Spotify Data Analysis

Onderscheid maken tussen genres in Spotify

Hoofdvraag

Is het mogelijk om muziekgenres statistisch te discrimineren?

Zijn muziekgenres statistisch te discrimineren aan de hand van features?

Mogelijke implicaties:

- Aanbevelingen kunnen doen op basis van luistergeschiedenis
- Statistische diversiteit tussen/binnen genres bewijzen
- De muziek-features die het grootste onderscheid tussen genres geven bepalen

Features

Spotify's machine learned features

Danceability	Energy	Speechiness
Acousticness	Instrumentalness	Liveness
Valence	Loudness	

Vaste features

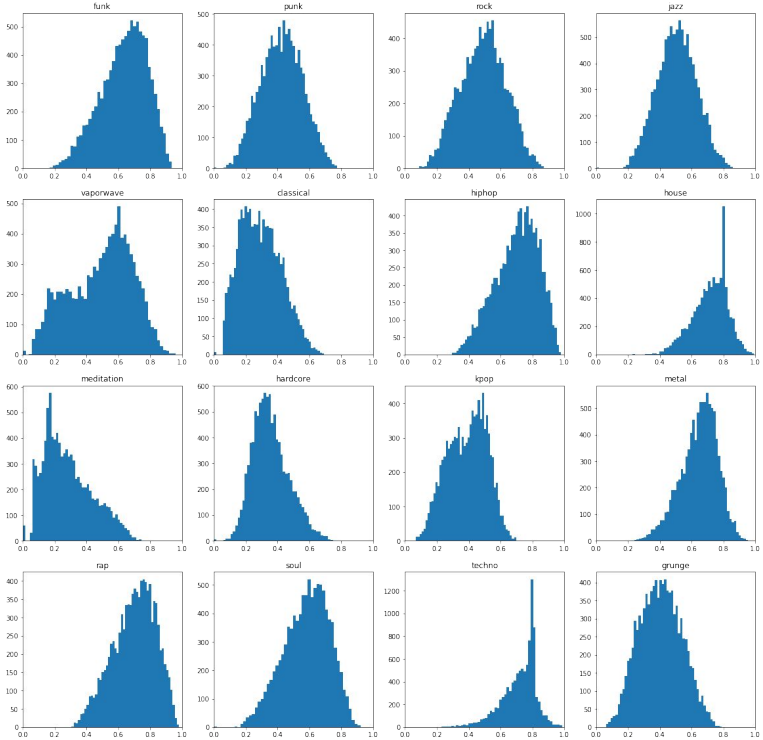
Key	Mode
Tempo	Duration
Time signature	

Tekstuele features

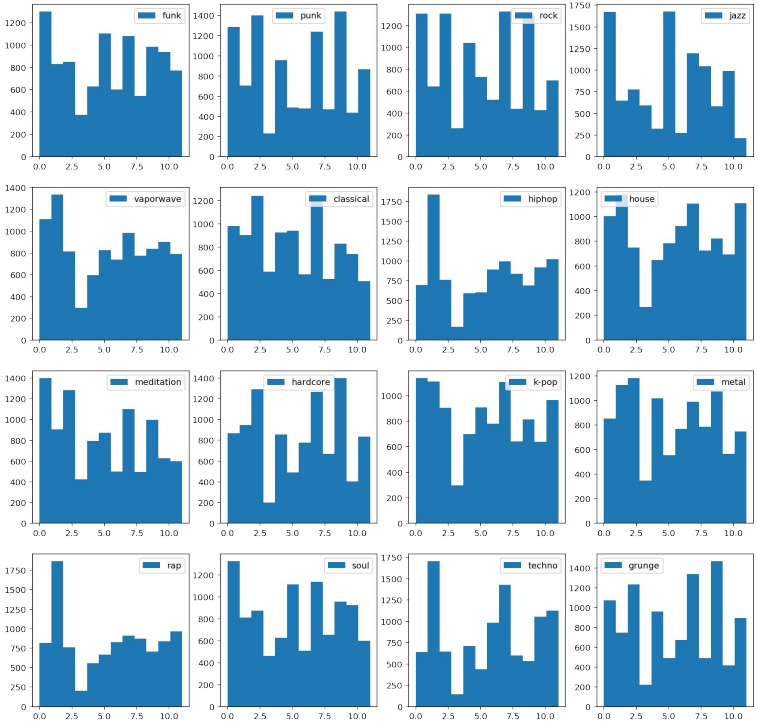
Titel	Songtekst

Feature visualisatie

Danceability

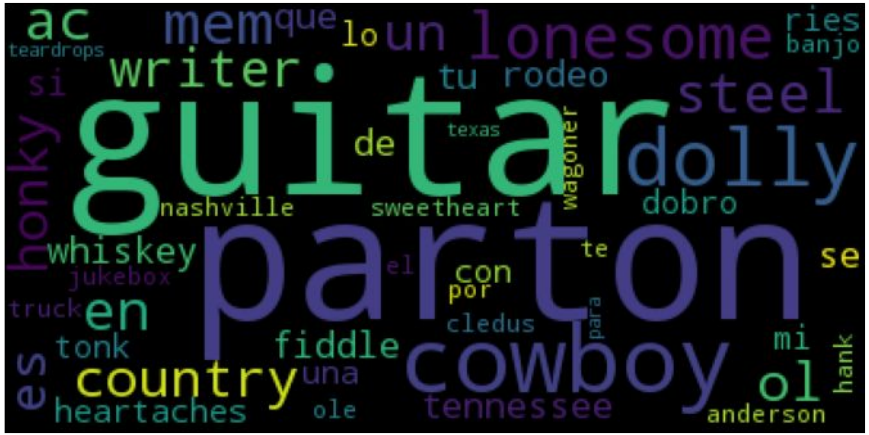


Key

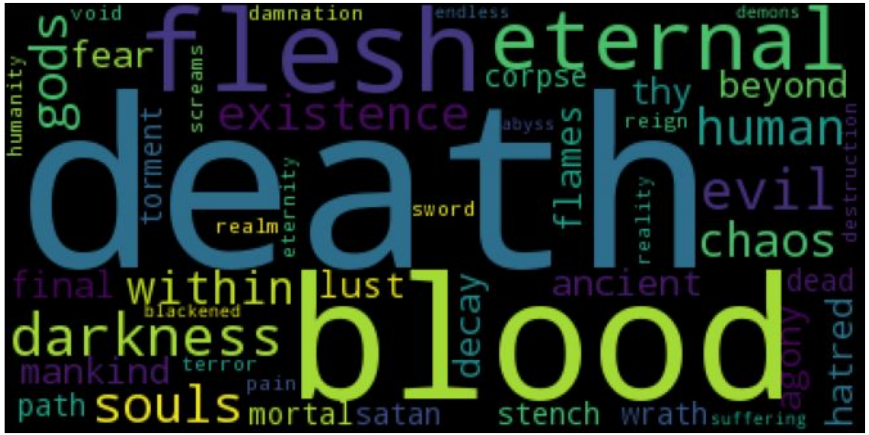


Feature visualisatie - songtekst

?

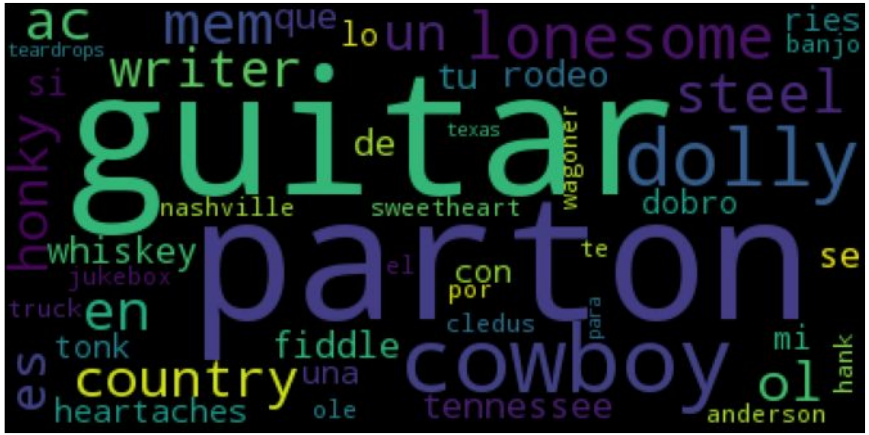


?

