approximate distance on the form between two texts, without really considering the meaning of the words (even if we tried to limit this problem with lemmatization). To overcome this problem, we decided to carry out a frequency analysis of the vocabulary of the test texts to extract the most redundant words. We could then calculate the NGD between these words and the name of each of the ten artists, to find out who was most likely to have written these words. However, we were hampered by Google's limit on the number of queries per IP address. Implementing NGD with Bing or Yahoo could be a solution.

The second improvement concerns the execution time of our algorithm. This is around 1 hour for each word processing method. The biggest problem is that each time two pieces of music are compared, the compressed size of each of the two texts alone is recalculated, whereas it would be sufficient to calculate the compressed size of the concatenation. A memoization method, or even more simply, an ancillary program that would create a dictionary associating each music with its compressed size would be easy to implement and welcome to drastically reduce program execution time.

## Bibliography

[1]: Rap Minerz graph on Diversity and Originality in French Rap landscape: <https://apps.rapminerz.io/diversity-originality/>