FLATIRON SCHOOL

Data Science Bootcamp

FINAL PROJECT

Hip-Hop Clans

Using Natural Language Processing (NLP) and Machine Learning to study 27k+ US rap songs

Fabrice MESIDOR• 24 Oct 2019

Overview

JUSTIFICATION

Music genres classification is very subjective

Music is more about taste - preferences diverge

Finding similarities among songs using ML

OBJECTIVES

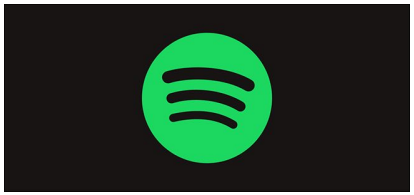
Find out different categories / topics of rap songs

HYPOTHESES

There are differences in Hip-Hop songs

Nature of the music vary over the years

What Data I am working with?



Ranker

Billboard HOT 100

55k data from Genius

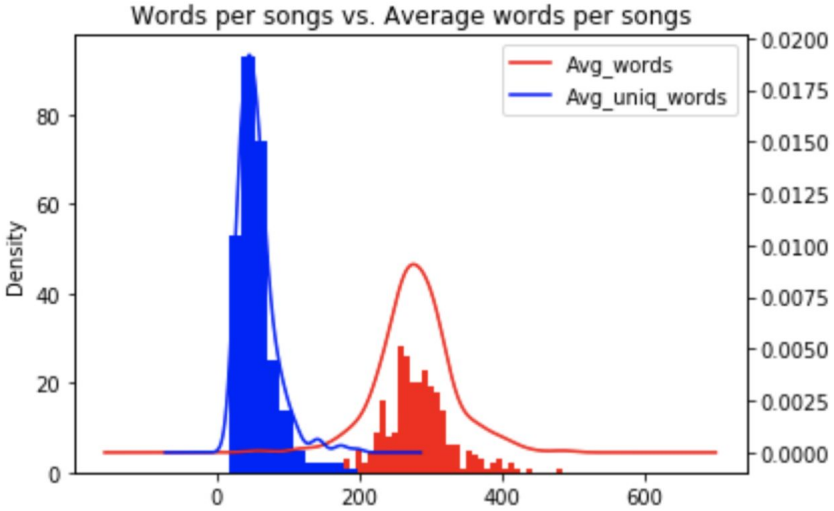
28k with Spotify features

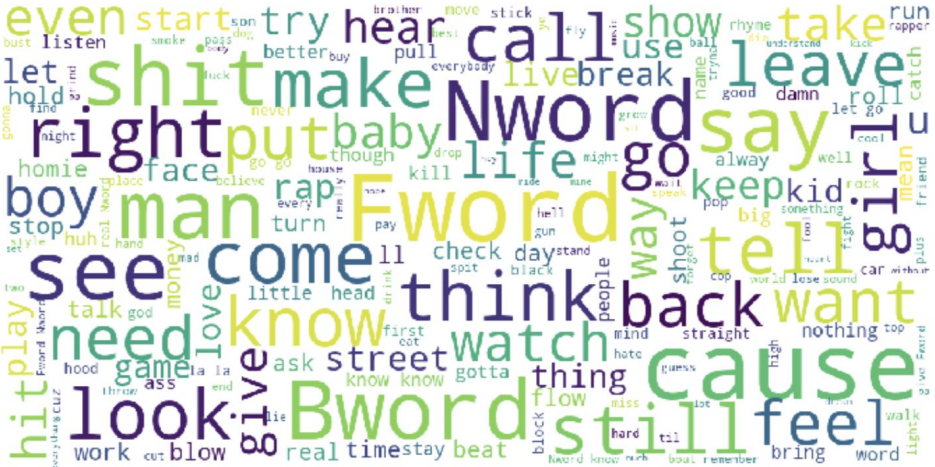
27k in English

From 1979 to 2019

- Lyrics
- Number of words
- Acousticness
- Popularity
- Danceability --> how suitable a track is for dancing
- Energy --> perceptual measure of intensity and activity.
- Instrumentalness --> Rap or spoken word tracks are clearly “vocal”.
- Liveness --> presence of an audience loudness --> in decibels (dB).
- Speechiness --> presence of spoken words in a track.
- Tempo --> beats per minute (BPM). time_signature --> beats / bar (or measure).
- Valence --> the musical positiveness