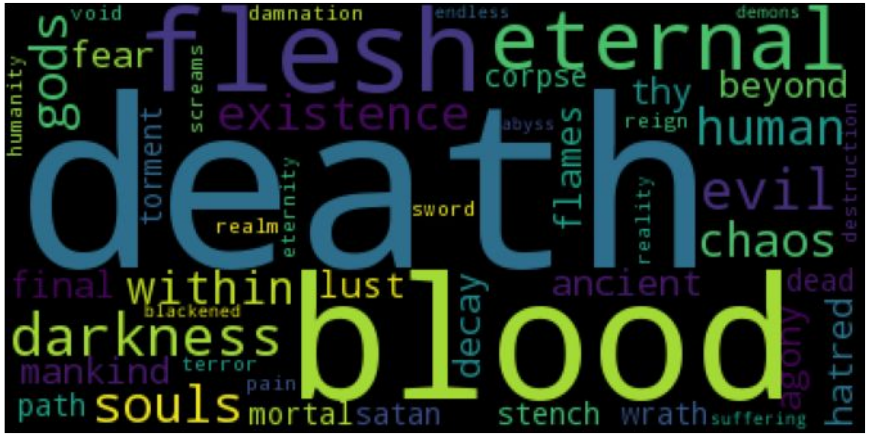


Feature visualisatie - songtekst

Country

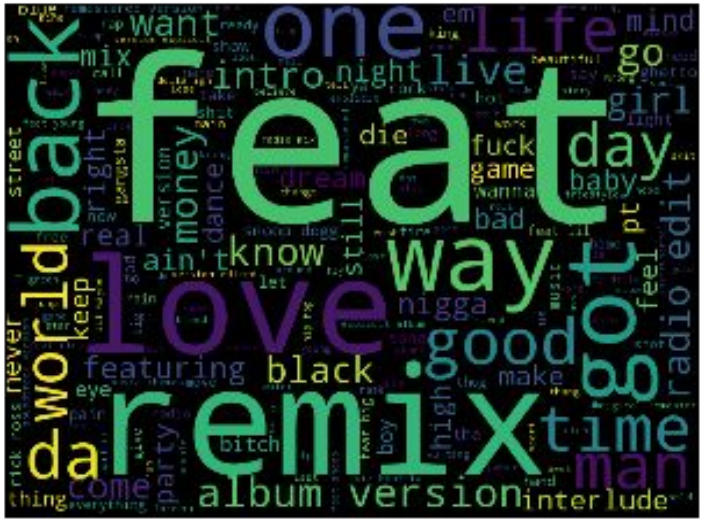


Metal

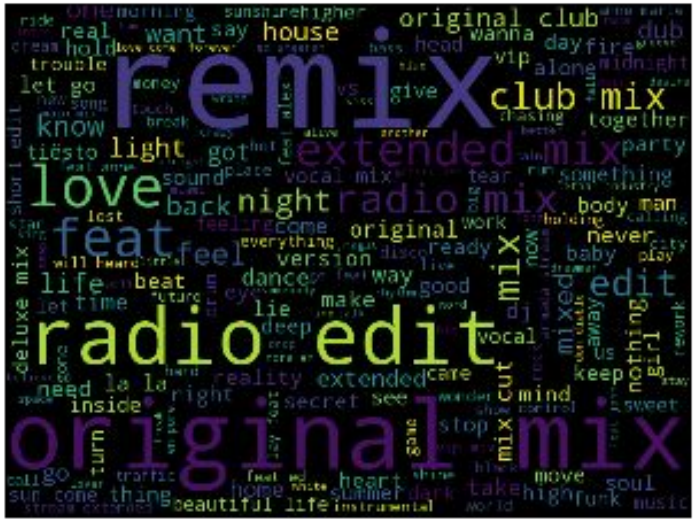


