

# Hip-Hop Clans



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



# 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

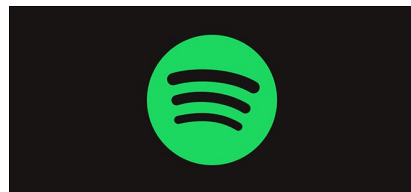
## HYPOTHESES

There are differences in Hip-Hop songs

Nature of the music vary over the years

---

# What Data I am working with?



**Billboard**  
**HOT 100**



55k data from  
Genius

28k with Spotify  
features

27k in English

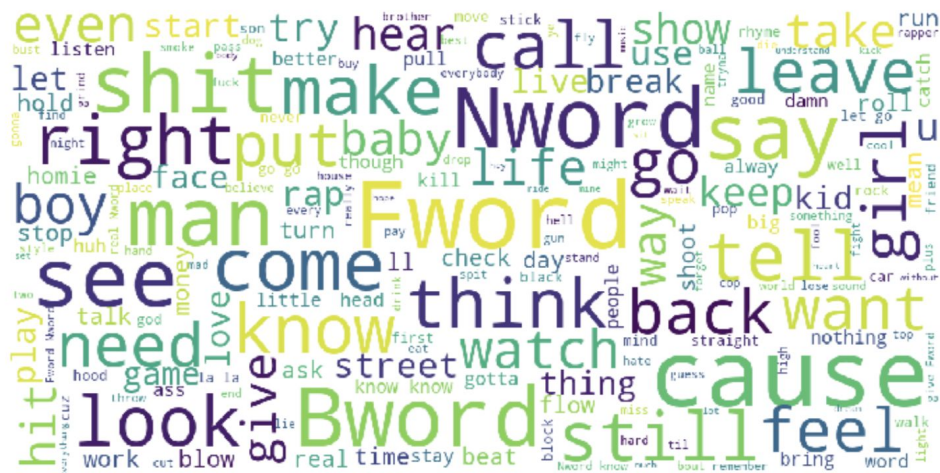
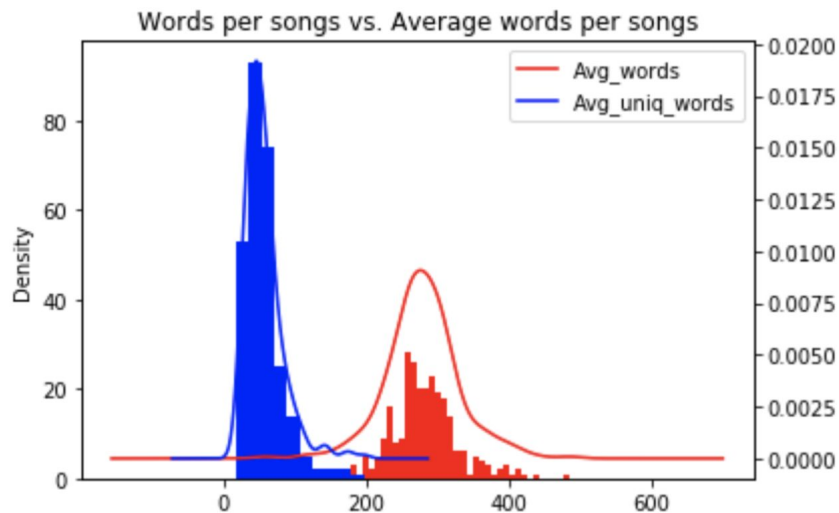
From 1979 to 2019