Understanding rappers vocabulary





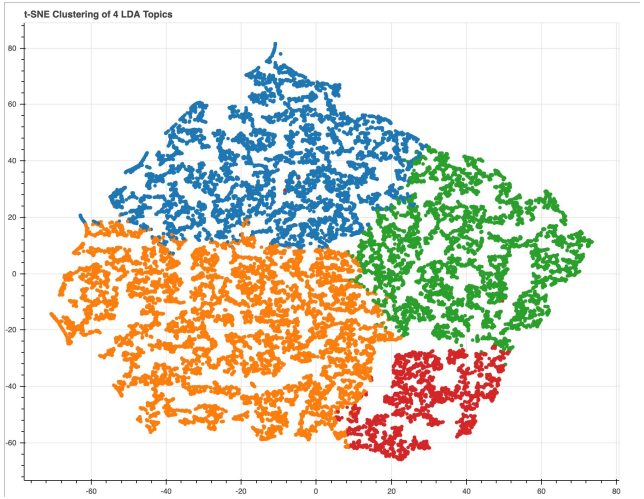
Nearly a sixth of the words used in rap are unique

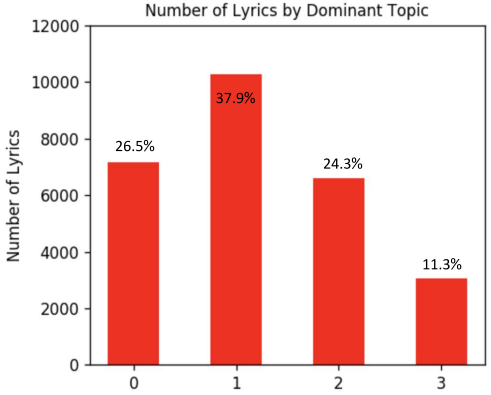
~300 words used per songs in average vs. 50 unique in average

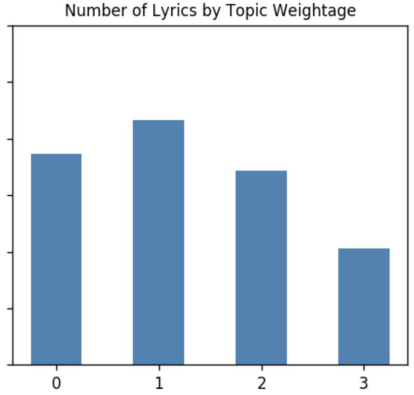
Lot of coarse words

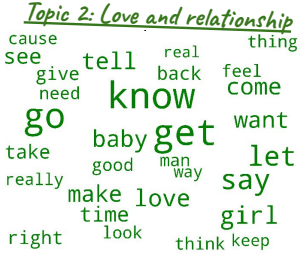
Repetition of words/sentences (go go go go shawty…) to make the songs catchy