Feature visualisatie - Titels

Hiphop

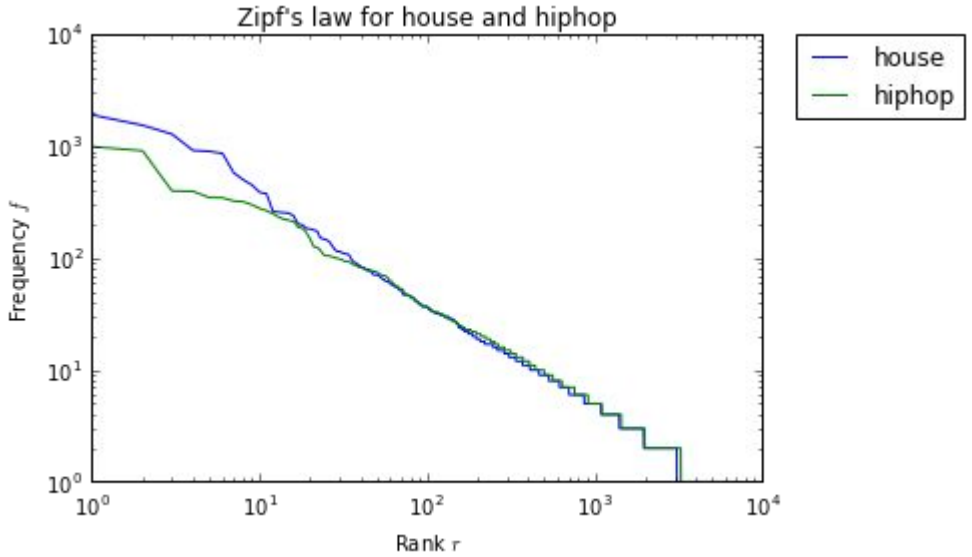


House

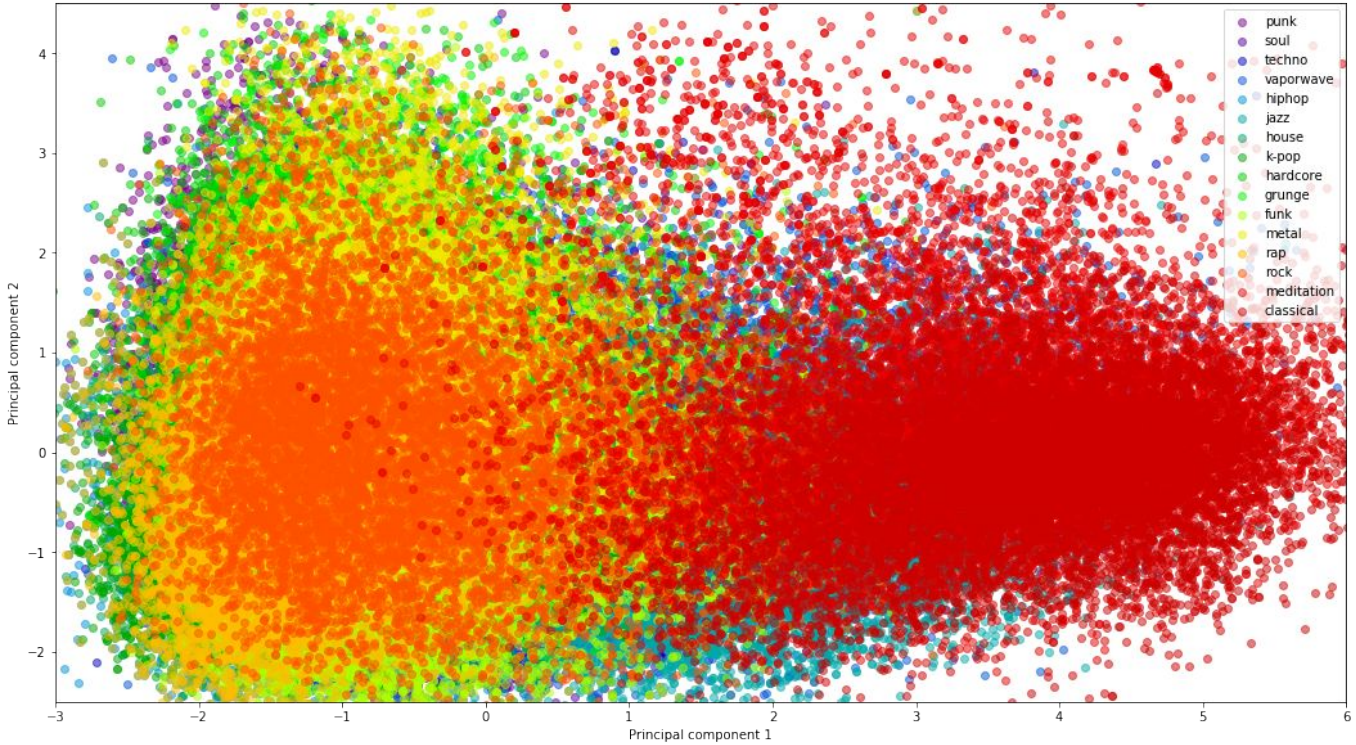


Titels

	Aantal woorden	Aantal unieke woorden
