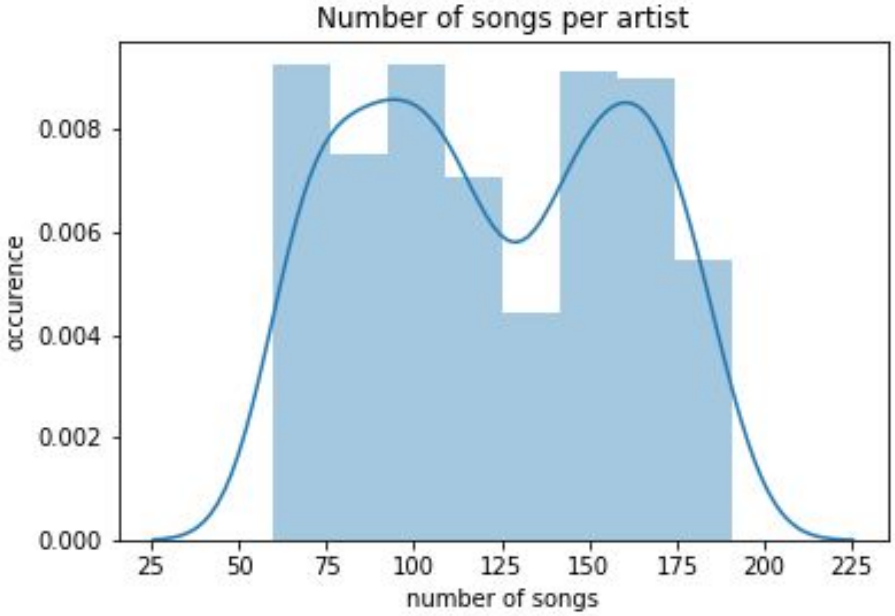
Clustering

Ist das Genre abhängig von Lyrics?

Erster Datensatz

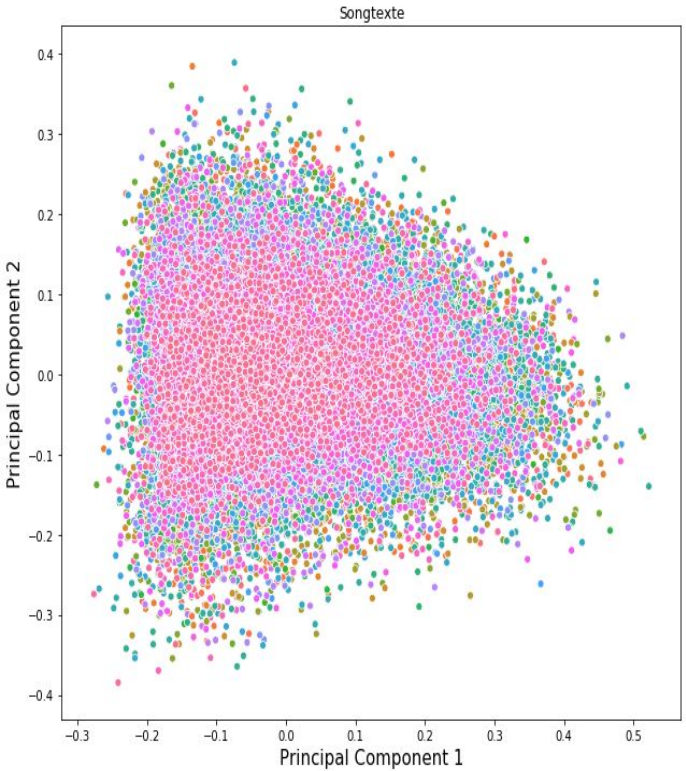


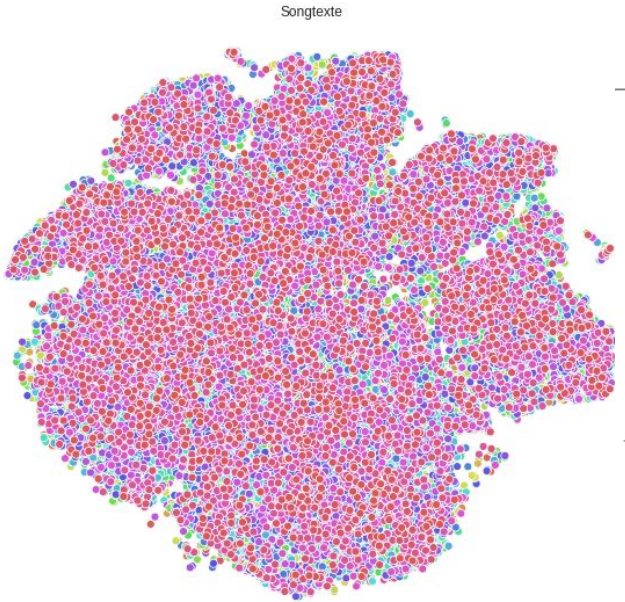
- 643 Bands
- 57 650 Songs

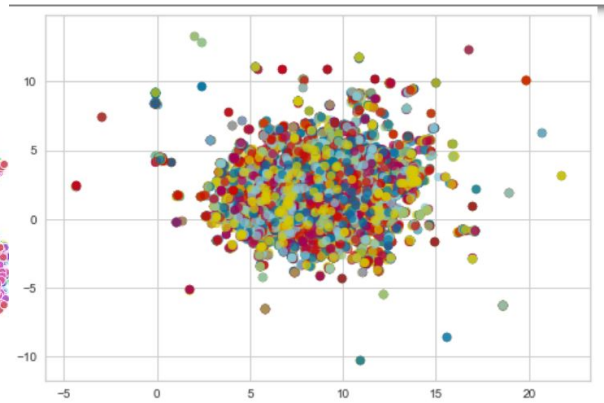
PCA

TSNE

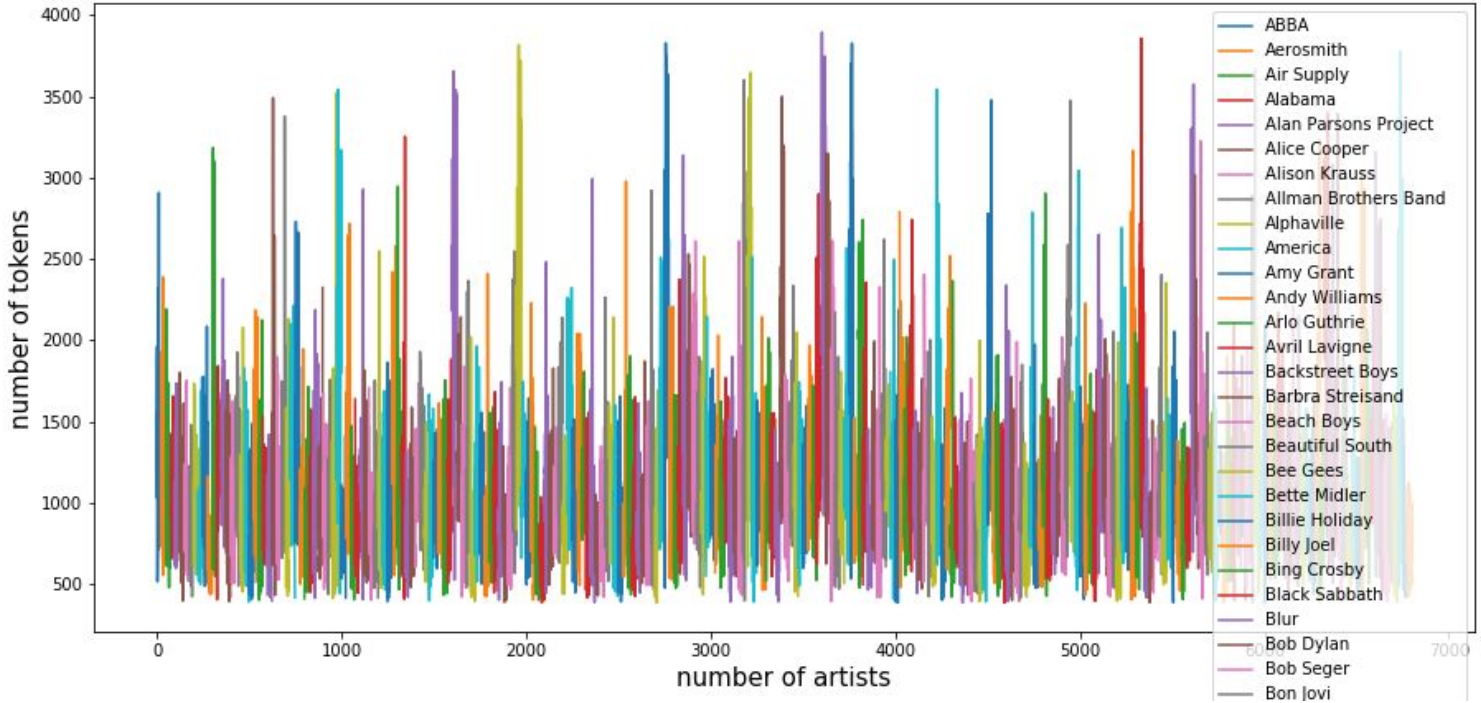
UMAP



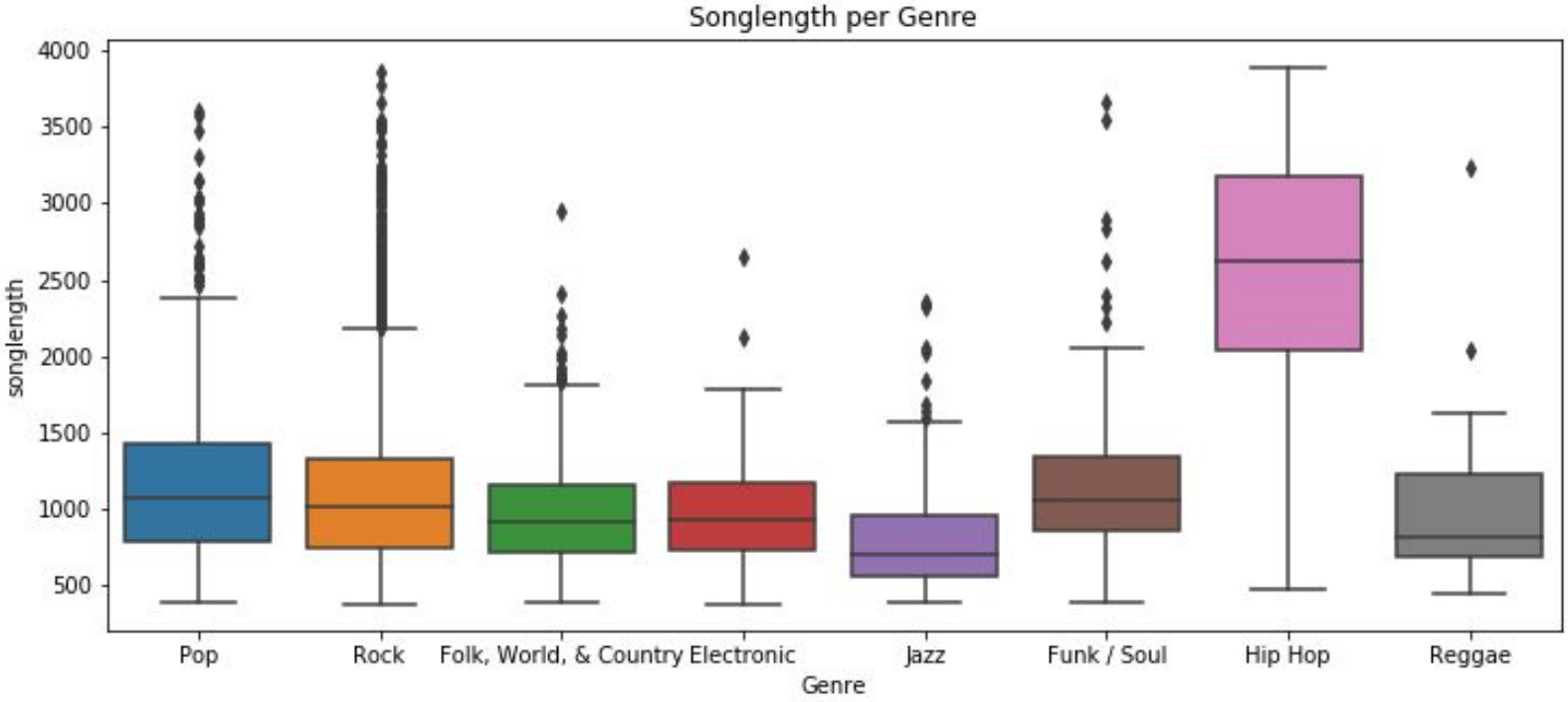




Datensatz anpassen

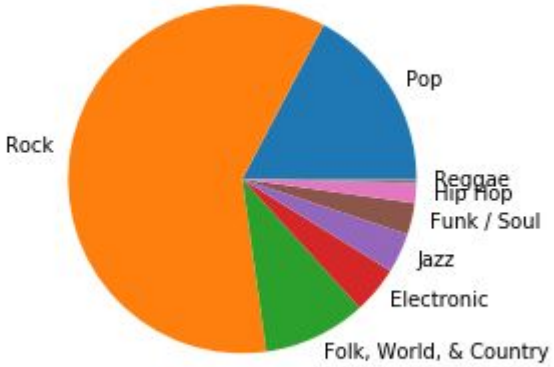


Genres

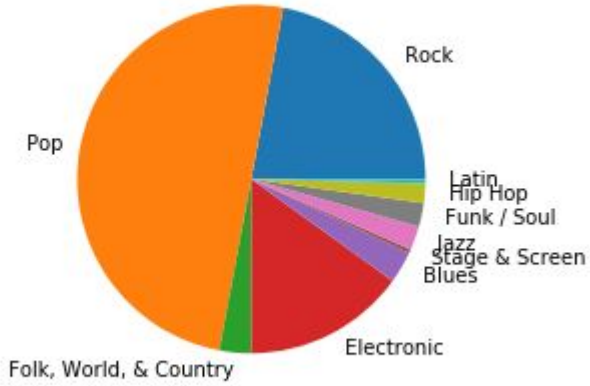

Songlength per Genre
songlength
4000
3500
3000
2500
2000
1500
1000
500
Pop
Rock
Folk, World, & Country
Electronic
Jazz
Funk / Soul
Hip Hop
Reggae
Genre