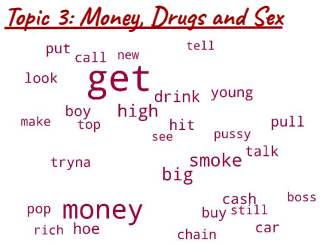
Most frequent words & Topics



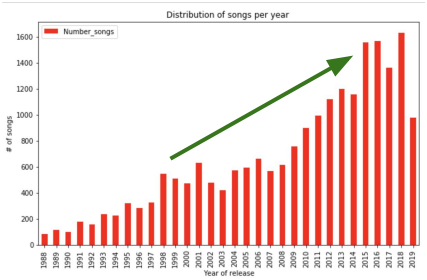


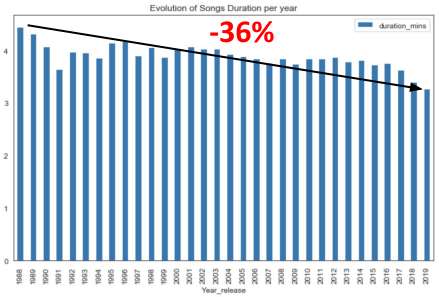


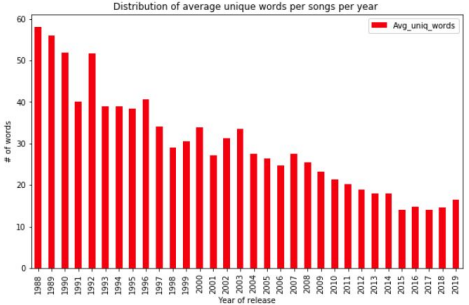




Rap over the years







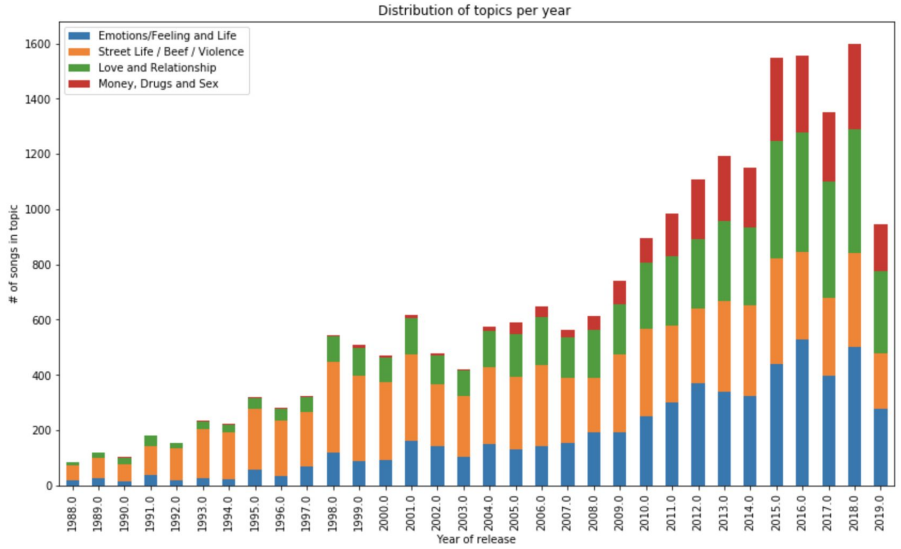
+ Increase of music production
However, less and less words

+ Shorter Duration →

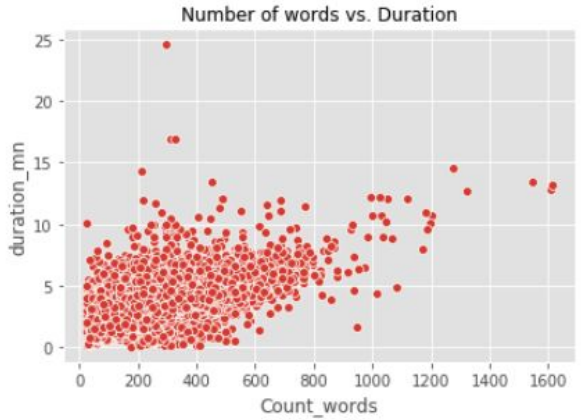
THE STREAMING EFFECT

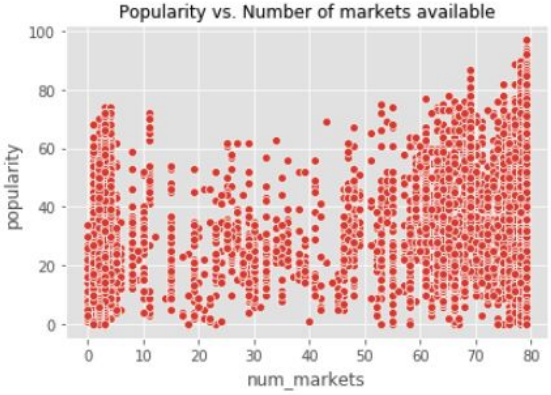
+ Stuffing more diminutive songs into an album is simply more remunerative than having a bunch of long ones