Hiphop	33462	7473
House	41201	6879

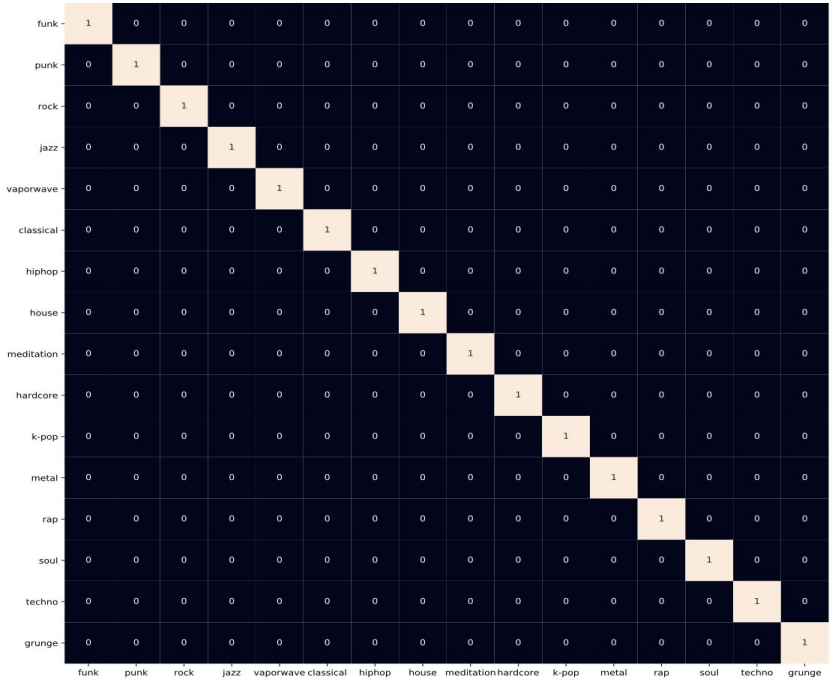


Feature visualisatie
Dimensionality reduction

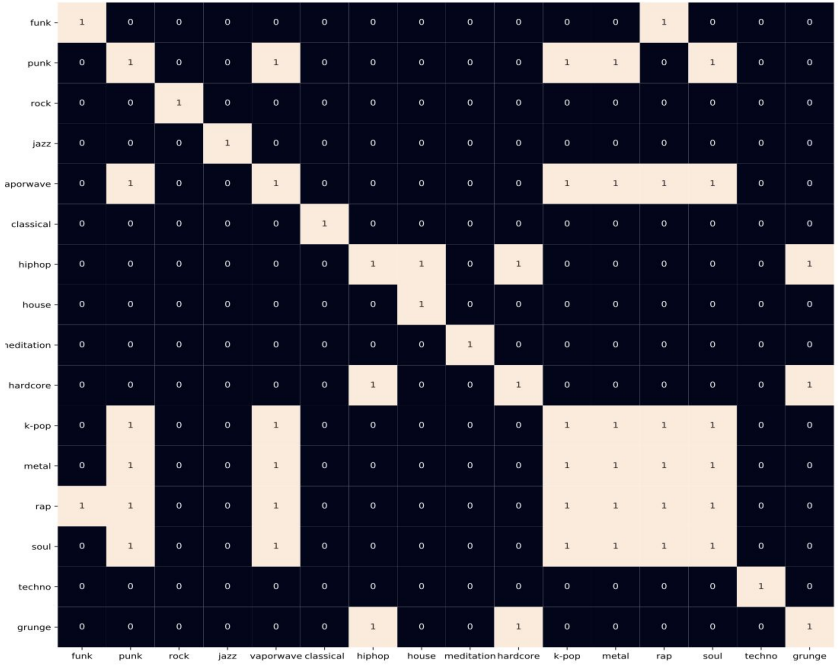