Genres Verteilung

Genre1

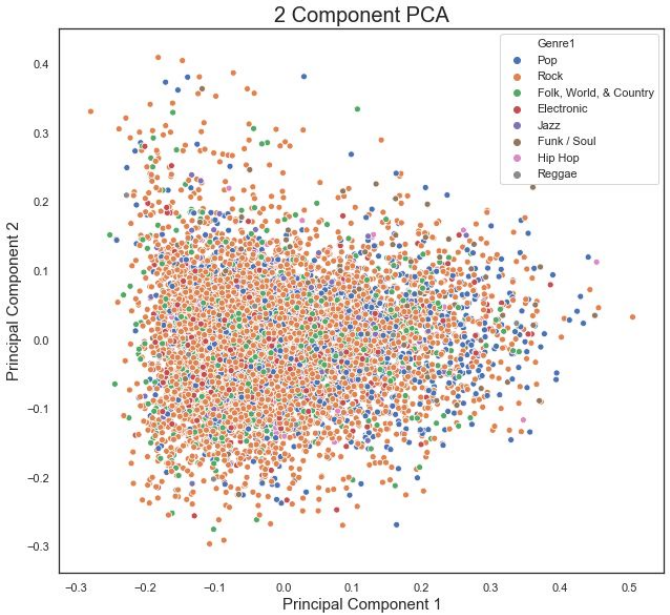


Genre2

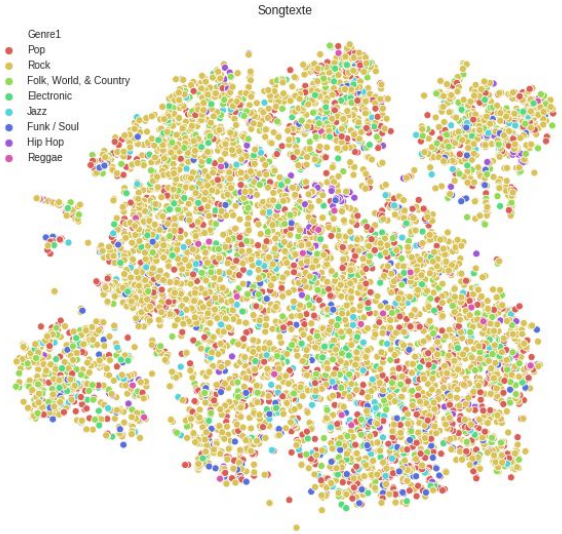


Lyrics

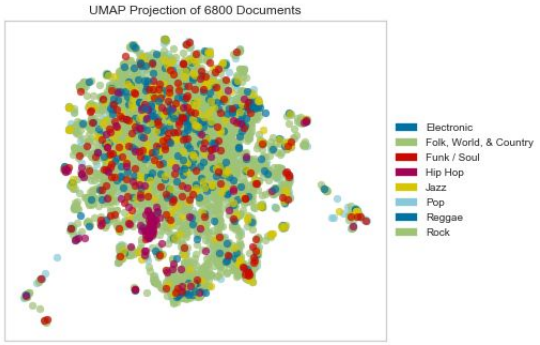
PCA



t-SNE

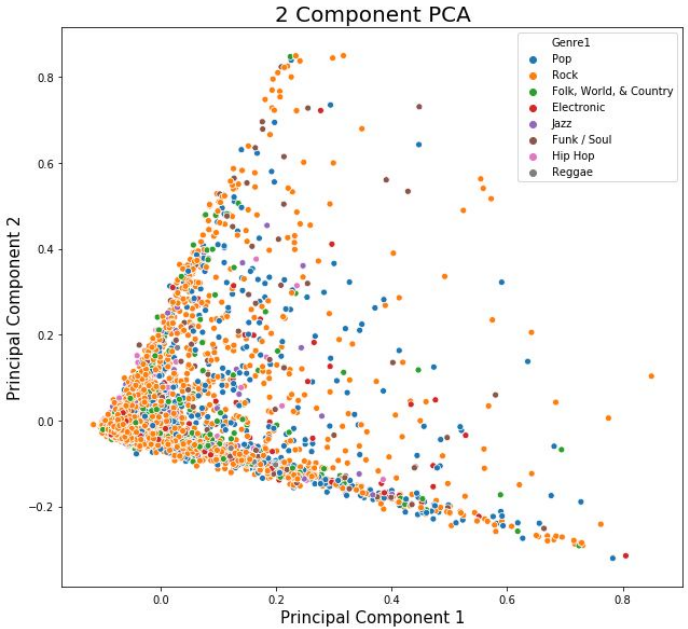


UMAP



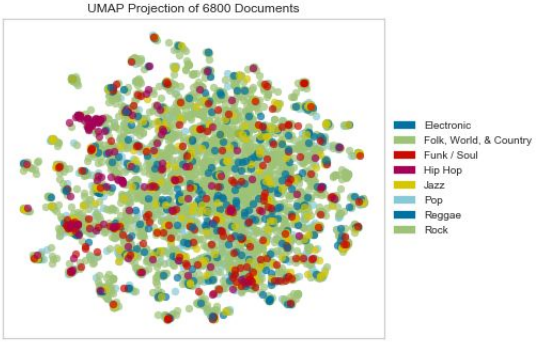
POS

PCA

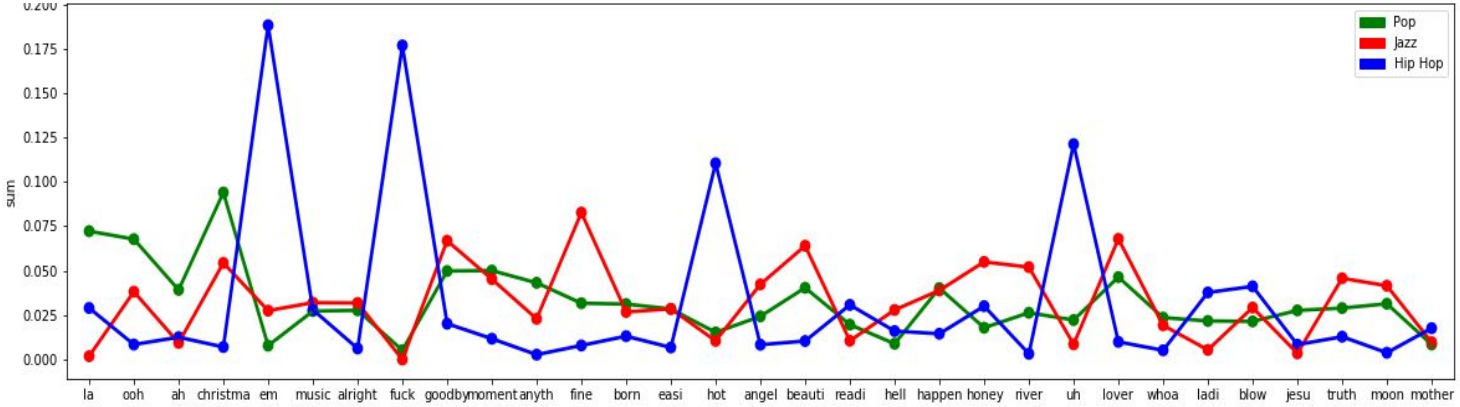


t-SNE

UMAP



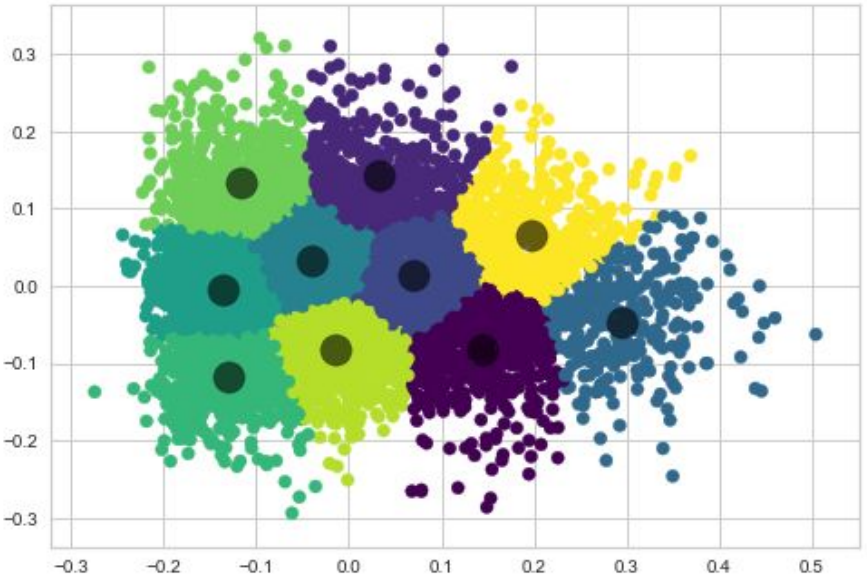
30 häufigste Wörter nach Genre



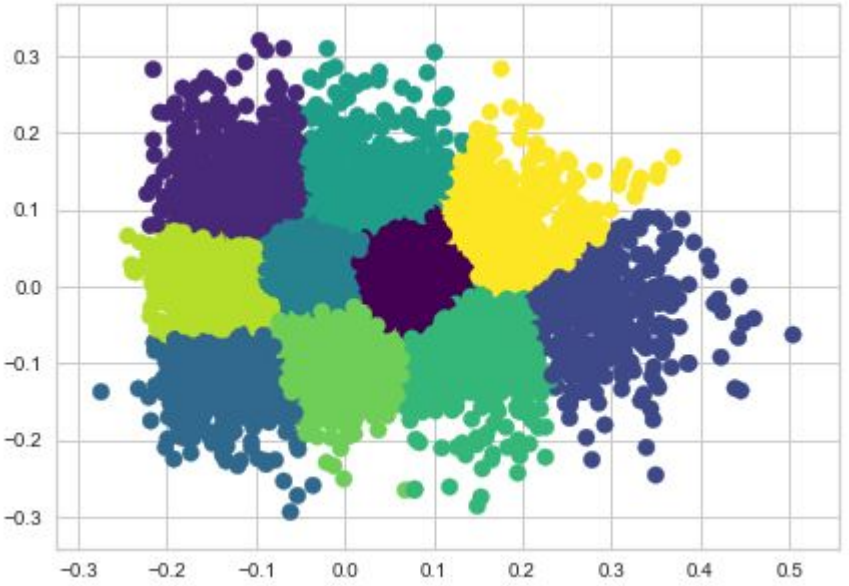
Wordclouds Country - Hip Hop

K-Means

PCA -- eigene Implementierung

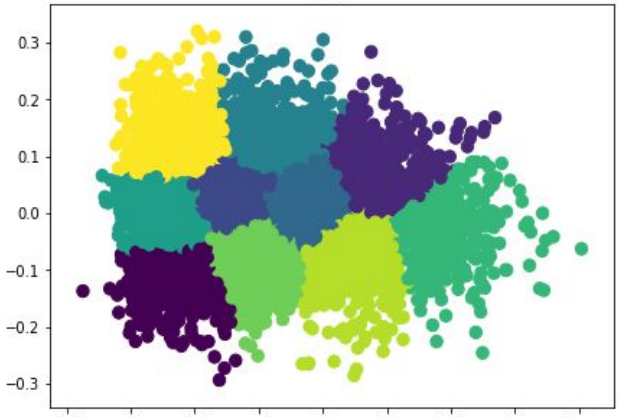


PCA -- SCIKIT-LEARN

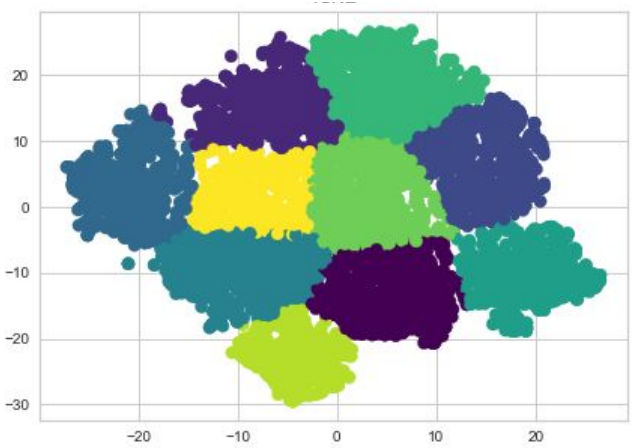


K-Means Lyrics

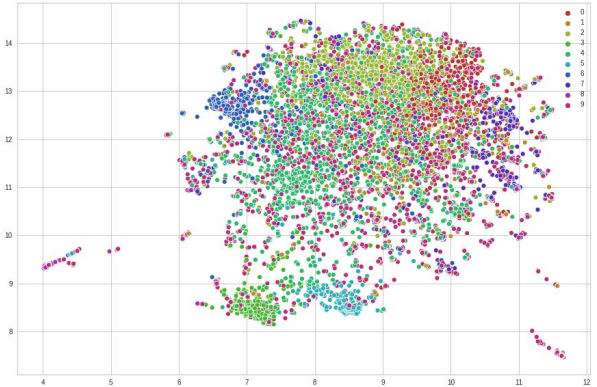
PCA



t-SNE

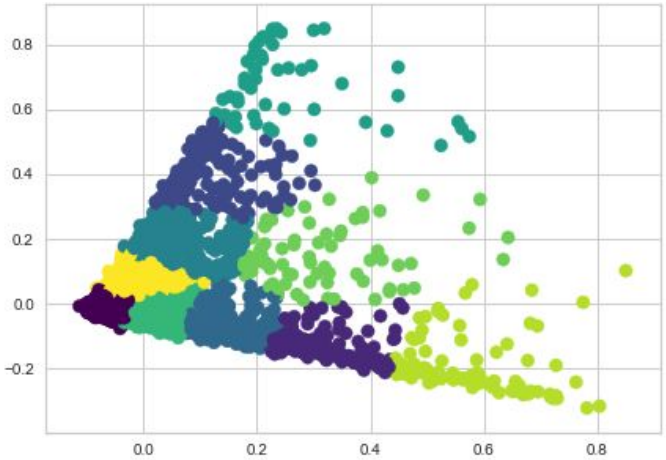


UMAP