## Features

- ❖ 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

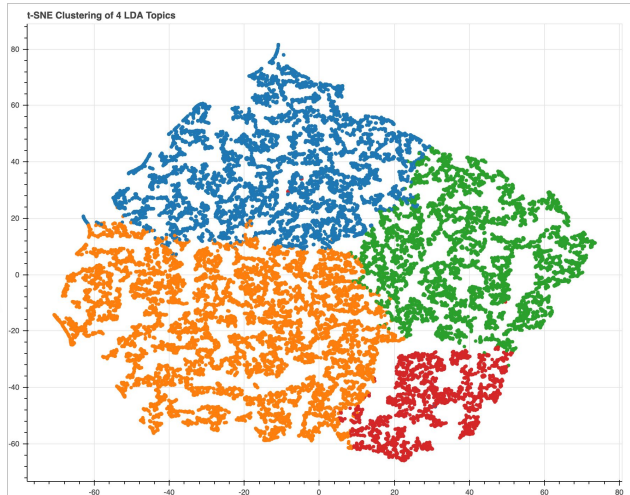
# Most frequent words & Topics

## Topic 0: Emotions/Feeling and Life

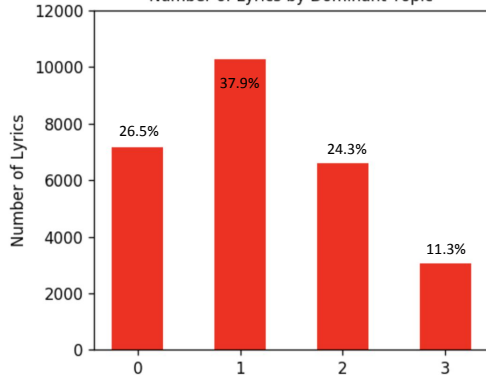
lose see live make die  
mind tell look day try  
think feel thing  
many man say never  
life take know time  
way find world  
people use still black

## Topic 1: Street Life / Beef / Violence

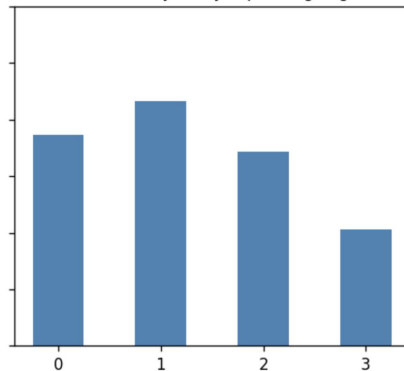
put gun game ass stop  
head know hear back play  
go let real  
cause keep kill make  
take run see  
street man beat flow try  
give big hit



Number of Lyrics by Dominant Topic



Number of Lyrics by Topic Weightage



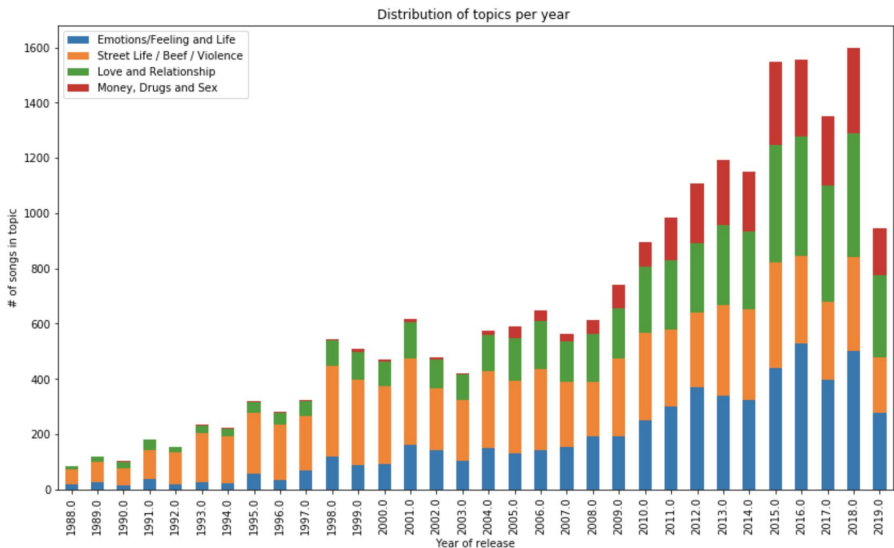
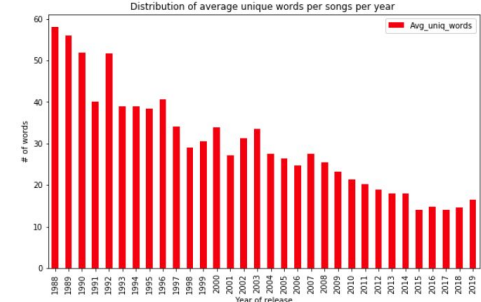
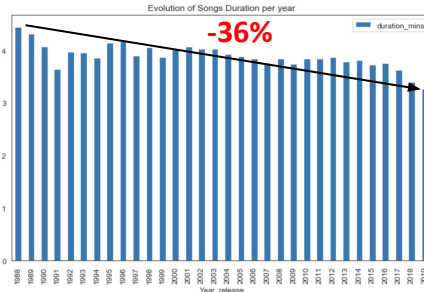
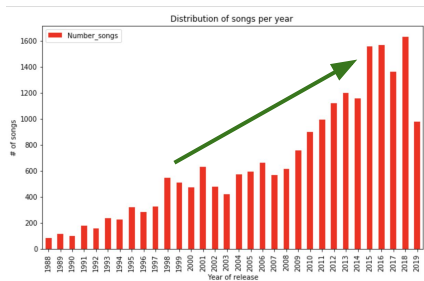
## Topic 2: Love and relationship