Two-sided Mean Test

Acousticness

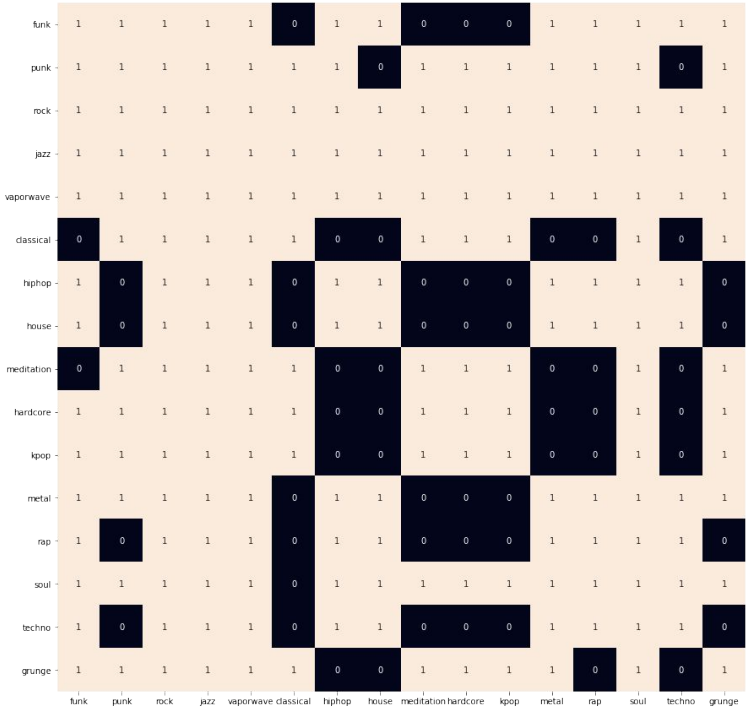


Key

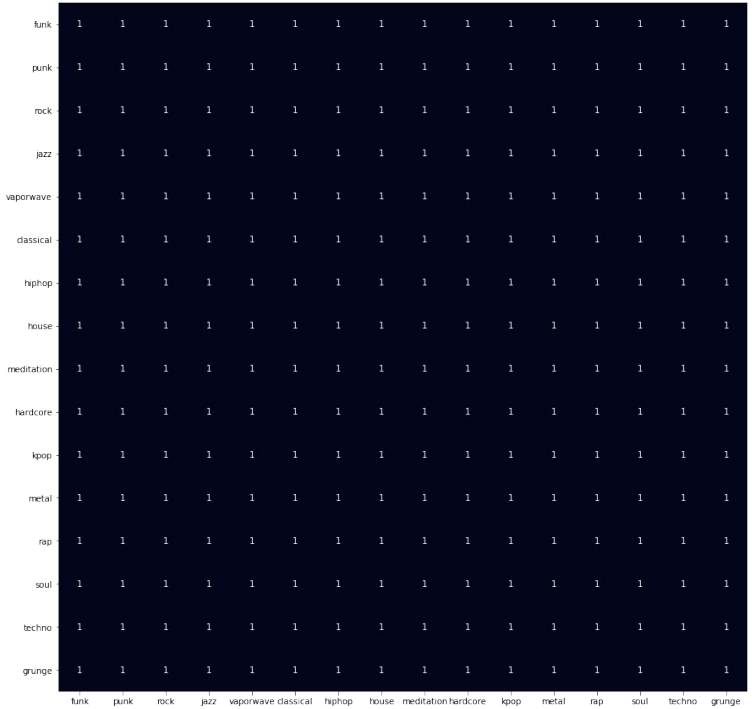


CI test: Welke features discrimineren genres?