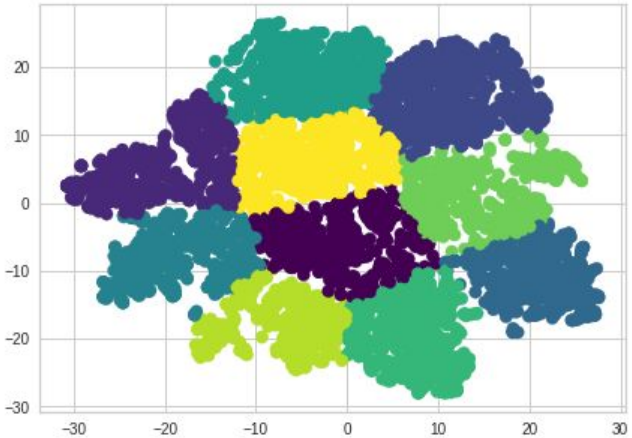


K-Means POS

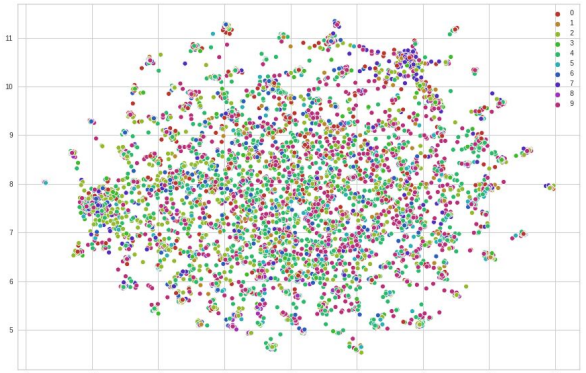
PCA



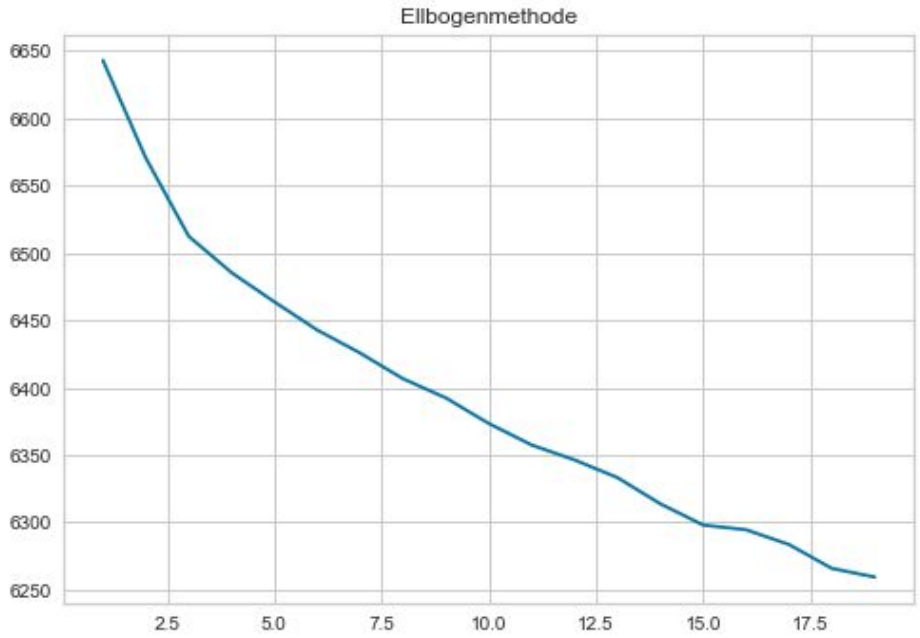
t-SNE

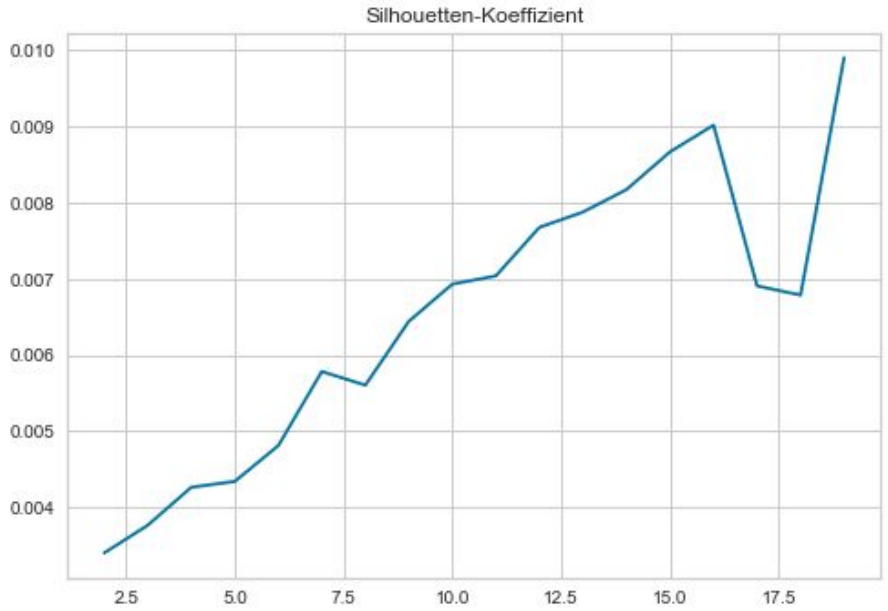


UMAP



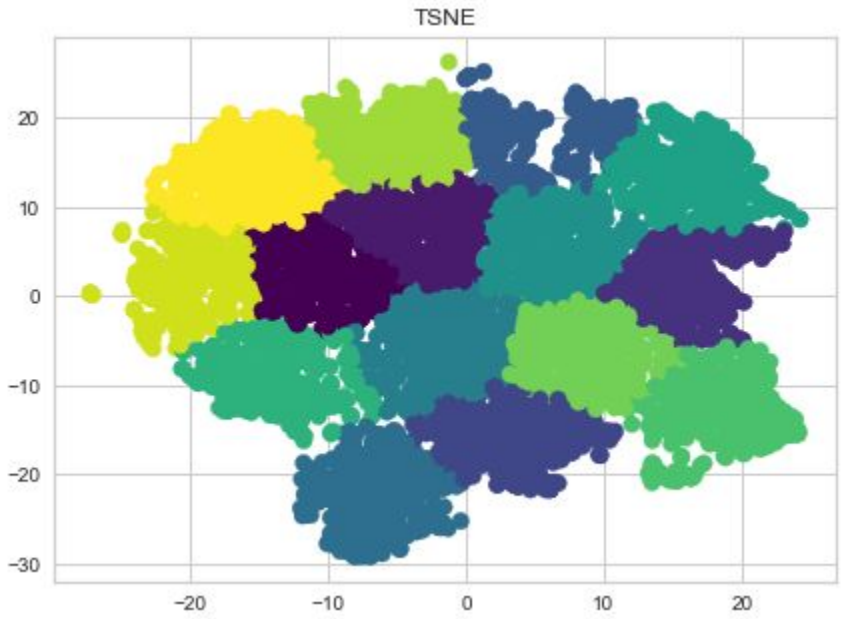
Anzahl der Cluster



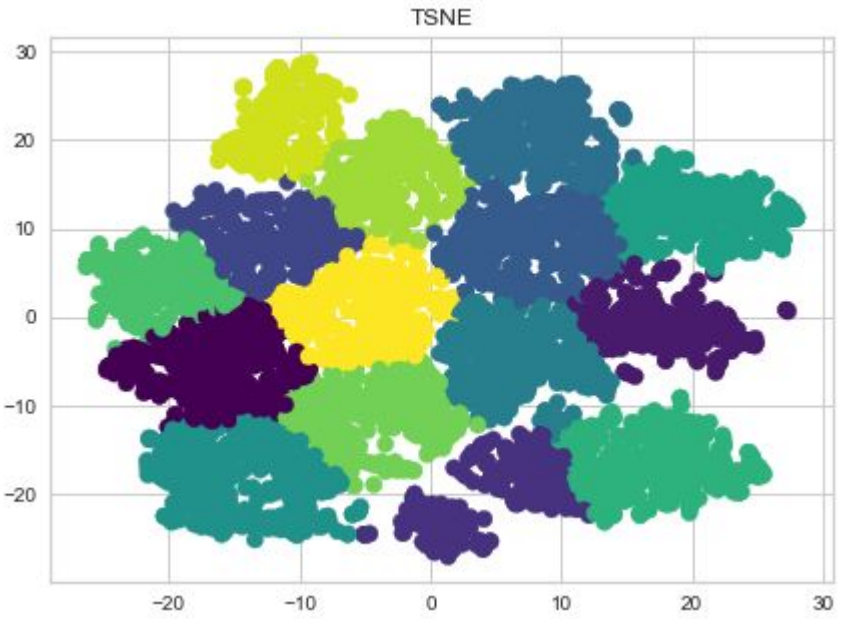


t-SNE mit k = 15

Lyrics

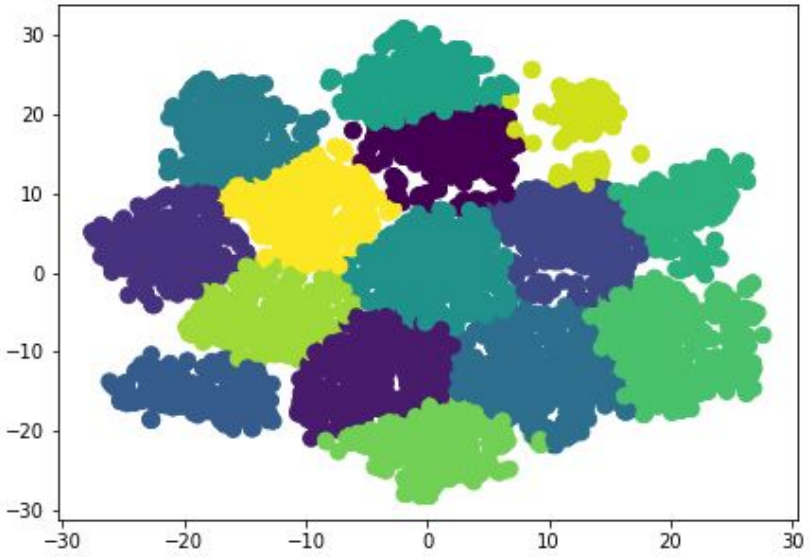


POS

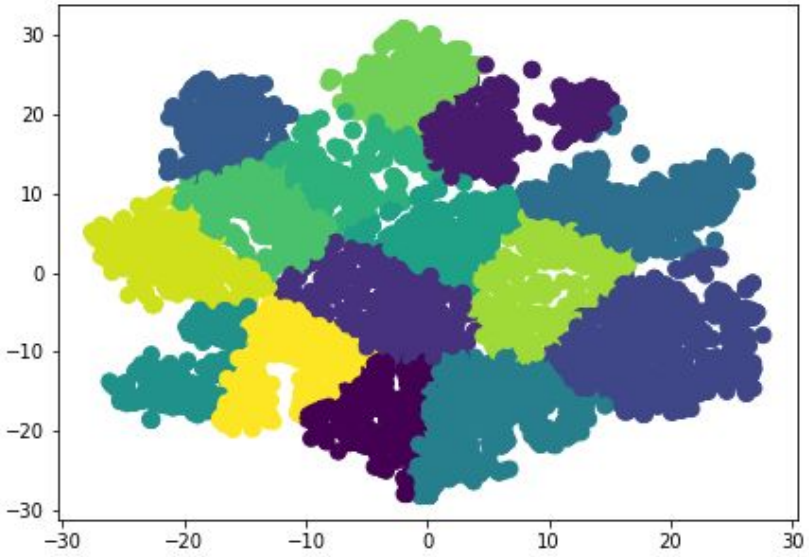


Mini-Batch mit t-SNE und POS

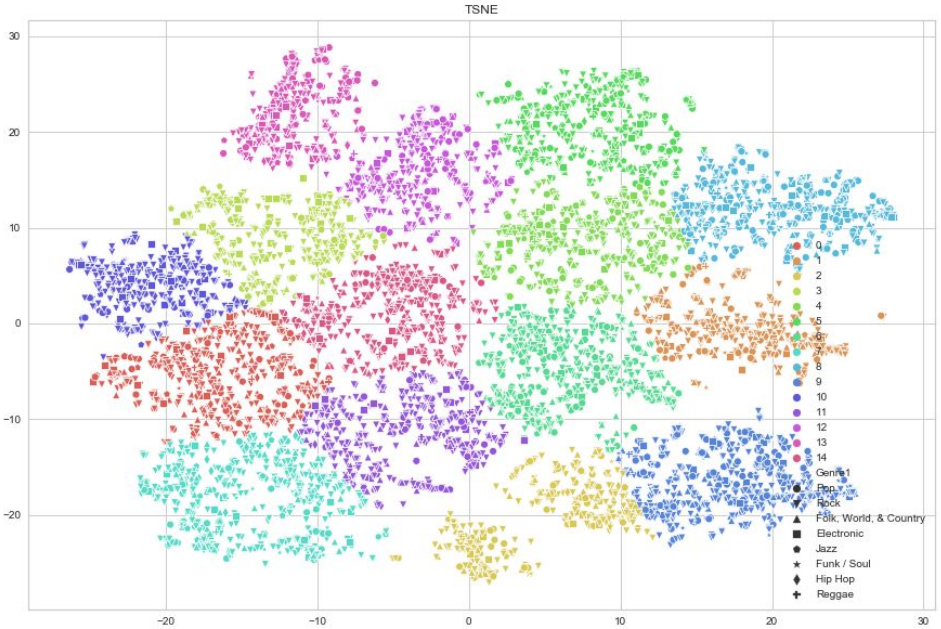
K-Means



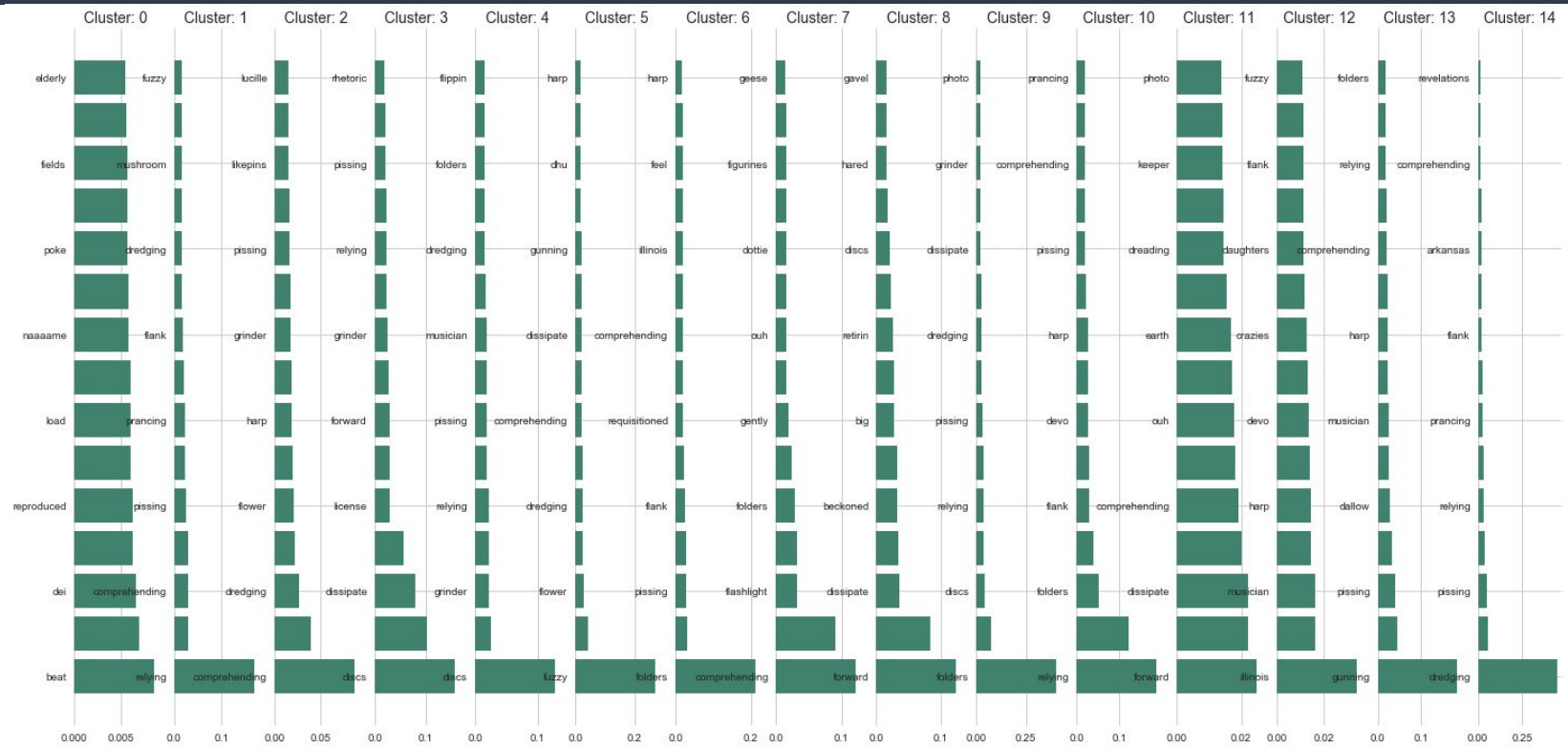
Mini-Batch



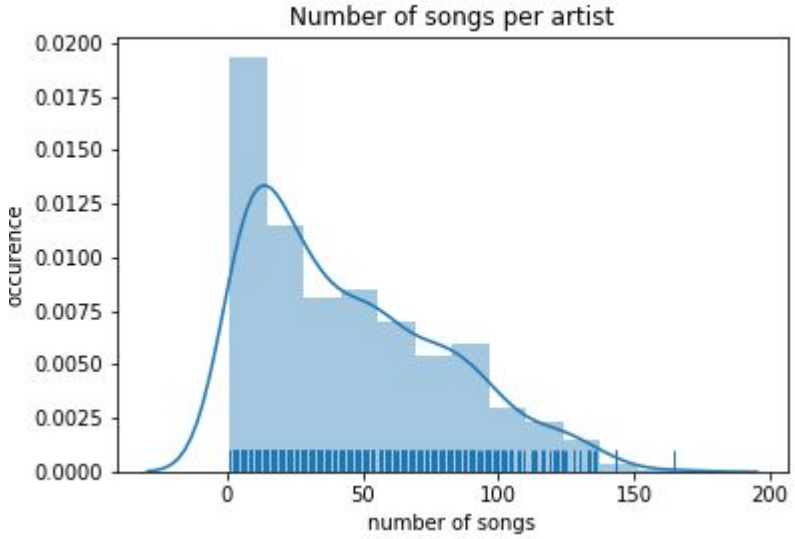
t-SNE mit Scikit-Learn und Genres



Features in Cluster

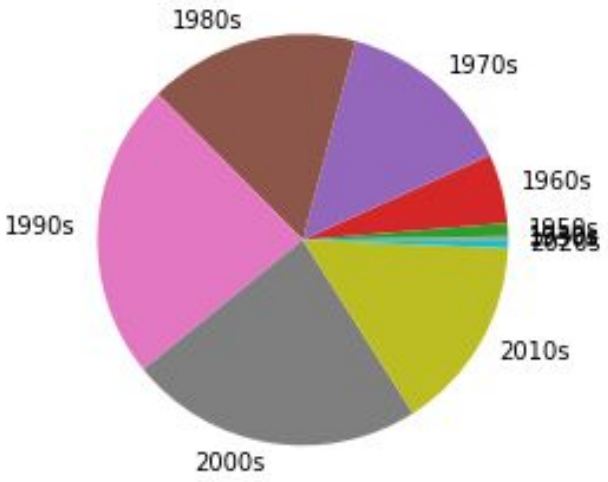


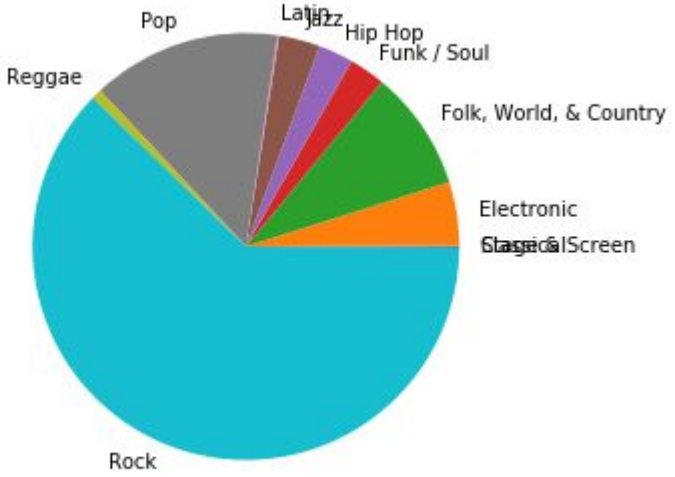
neuer Datensatz