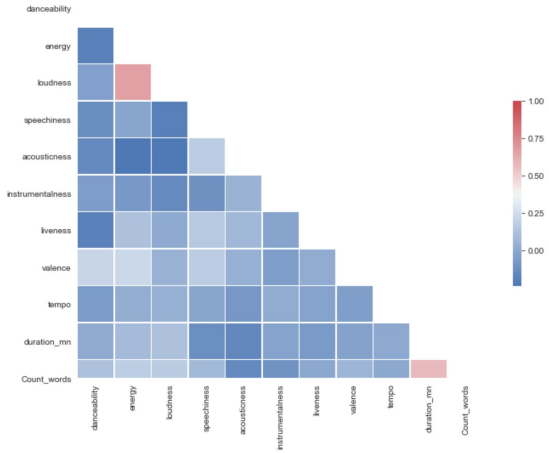
+ Increase of songs related to money/drug around 2004

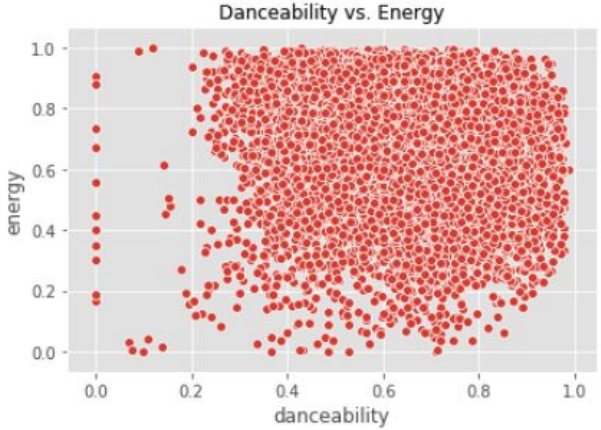


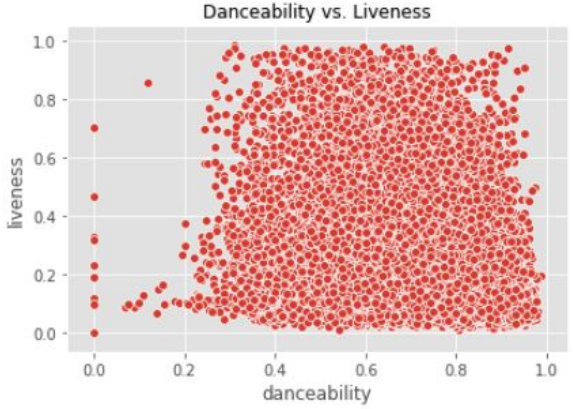
What songs are we listening to?











- Choice between loudness and energy
- Popularity can be considered as an initial cluster factor
- Different energy and liveness in hip hop

What songs are you listening to?