Danceability

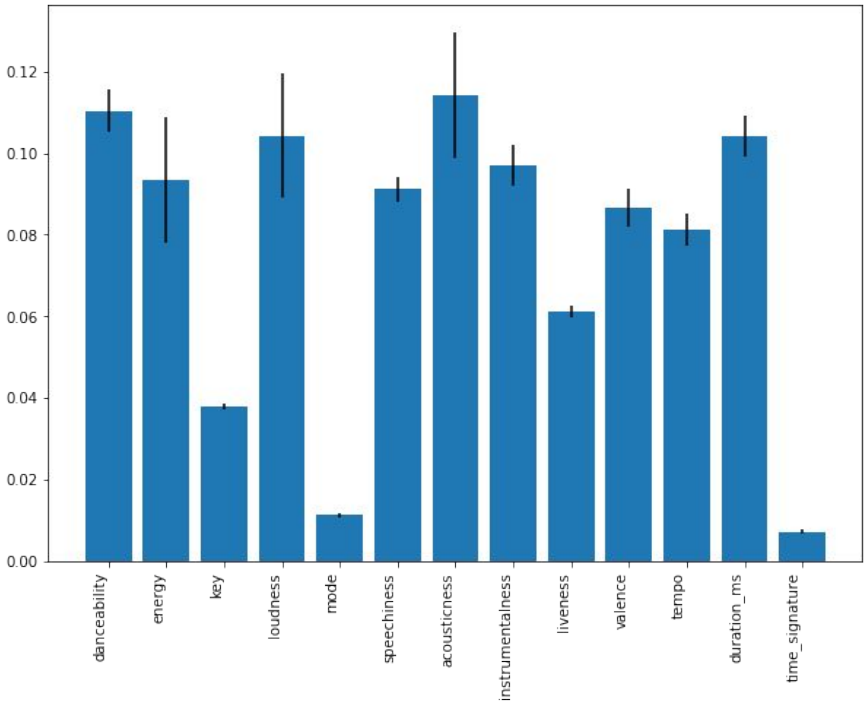


Key



Feature importance

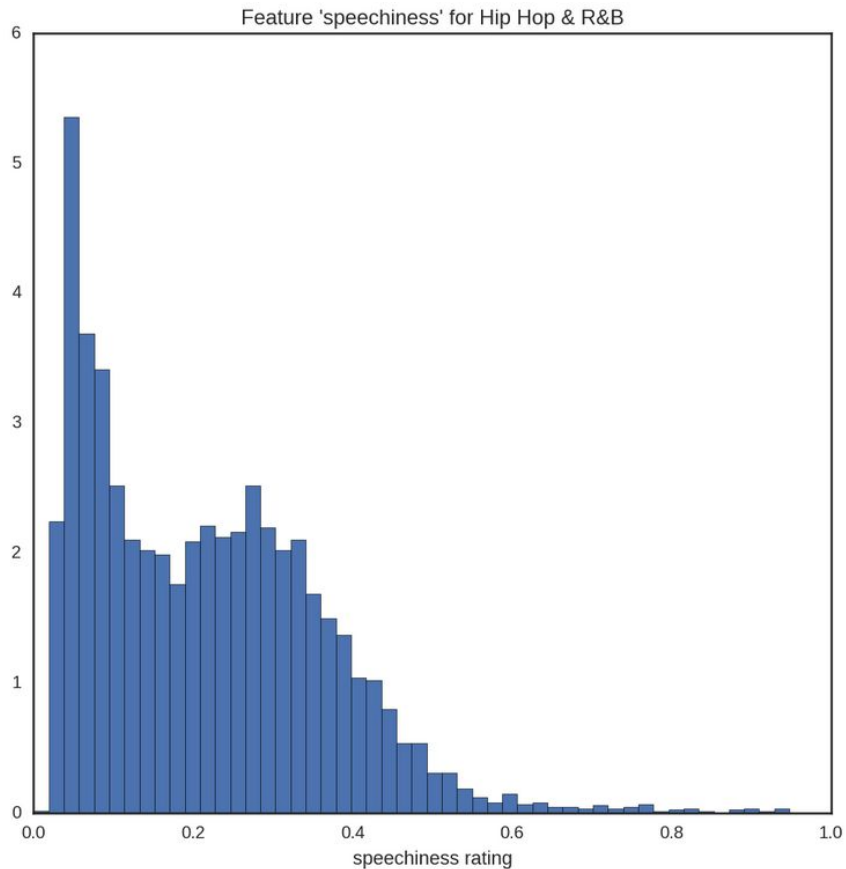
- Everything below valence gives a full-ones matrix



Speechiness in hiphop liedjes

10.000 hiphop liedjes gesampled uit spotify

Er lijken twee distributies in het histogram te zitten.