- 541 Bands
- 24.443 Songs

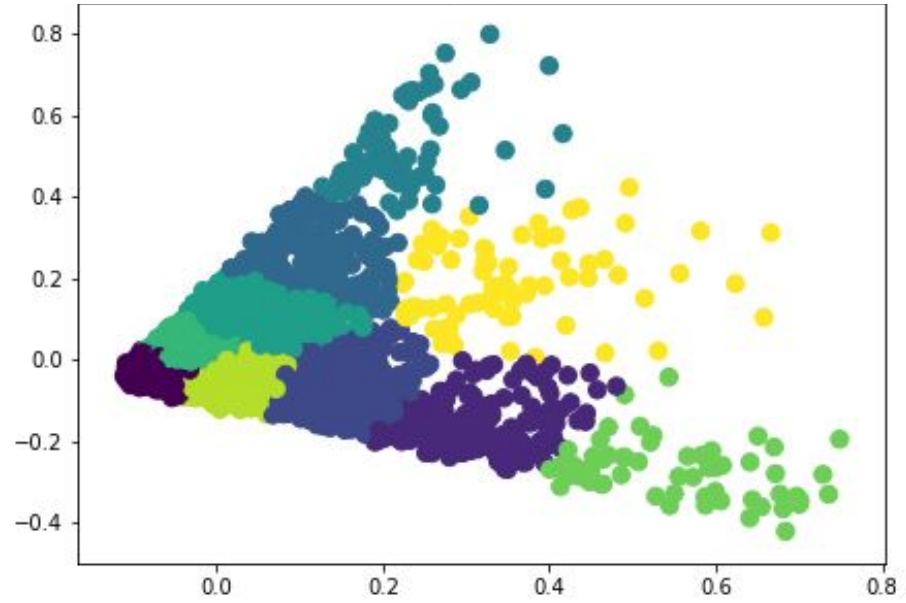
neuer Datensatz



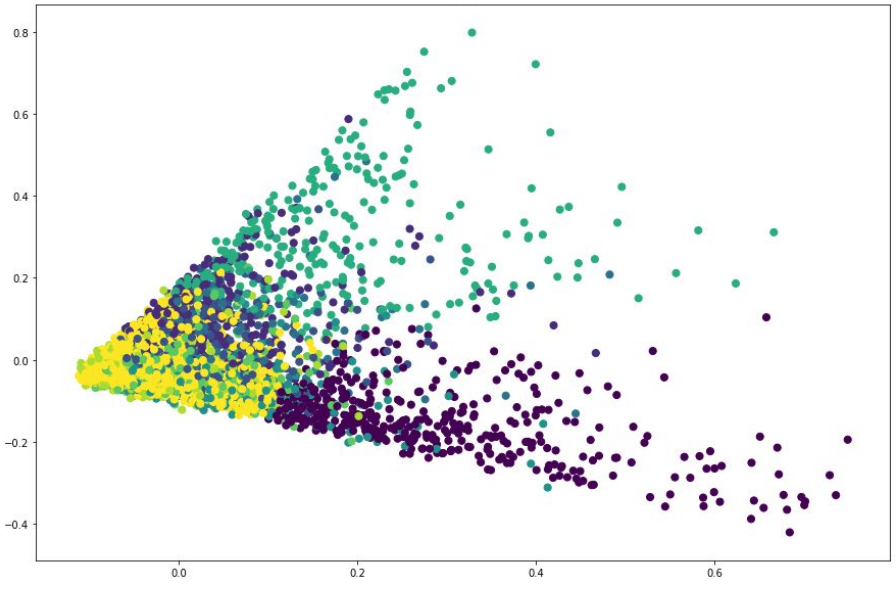


K-Means mit PCA

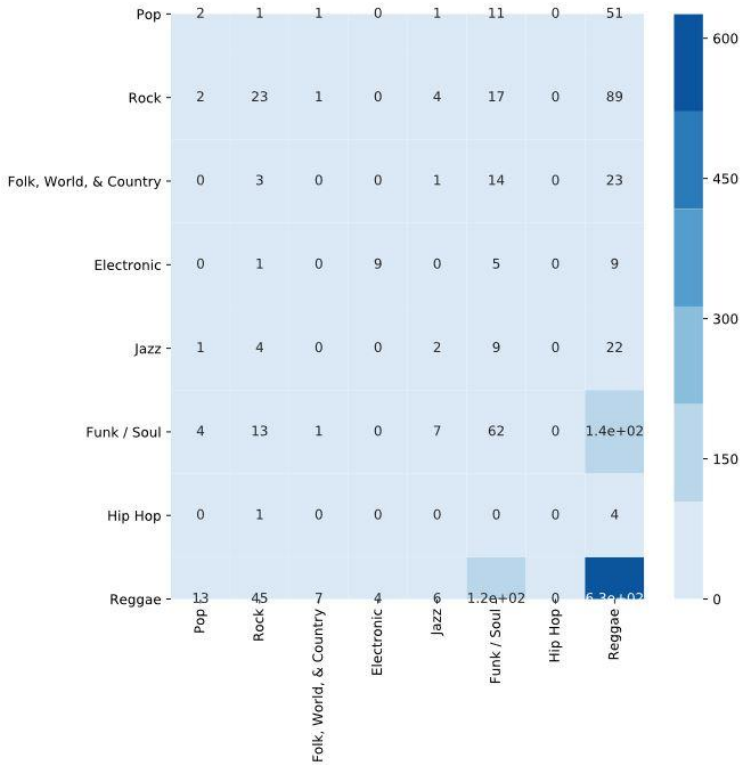
PCA vor KMeans



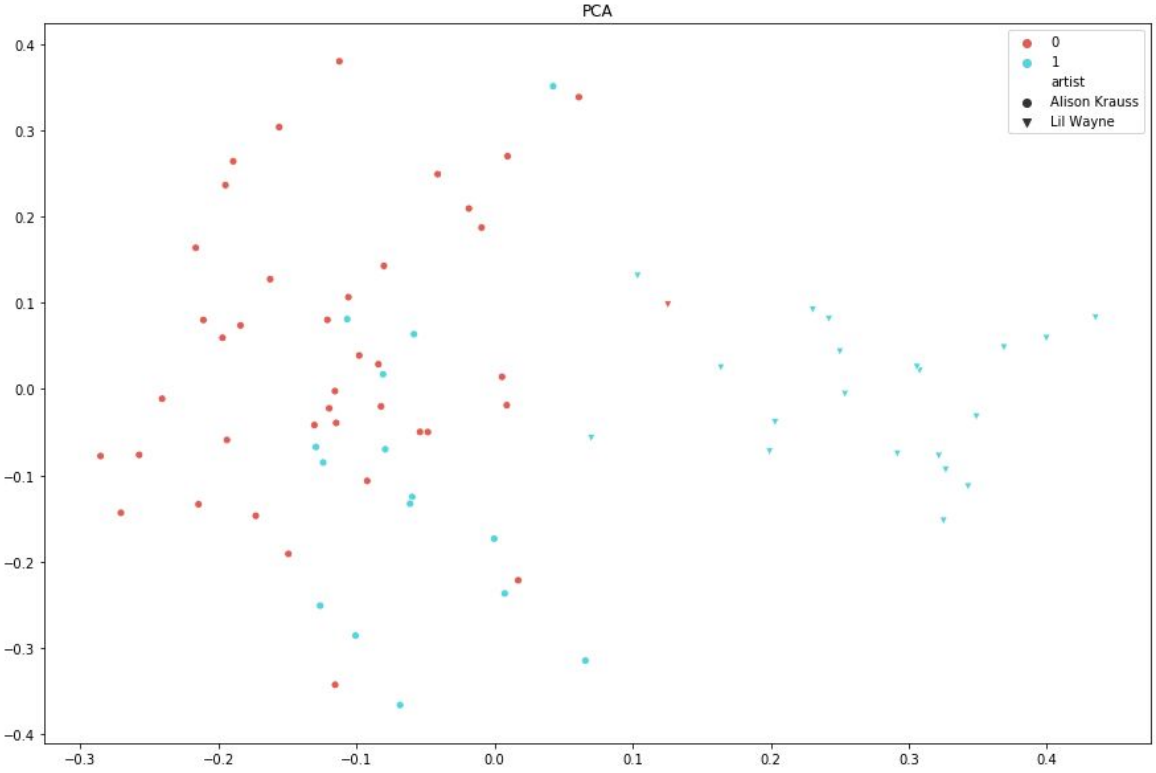
PCA nach KMeans



SVM

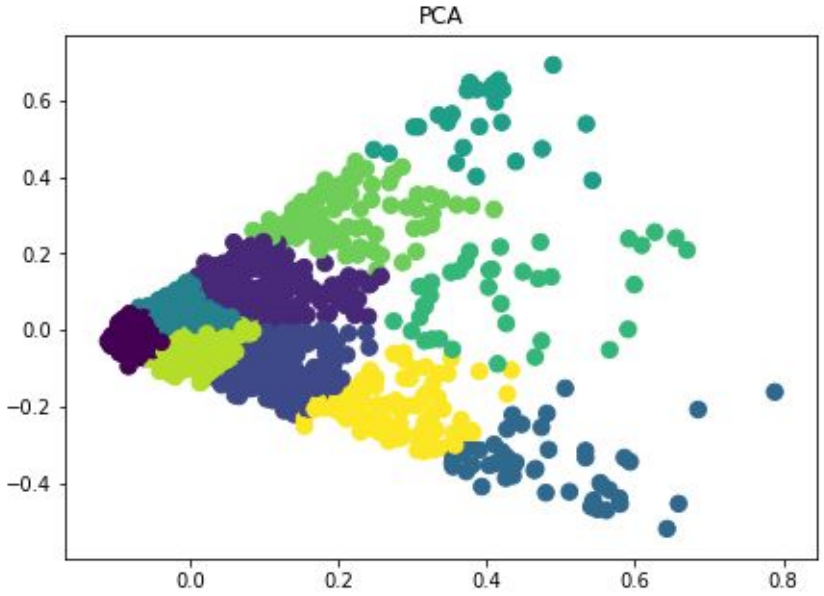

Pop
Rock
Folk, World, & Country
Electronic
Jazz
Funk / Soul
Hip Hop
Reggae
2 1 1 0 1 11 0 51
2 23 1 0 4 17 0 89
0 3 0 0 1 14 0 23
0 1 0 9 0 5 0 9
1 4 0 0 2 9 0 22
4 13 1 0 7 62 0 1.4e+02
0 1 0 0 0 0 0 4
13 45 7 4 6 1.2e+02 0 6.3e+02
600
450
300
150
0

Künstler

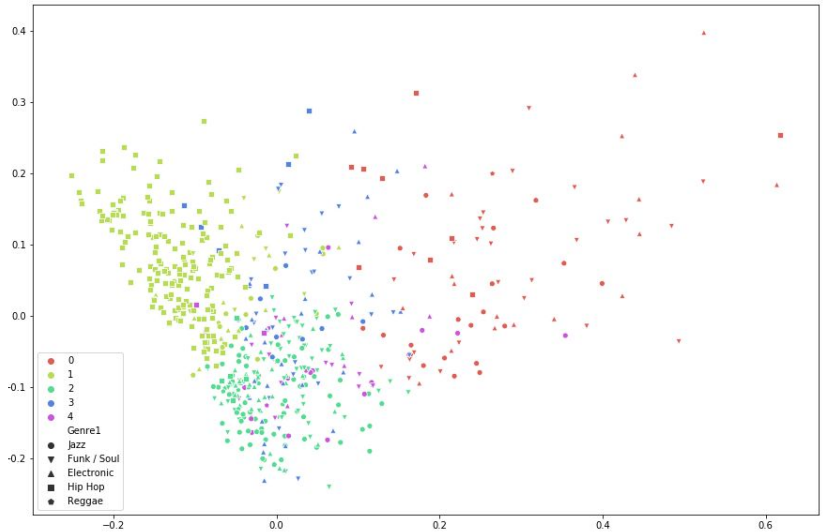

PCA
0
1
artist
Alison Krauss
Lil Wayne
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.4
−0.3
−0.2
−0.1
0.0
0.1
0.2
0.3
0.4