Cluster0

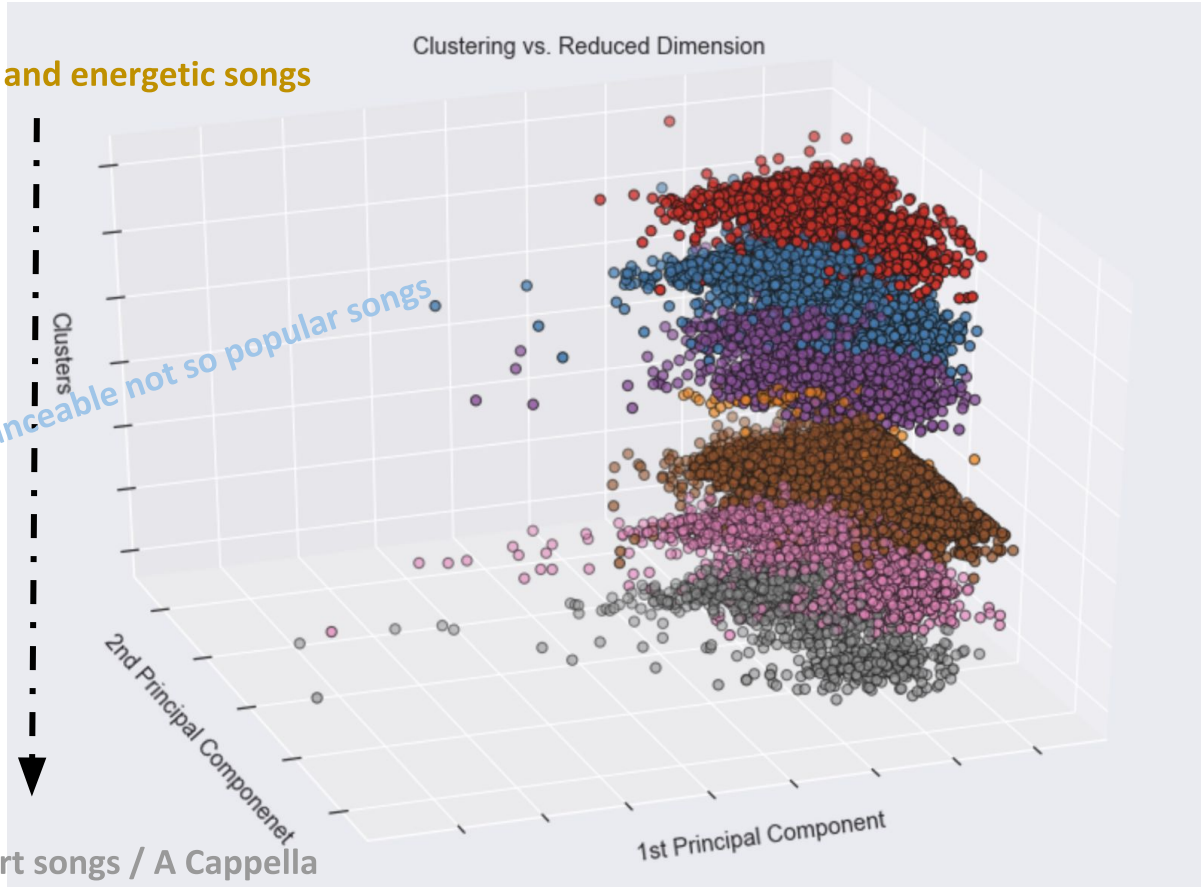
	Artist	Title	duration_mins	loudness	cluster
2367	Tech N9ne	No Can Do	3.94	-3.605	0
20560	Cypress Hill	A Man	4.20	-13.326	0
26275	Boosie Badazz	Roller Coaster Ride	3.51	-7.830	0
26002	Nicki Minaj	Miami	3.18	-7.084	0
968	DMX	Where The Hood At (High On X Remix)	4.91	-5.891	0

- ★ K-Means clustering was used on Spotify scaled features and IF-IDF
- ★ 7 Clusters identified with Elbow and Silhouette methods

Cluster5

	Artist	Title	duration_mins	loudness	cluster
12271	Ja Rule	Intro Superstar	0.27	-22.209	5
12913	Atmosphere	Secret	0.32	-26.830	5
749	Chamillionaire	The Final Chapter	0.35	-12.375	5
7920	A$AP Ferg	Uncle	0.38	-5.593	5
22794	Kurupt	Intro	0.64	-17.699	5
7205	Missy Elliott	Religious Blessings - Outro	0.65	-18.310	5
16642	Cormega	Reflection	0.69	-6.677	5
27018	Atmosphere	The Ocean	0.71	-17.234	5
8811	Pitbull	Outro	0.73	-8.116	5
800	Eminem	Hazardous Youth	0.73	-9.502	5
15260	Cam'ron	Intro 2	0.74	-12.797	5
19327	2Pac	Starry Night	0.80	-8.591	5
20938	Busdriver	Drivers Manual	0.83	-13.961	5

Long popular and energetic songs



Catchy Short songs / A Cappella

- ★ Establish 4 topics on 27k+ lyrics using Latent Dirichlet Allocation
- ★ 7 Clusters found using K-Means
- ★ Evolution of the music industry over the years

next

- → Recommendation system in order to find similar songs (topics and audio features)
- → RNN to create part of songs