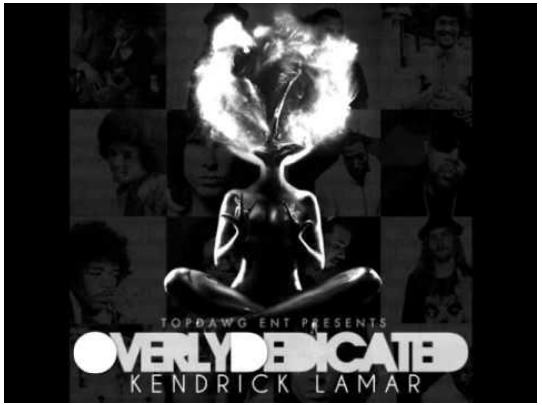


Clusteren hiphop liedjes met K means

Er waren geen labels, enkele uitkomsten zijn handmatig nagekeken

- 1 cluster met rustige hiphop, meer richting pop muziek
- Een ander cluster met meer rap muziek

Pop:



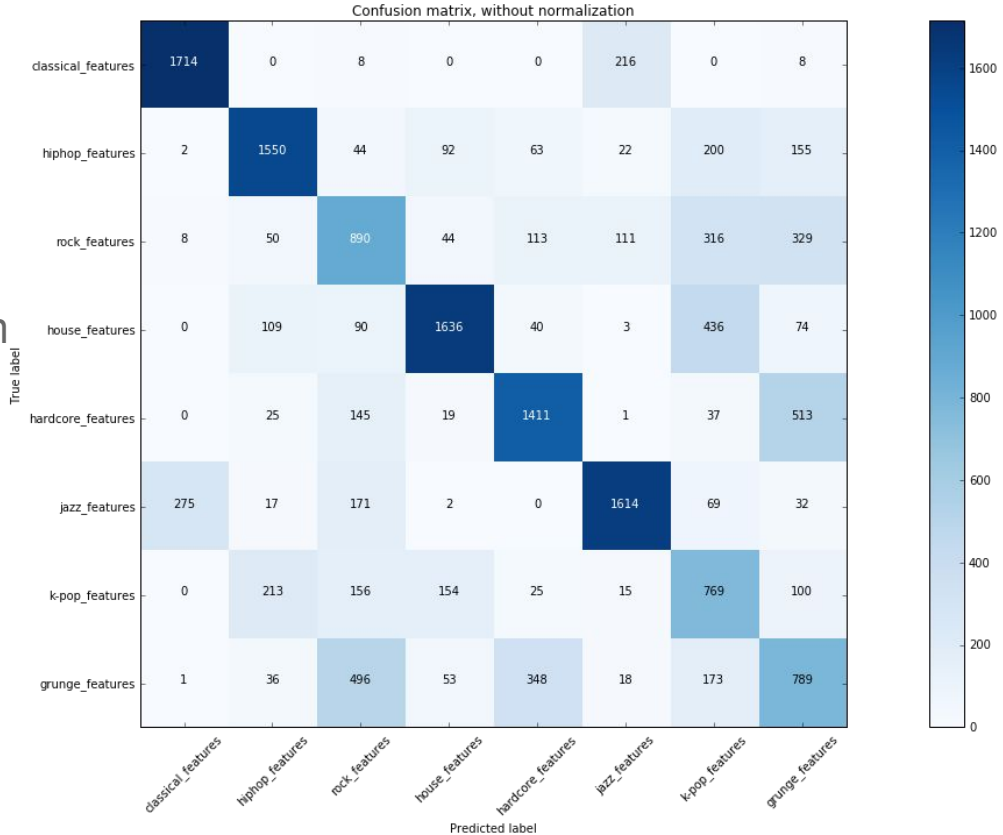
Rap

K nearest neighbors

10.000 samples per genre

- 60% voor training
- 20% voor zowel validatie als testen

65% accuracy, en veel herhaalde fouten



Playlists classificeren

Hypothese:

- Playlists bestaan vaak uit liedjes uit één genre.

Per genre 1000 playlists van 100 liedjes gegenereerd met de validatie-set

- classificeer de playlist
- bereken de cosine similarity met een rij uit de confusion matrix
- bereken confidence intervals

Geen van de echte playlists viel binnen het confidence interval.

MaxEnt logistic regression op basis van songtekst

- Kaggle dataset, voor aantal genres veel meer data beschikbaar dan voor de anderen
- Voor elk genre d.m.v. Mutual Information de 50 meest discriminerende woorden bepalen
- Met deze woord-features een MaxEnt classifier trainen, cross-validation gebruiken om confidence interval voor de nauwkeurigheid te bepalen