Genres

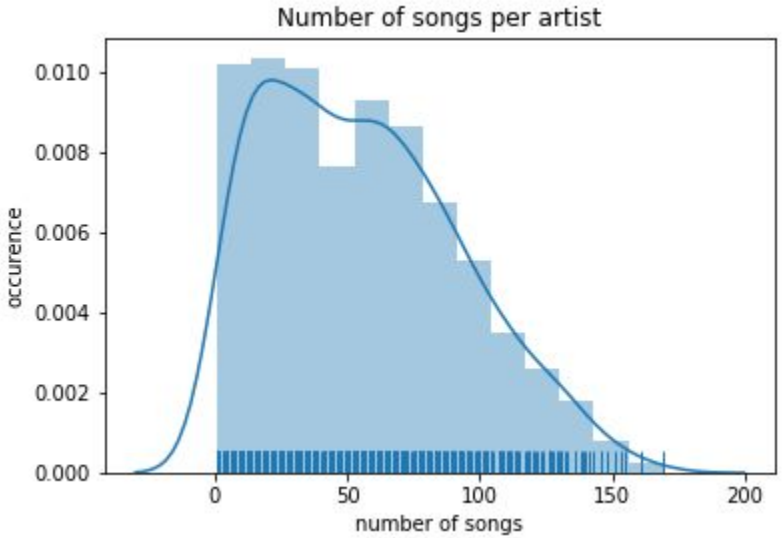
Rock



alle Genres außer Pop und Rock



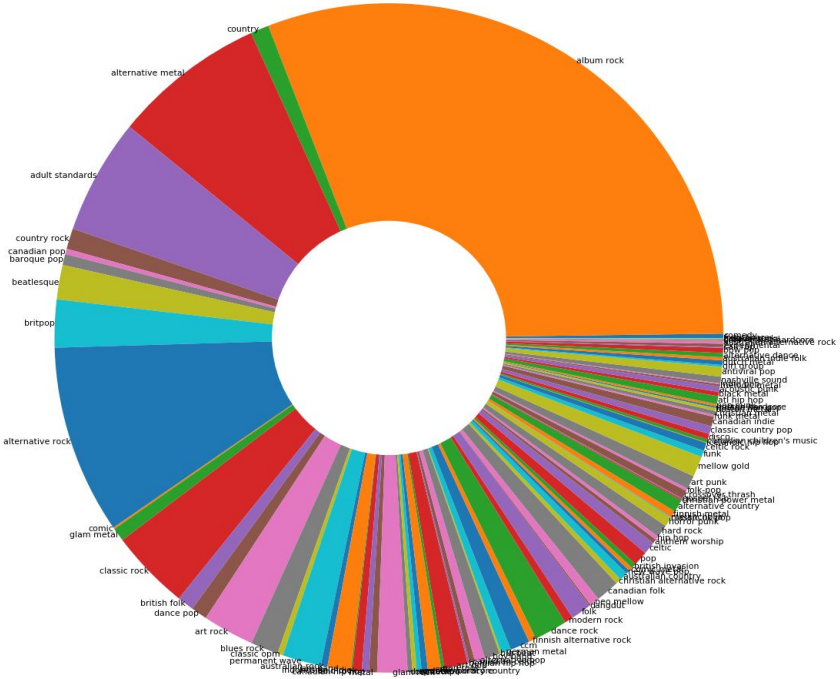
neuer Datensatz



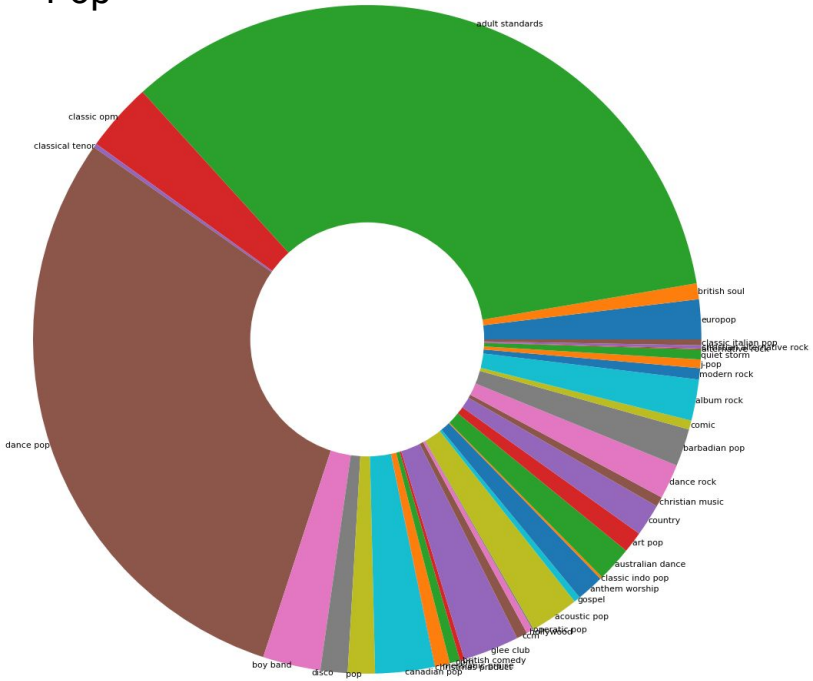
- 597 Bands
- 33.440 Songs

Subgenre

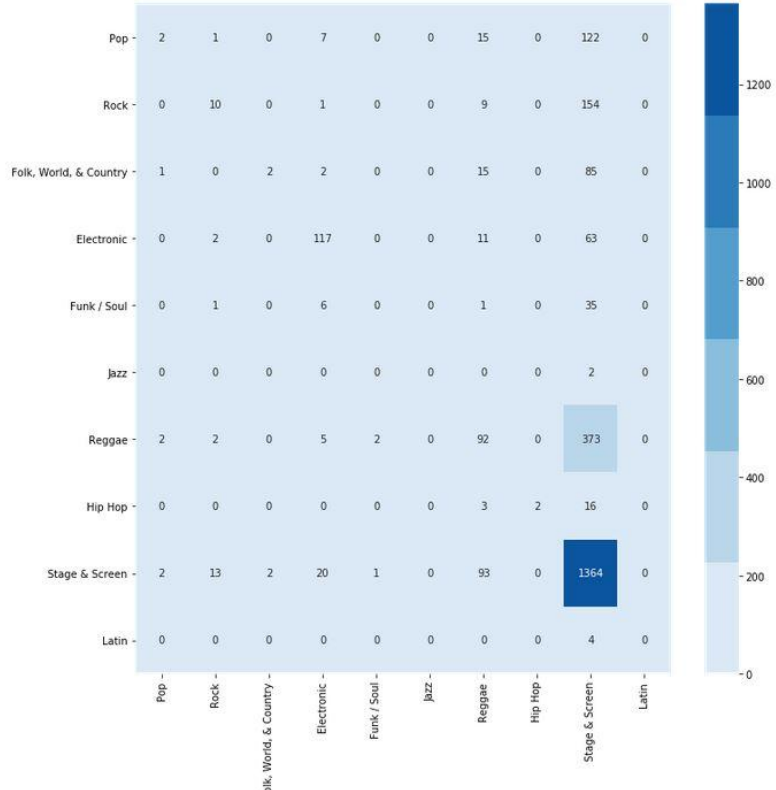
Rock

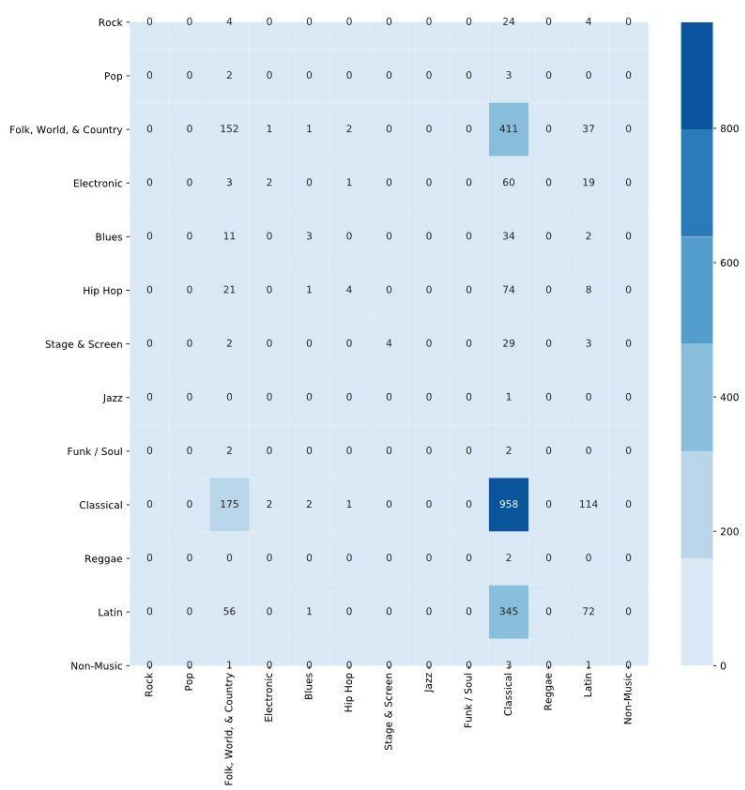


Pop

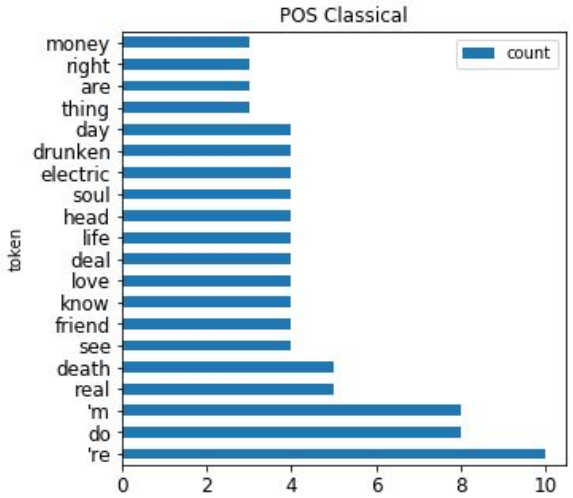


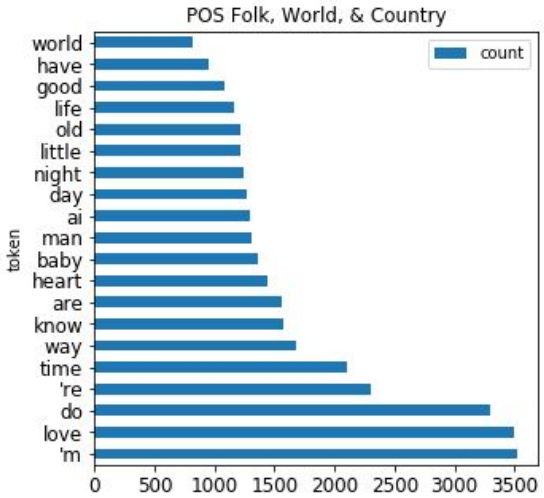
SVM

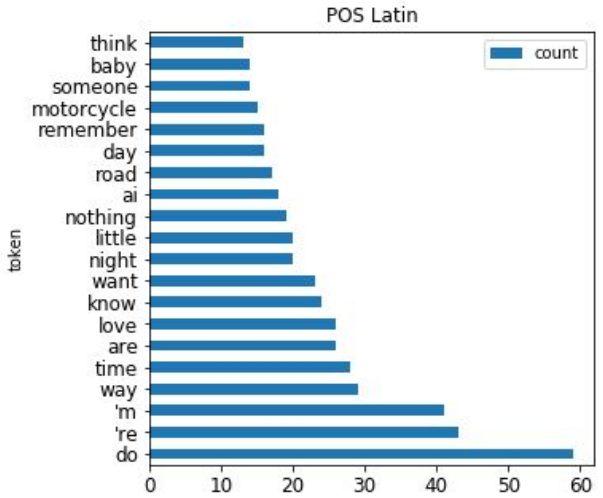

Pop 2 1 0 7 0 0 15 0 122 0
Rock 0 10 0 1 0 0 9 0 154 0
Folk, World, & Country 1 0 2 2 0 0 15 0 85 0
Electronic 0 2 0 117 0 0 11 0 63 0
Funk / Soul 0 1 0 6 0 0 1 0 35 0
Jazz 0 0 0 0 0 0 0 0 2 0
Reggae 2 2 0 5 2 0 92 0 373 0
Hip Hop 0 0 0 0 0 0 3 2 16 0
Stage & Screen 2 13 2 20 1 0 93 0 1364 0
Latin 0 0 0 0 0 0 0 0 4 0
Pop
Rock
Folk, World, & Country
Electronic
Funk / Soul
Jazz
Reggae
Hip Hop
Stage & Screen
Latin
1200
1000
800
600
400
200
0


Rock 0 0 4 0 0 0 0 0 0 24 0 4 0
Pop 0 0 2 0 0 0 0 0 0 3 0 0 0
Folk, World, & Country 0 0 152 1 1 2 0 0 0 411 0 37 0
Electronic 0 0 3 2 0 1 0 0 0 60 0 19 0
Blues 0 0 11 0 3 0 0 0 0 34 0 2 0
Hip Hop 0 0 21 0 1 4 0 0 0 74 0 8 0
Stage & Screen 0 0 2 0 0 0 4 0 0 29 0 3 0
Jazz 0 0 0 0 0 0 0 0 0 1 0 0 0
Funk / Soul 0 0 2 0 0 0 0 0 0 2 0 0 0
Classical 0 0 175 2 2 1 0 0 0 958 0 114 0
Reggae 0 0 0 0 0 0 0 0 0 2 0 0 0
Latin 0 0 56 0 1 0 0 0 0 345 0 72 0
Non-Music 0 0 1 0 0 0 0 0 0 4 0 1 0
Rock
Pop
Folk, World, & Country
Electronic
Blues
Hip Hop
Stage & Screen
Jazz
Funk / Soul
Classical
Reggae
Latin
Non-Music
800
600
400
200
0

