R lab

kmeans

Author: François Amat

Contact: amat.francois@gmail.com

Introduction:

In this lab, we want to use the kmeans algorithm to simplify some audio file by replacing the value by the centroid values.
We will test this approach on two differents files and with differents values of "k" in the kmeans algorithm.

I used the ZIP compression, I also modified the write script in order to automatise the compression process.

Script usage:

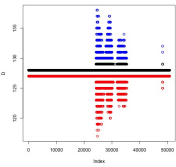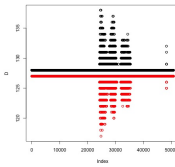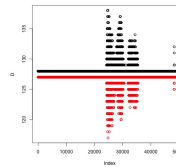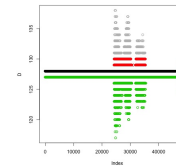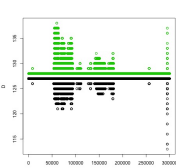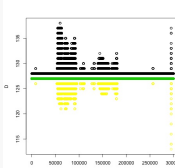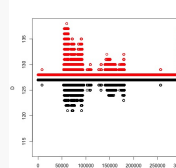I created the `compress.sh` script in the src folder (all the files can be found here)
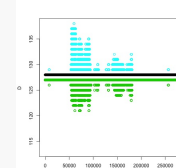
this script will take all raw audio files in the folder "/data/audio" and will compute the compressed file with the kmeans for all the differents k.

Results :

File	original	k = 5	k = 7	k = 15	k = 25
Aurevoir2.zip	7.1k	0.802	0.77	0.77	0.83
Trains.zip	23k	0.826	0.870	0.826	0.870

File	k = 5	k = 7	k = 15	k = 25
Aurevoir2				
Trains				

We can notice that sometimes an higher K does not mean a worse compression rate, that's because sometimes we found empty cluster in the algorithm. And therefore a k=25 can have similar results of the k=7 case such as the "Trains" file.

Conclusion:

I have been really surprised by the outputed quality after the compression effect with this algorithm, even with a low k we can understand the word prononced.
I also have been surprised by the compression rate, the gain can be very high (up to 23% mesured !).
This lab helped me to get started with R, and explore the kmeans algorithm.