cause see tell real thing  
give need know back feel  
go baby get want  
take good man way let  
really make love say  
time look girl  
right think keep

## Topic 3: Money, Drugs and Sex

put call new tell  
look get drink young  
make boy top high hit pussy pull  
tryna see smoke talk  
big  
pop money cash boss  
rich hoe buy still car  
chain

# Rap over the years



+ Increase of music production  
However, less and less words

+ Shorter Duration →

## THE STREAMING EFFECT

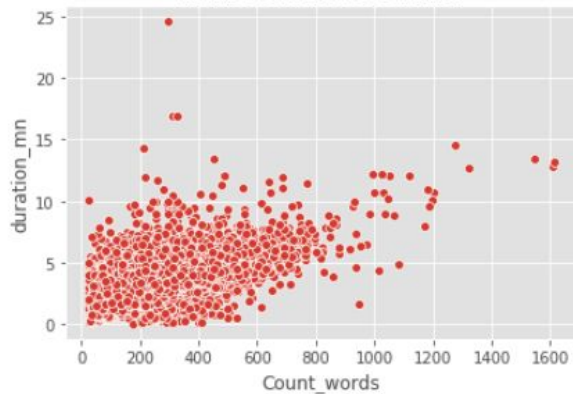
+ Stuffing more diminutive songs into an album is simply more remunerative than having a bunch of long ones