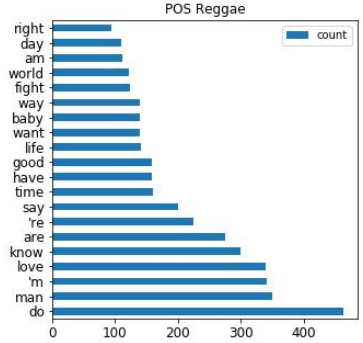
Nach SVM -- most common words

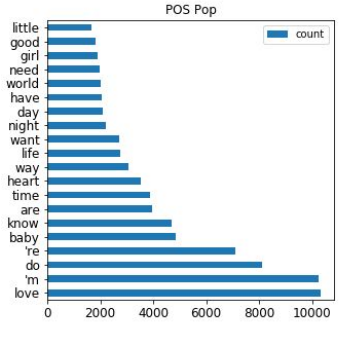


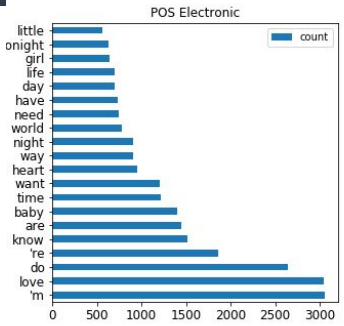


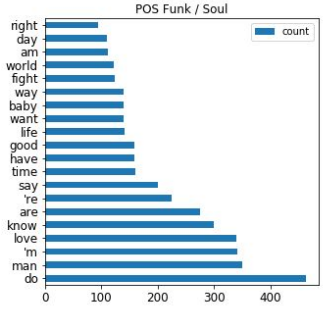


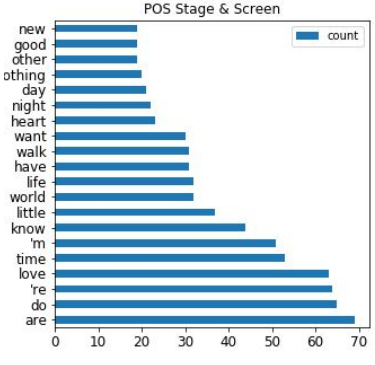
Genres – most common words

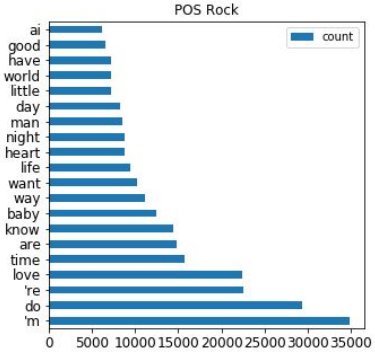


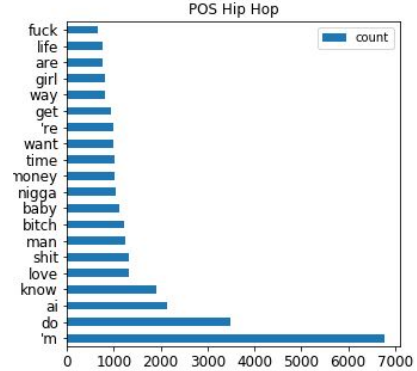




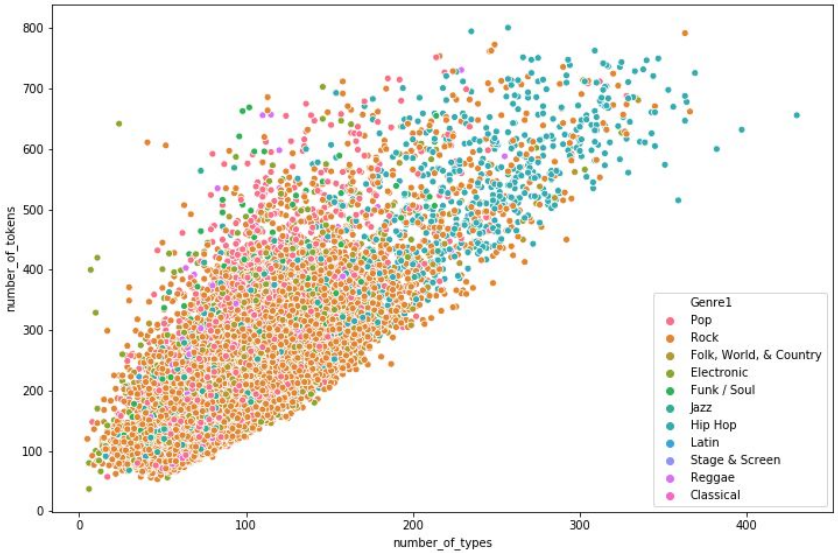


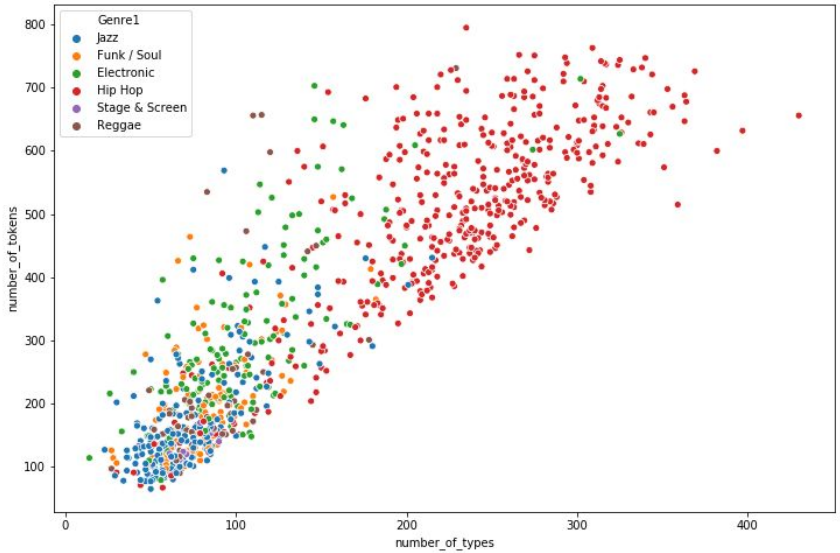






type-token ratio

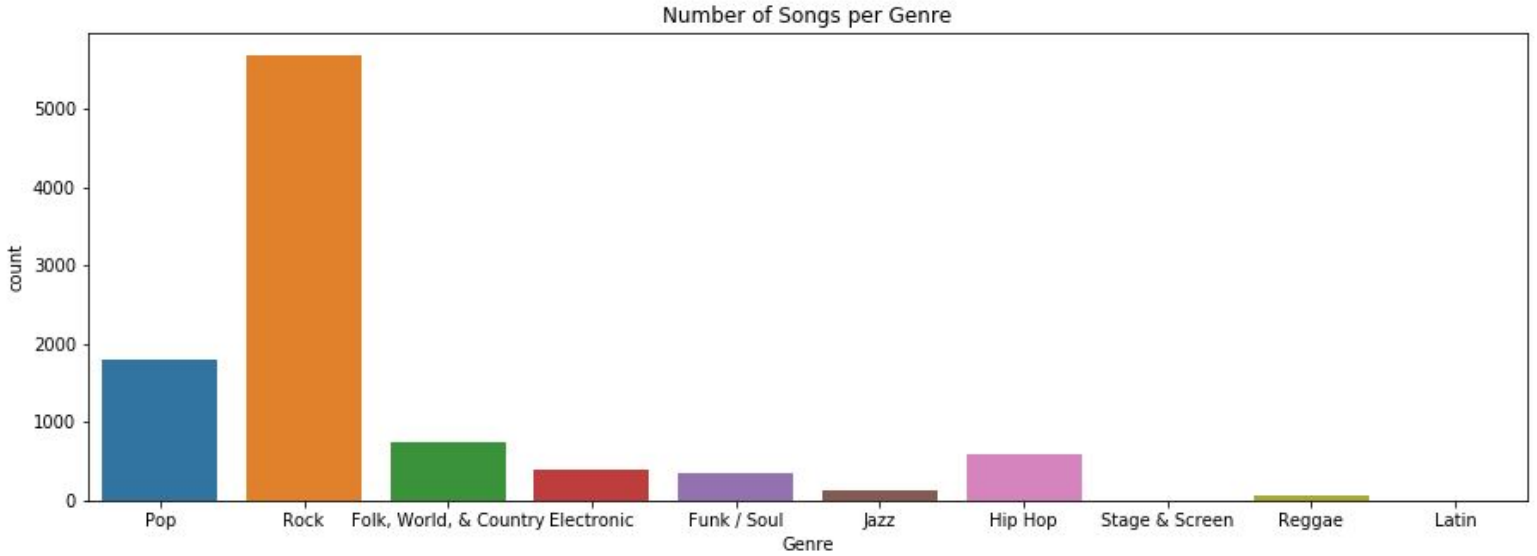




verkleinerter Datensatz

Datensatz: nur überdurchschnittlich lange Texte

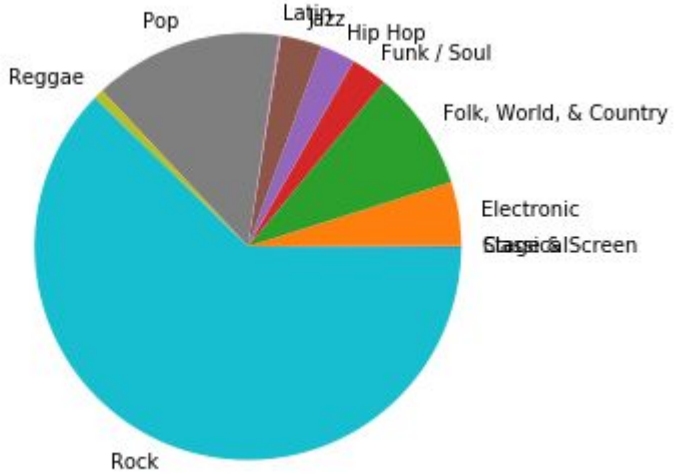
nur Texte über der Durchschnittslänge: hier ca. 222 Tokens



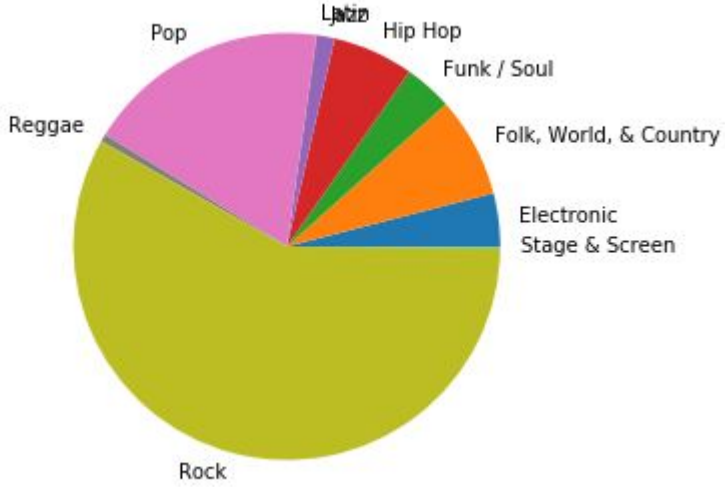
Verteilung der Genres

Datensatz: nur überdurchschnittlich lange Texte

vor Verkleinerung



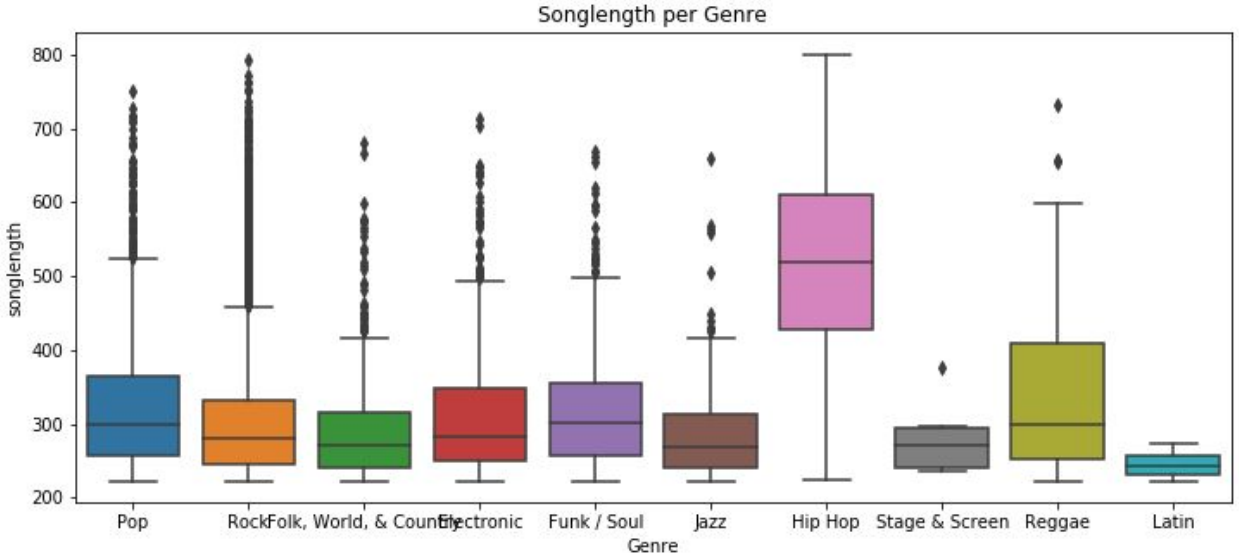
nach Verkleinerung



Textlängen pro Genre

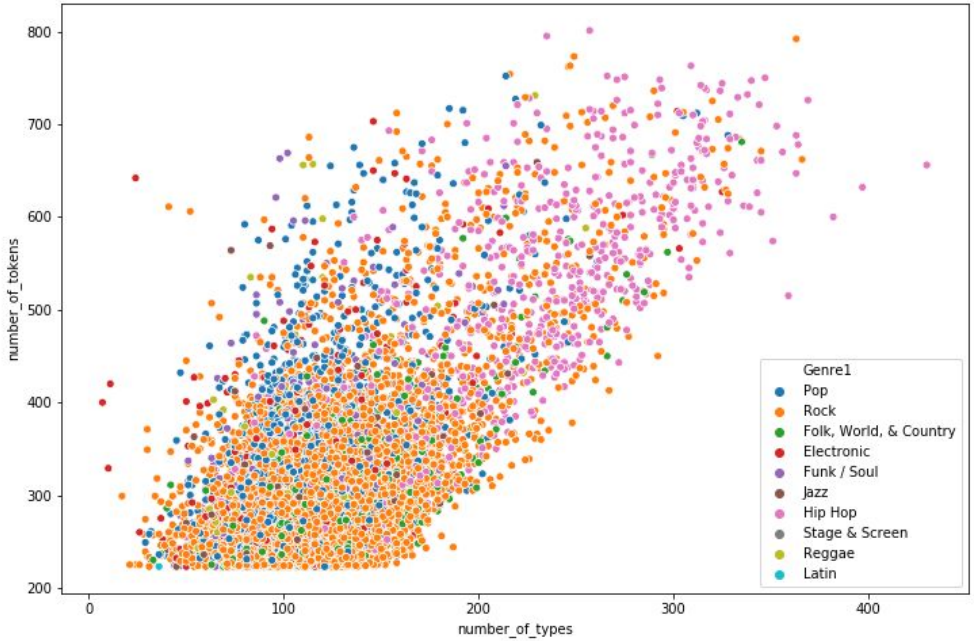
Datensatz: nur überdurchschnittlich lange Texte

Textlänge von HipHop fällt noch stärker auf als vor Verkleinerung des Datensatzes

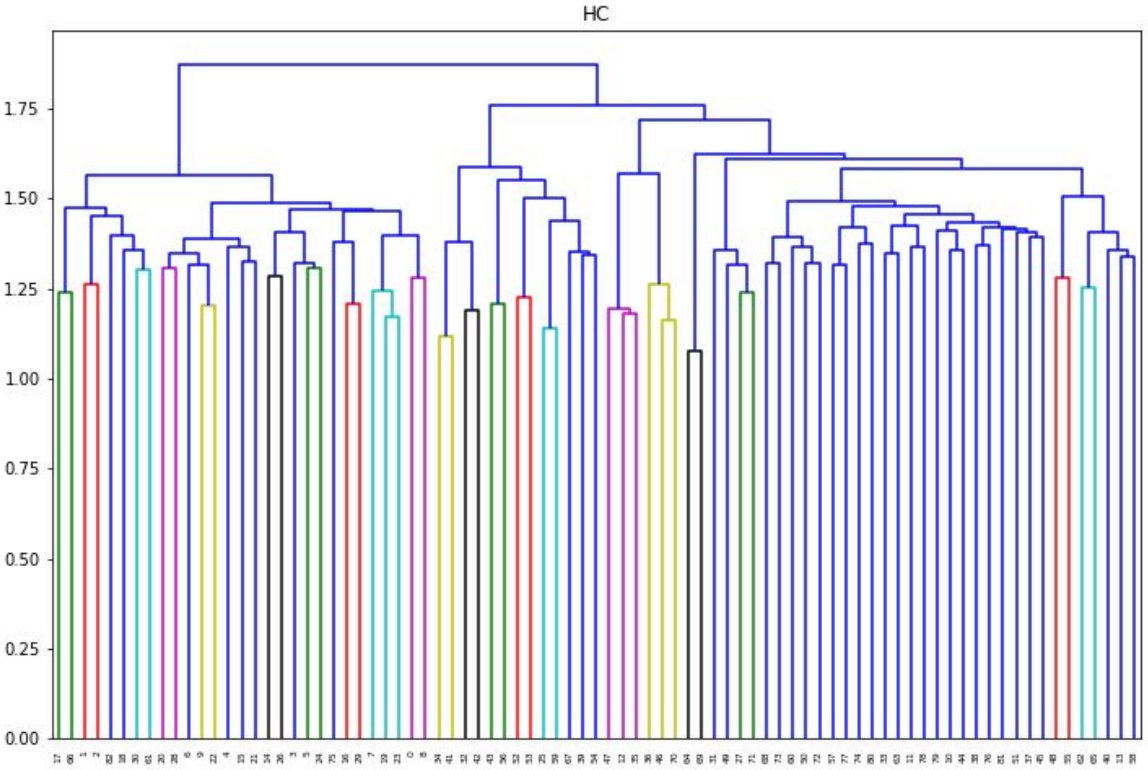


type-token-ratio

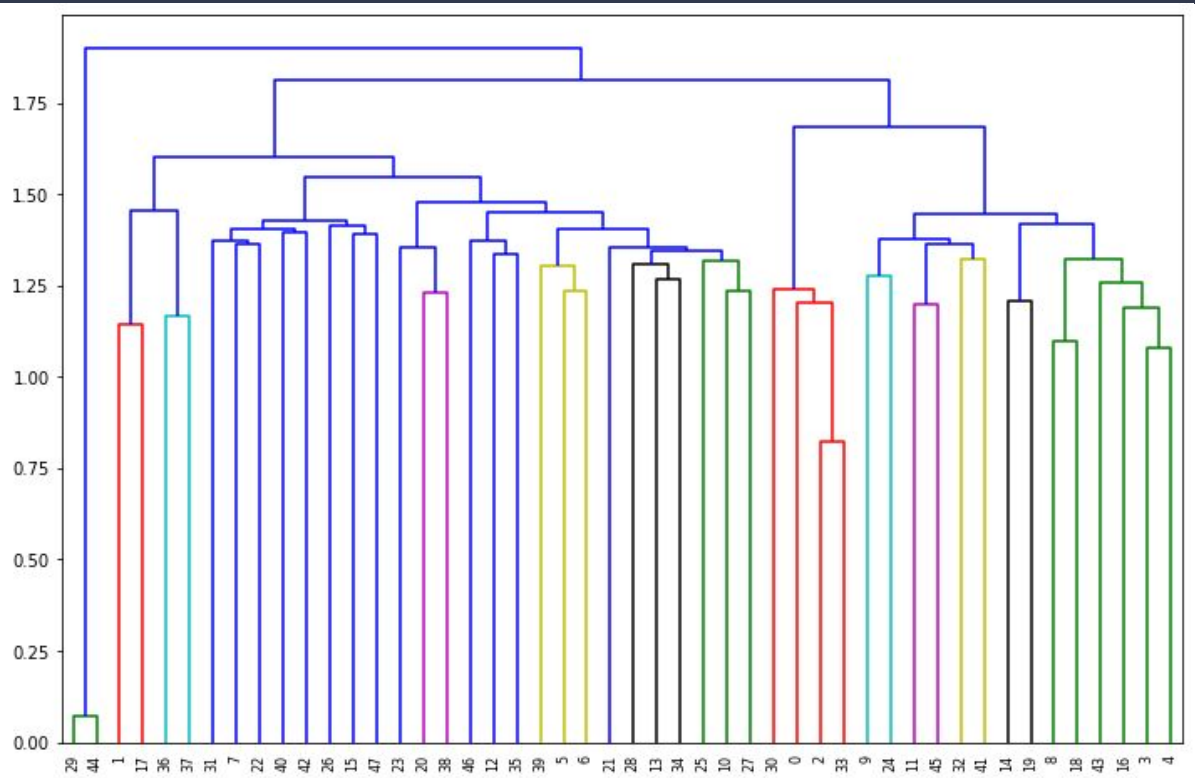
Datensatz: nur überdurchschnittlich lange Texte



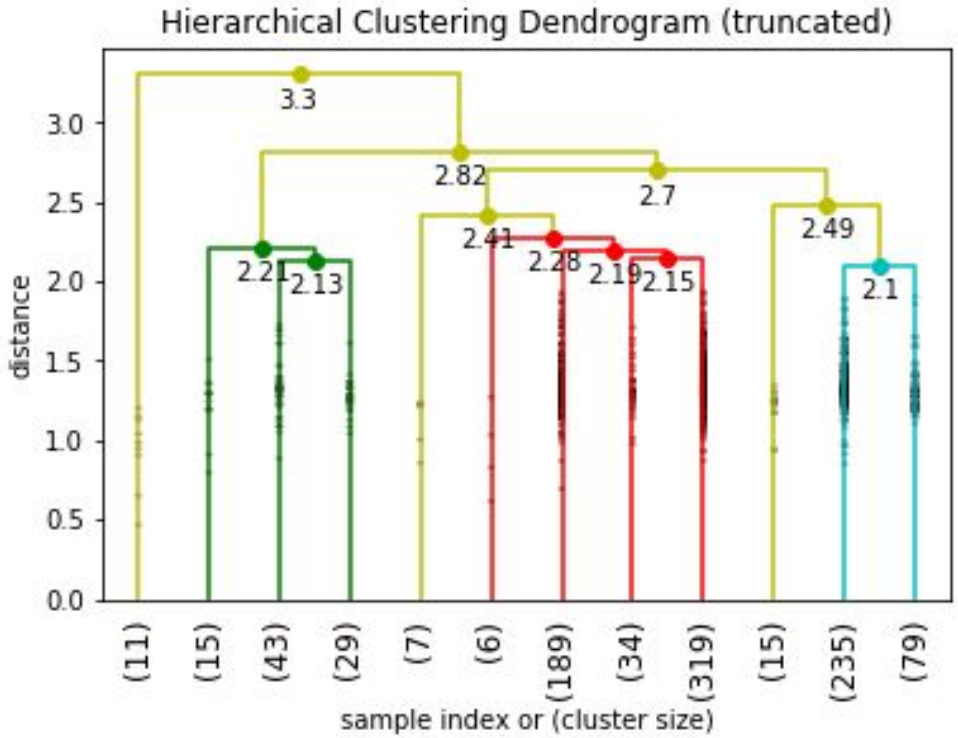
Hierarchical Clustering - Eminem & ABBA



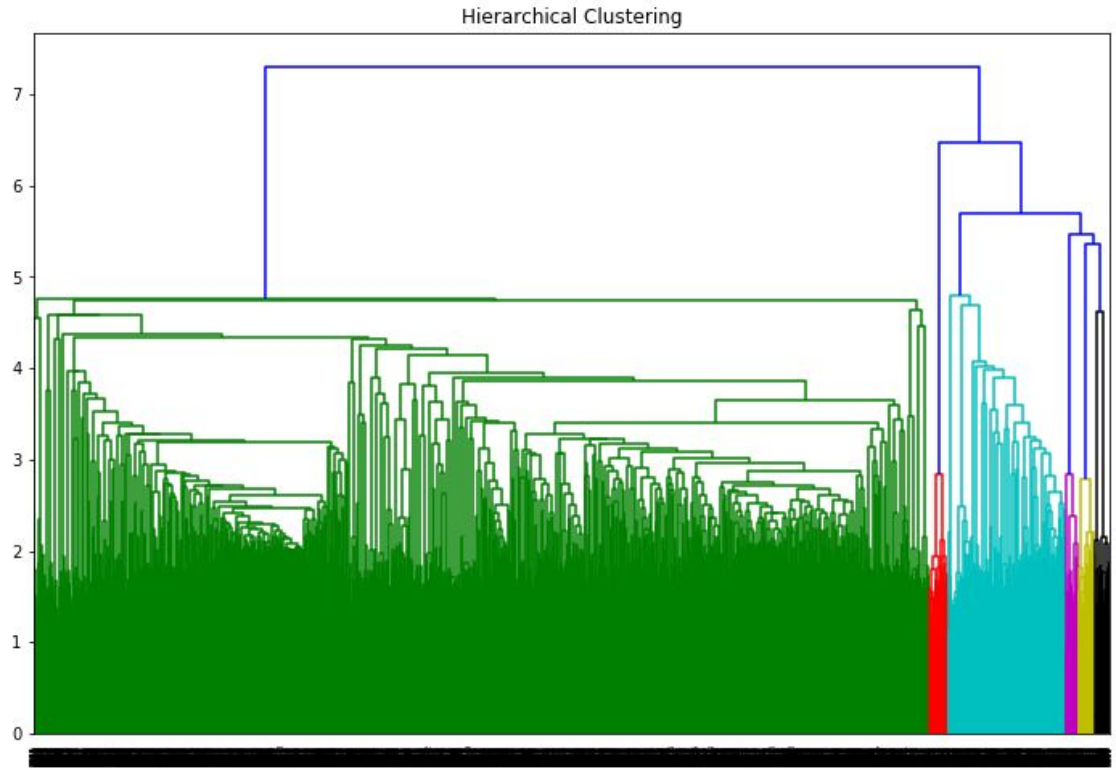
Hierarchical Clustering – 1980: Electronic & Reggae



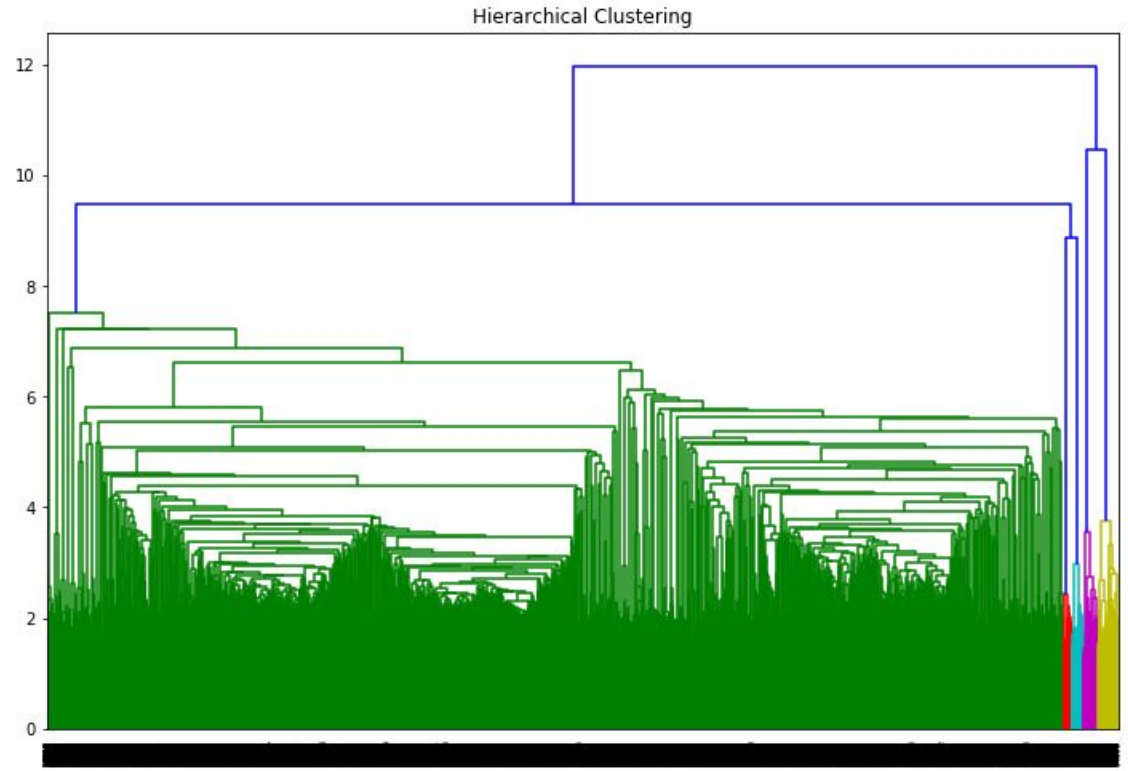
Hierarchical Clustering - Latin & HipHop



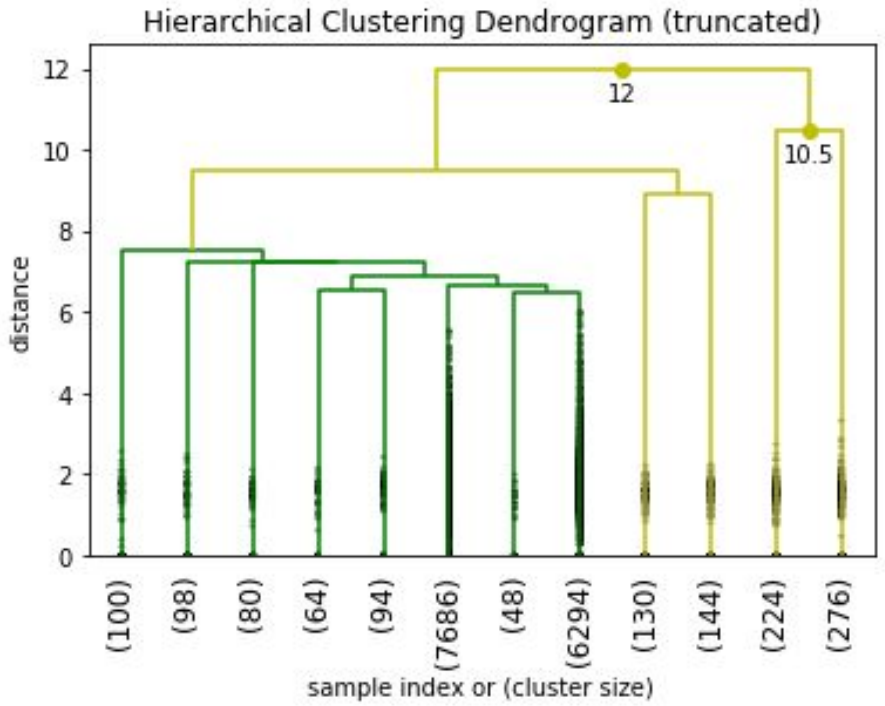
Hierarchical Clustering - Pop (POS)



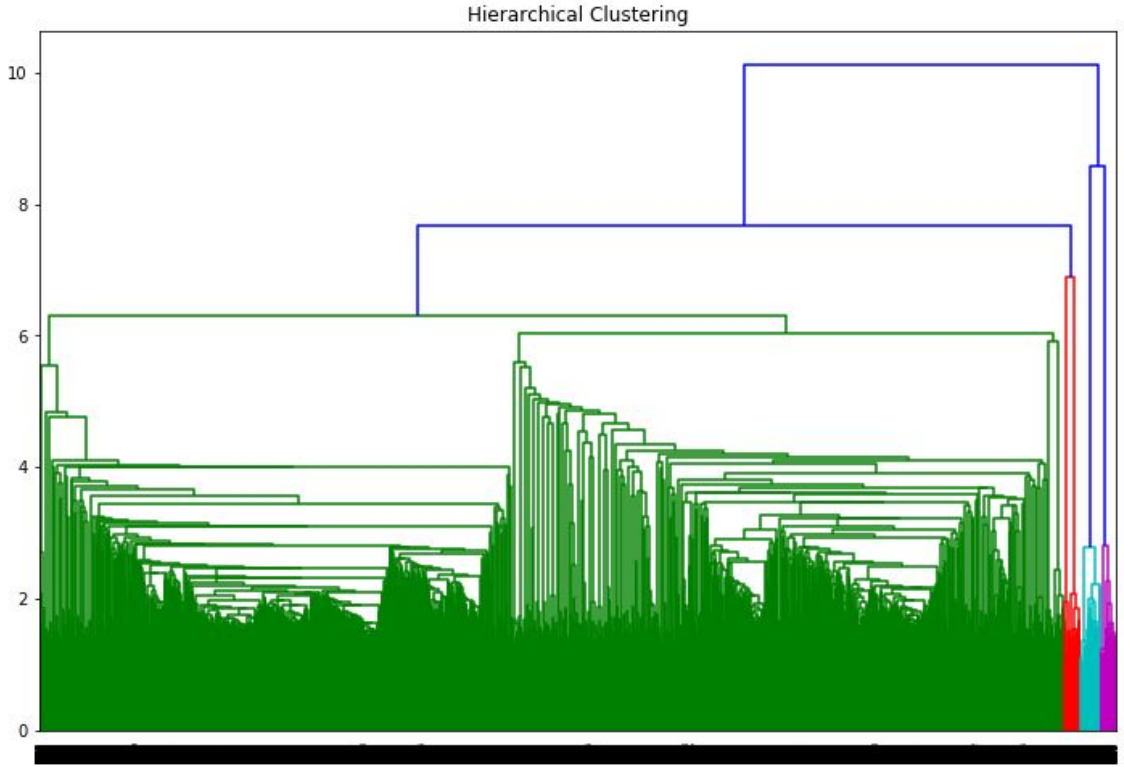
Hierarchical Clustering - Rock (POS)