+ Increase of songs related to money/drug around 2004

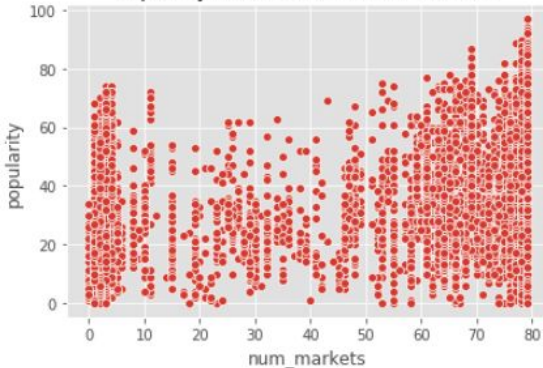


# What songs are we listening to?

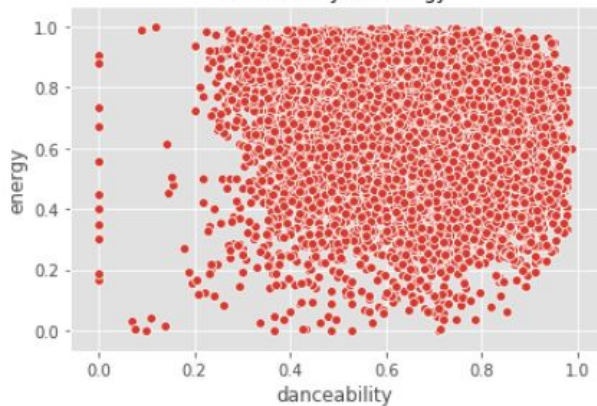
Number of words vs. Duration



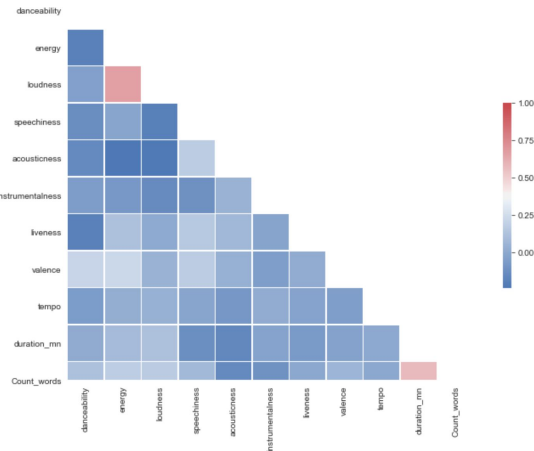
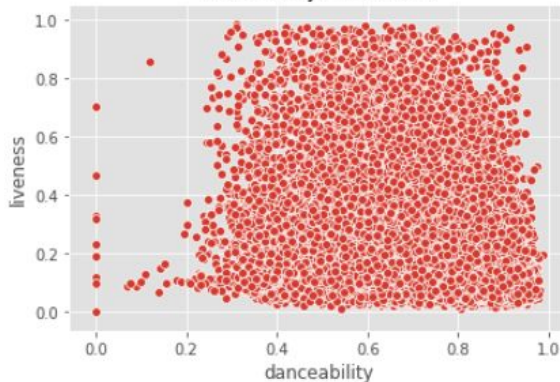
Popularity vs. Number of markets available



Danceability vs. Energy



Danceability vs. Liveness



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

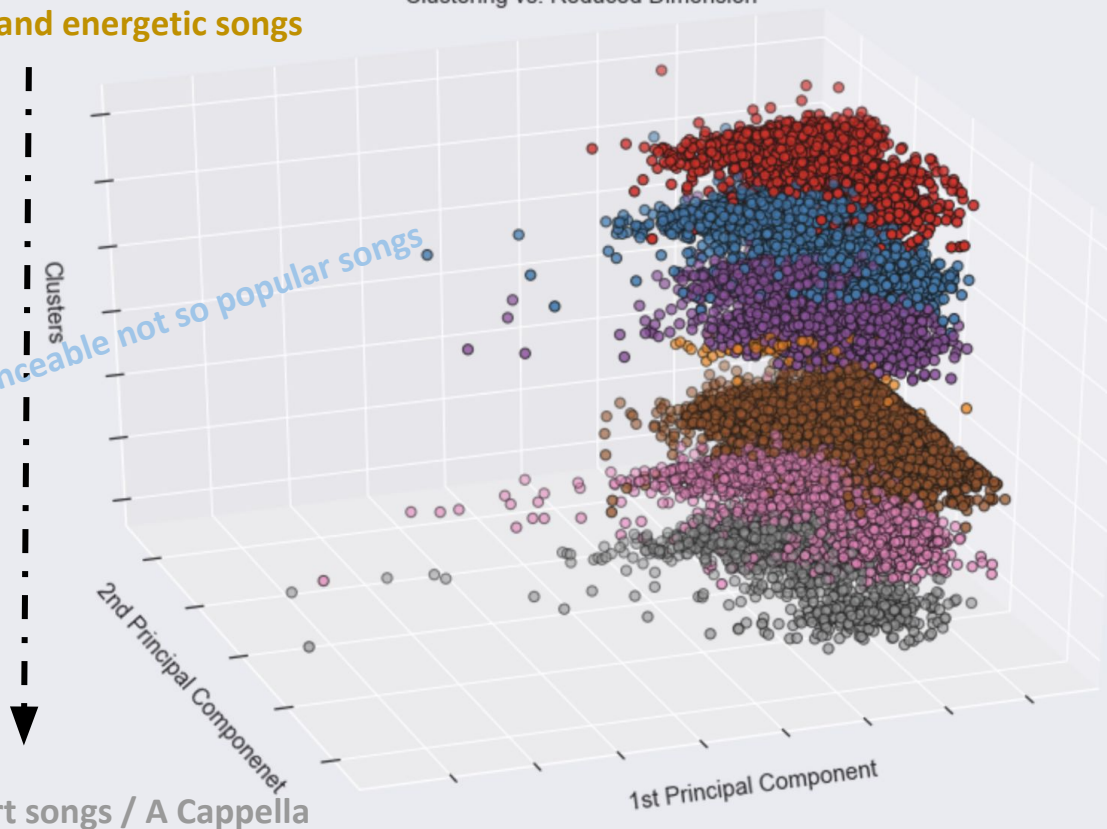
# What songs are you listening to?

## Long popular and energetic songs

Danceable not so popular songs

Catchy Short songs / A Cappella

Clustering vs. Reduced Dimension



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
- ★ 6 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	ASAP 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
15280	Cam'ron	Intro 2	0.74	-12.797	5
19327	2Pac	Stann Night	0.80	-8.591	5
20938	Busdriver	Drivers Manual	0.83	-13.961	5





- ★ 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