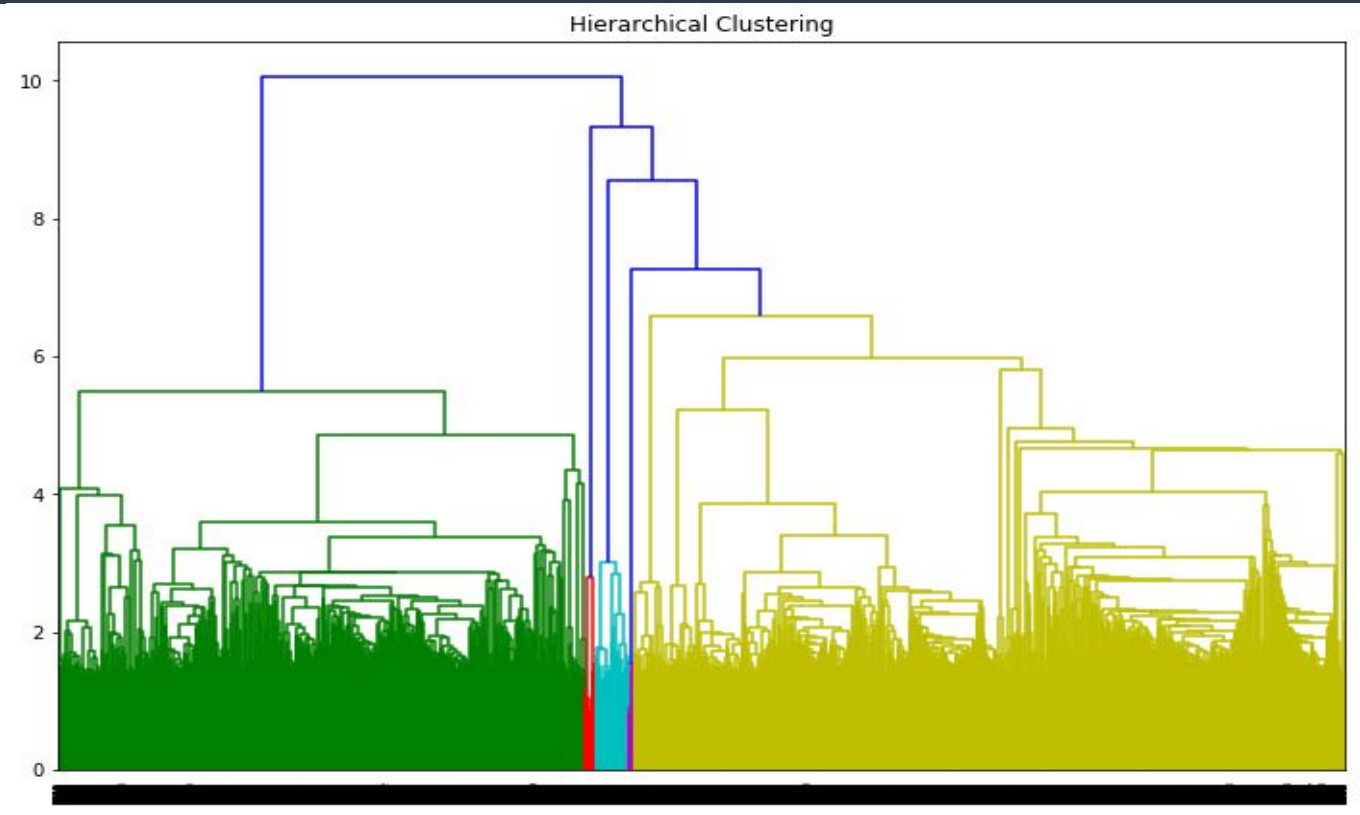
Hierarchical Clustering - Rock (POS)



Hierarchical Clustering – Rock & Pop (POS)



Hierarchical Clustering - Rock & Pop (Text)



Code – HC

In [139]:
```
Z = linkage(tf_idf_array, 'ward')
Z[:20]
```

Out[139]:
```
array([[4.28000000e+02, 5.79000000e+02, 4.66649503e-01, 2.00000000e+00],
       [7.78000000e+02, 7.79000000e+02, 6.15168343e-01, 2.00000000e+00],
       [2.39000000e+02, 9.82000000e+02, 6.50724643e-01, 3.00000000e+00],
       [4.09000000e+02, 5.80000000e+02, 6.94752655e-01, 2.00000000e+00],
       [7.14000000e+02, 7.19000000e+02, 7.95616619e-01, 2.00000000e+00],
       [7.75000000e+02, 8.66000000e+02, 8.29488222e-01, 2.00000000e+00],
       [3.24000000e+02, 6.71000000e+02, 8.50526308e-01, 2.00000000e+00],
       [6.82000000e+02, 6.87000000e+02, 8.56495952e-01, 2.00000000e+00],
       [2.90000000e+02, 5.78000000e+02, 8.69025799e-01, 2.00000000e+00],
       [3.41000000e+02, 4.02000000e+02, 8.70202274e-01, 2.00000000e+00],
       [1.59000000e+02, 9.77000000e+02, 8.85721802e-01, 2.00000000e+00],
       [6.91000000e+02, 9.84000000e+02, 9.03952944e-01, 4.00000000e+00],
       [7.30000000e+02, 9.86000000e+02, 9.09266209e-01, 3.00000000e+00],
       [8.91000000e+02, 9.61000000e+02, 9.15191540e-01, 2.00000000e+00],
       [1.10000000e+02, 7.99000000e+02, 9.25897696e-01, 2.00000000e+00],
       [8.18000000e+02, 8.54000000e+02, 9.31039325e-01, 2.00000000e+00],
       [1.26000000e+02, 9.57000000e+02, 9.38272178e-01, 2.00000000e+00],
       [7.69000000e+02, 9.93000000e+02, 9.45615989e-01, 5.00000000e+00],
       [2.40000000e+01, 9.23000000e+02, 9.45918300e-01, 2.00000000e+00],
       [1.39000000e+02, 1.66000000e+02, 9.60723958e-01, 2.00000000e+00]])
```

In [141]:
```
model = AgglomerativeClustering(n_clusters=None, distance_threshold=0).fit(tf_idf_array)
link_matrix = linkage_matrix(tf_idf_array.shape[0], model.children_, model.distances_)
link_matrix[:20]
```

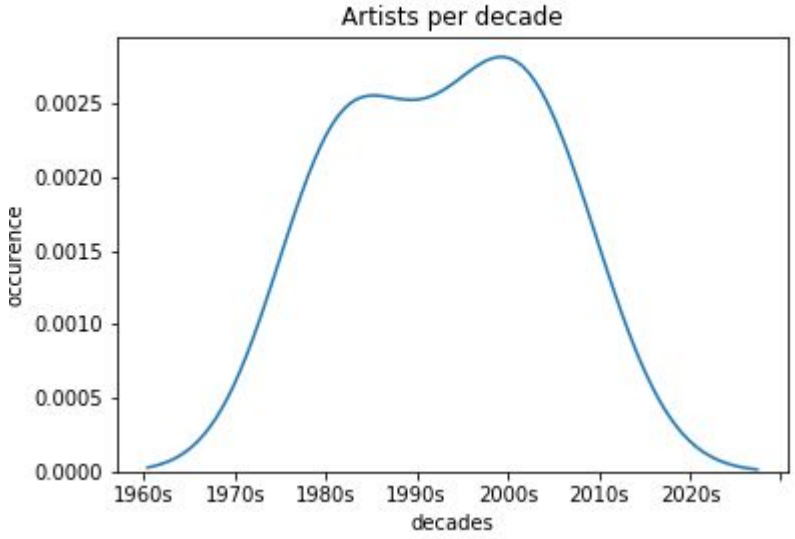
Out[141]:
```
array([[4.28000000e+02, 5.79000000e+02, 4.66649503e-01, 2.00000000e+00],
       [7.78000000e+02, 7.79000000e+02, 6.15168343e-01, 2.00000000e+00],
       [2.39000000e+02, 9.82000000e+02, 6.50724643e-01, 3.00000000e+00],
       [4.09000000e+02, 5.80000000e+02, 6.94752655e-01, 2.00000000e+00],
       [7.14000000e+02, 7.19000000e+02, 7.95616619e-01, 2.00000000e+00],
       [7.75000000e+02, 8.66000000e+02, 8.29488222e-01, 2.00000000e+00],
       [3.24000000e+02, 6.71000000e+02, 8.50526308e-01, 2.00000000e+00],
       [6.82000000e+02, 6.87000000e+02, 8.56495952e-01, 2.00000000e+00],
       [2.90000000e+02, 5.78000000e+02, 8.69025799e-01, 2.00000000e+00],
       [3.41000000e+02, 4.02000000e+02, 8.70202274e-01, 2.00000000e+00],
       [1.59000000e+02, 9.77000000e+02, 8.85721802e-01, 2.00000000e+00],
       [6.91000000e+02, 9.84000000e+02, 9.03952944e-01, 4.00000000e+00],
       [7.30000000e+02, 9.86000000e+02, 9.09266209e-01, 3.00000000e+00],
       [8.91000000e+02, 9.61000000e+02, 9.15191540e-01, 2.00000000e+00],
       [1.10000000e+02, 7.99000000e+02, 9.25897696e-01, 2.00000000e+00],
       [8.18000000e+02, 8.54000000e+02, 9.31039325e-01, 2.00000000e+00],
       [1.26000000e+02, 9.57000000e+02, 9.38272178e-01, 2.00000000e+00],
       [7.69000000e+02, 9.93000000e+02, 9.45615989e-01, 5.00000000e+00],
       [2.40000000e+01, 9.23000000e+02, 9.45918300e-01, 2.00000000e+00],
       [1.39000000e+02, 1.66000000e+02, 9.60723958e-01, 2.00000000e+00]])
```

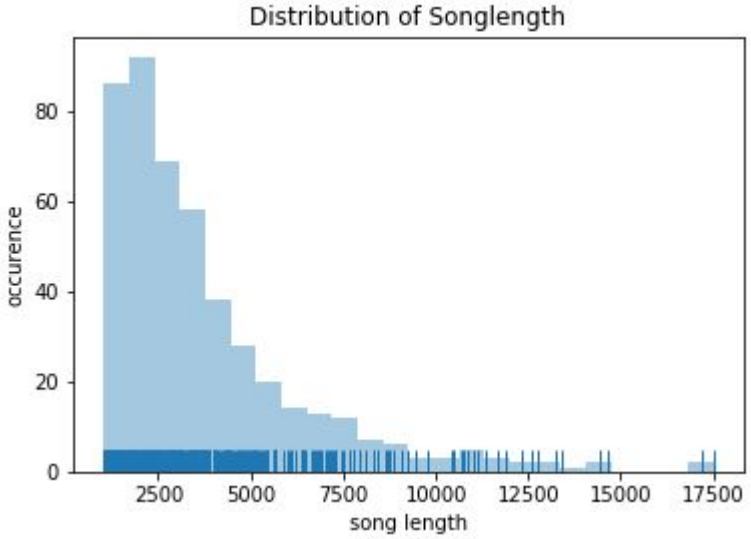
Überlegungen

- Genre und Subgenre neu sortieren
- word embedding
- SVM Features hinzufügen

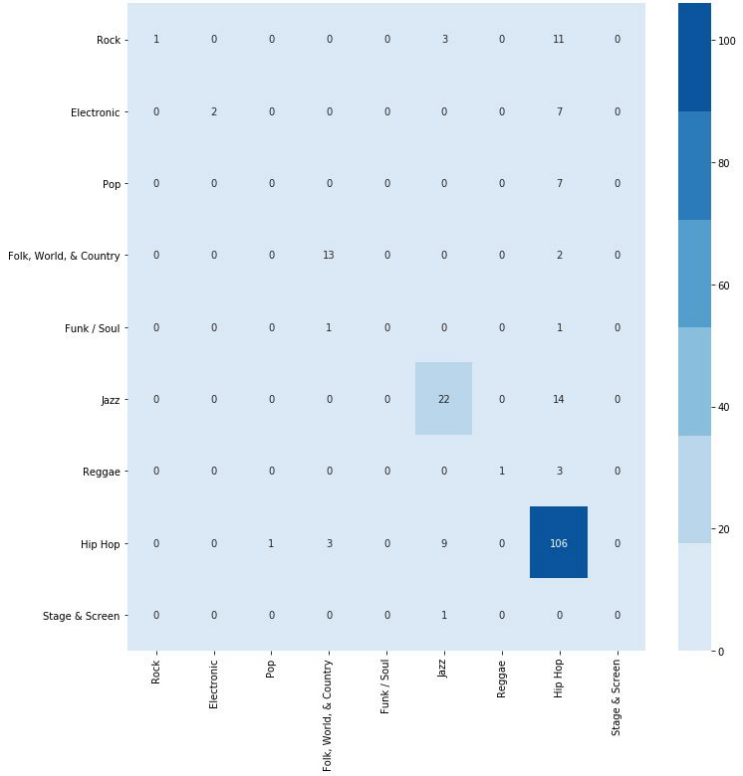
neuer Datensatz

- 466 Bands



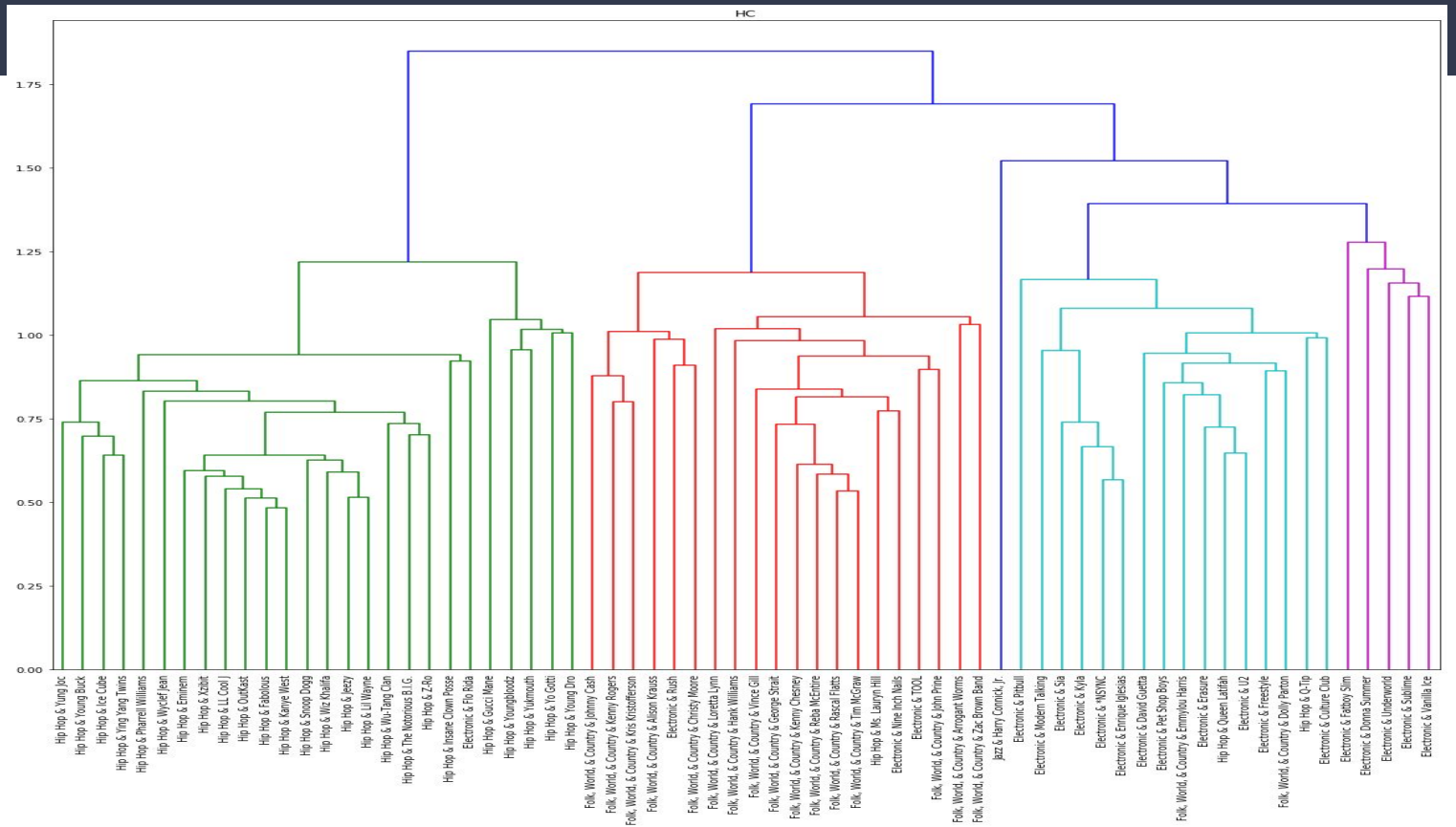


SVM



	precision	recall	f1-score	support
Rock	1.00	0.07	0.12	15
Electronic	1.00	0.22	0.36	9
Pop	0.00	0.00	0.00	7
Folk, World, & Country	0.76	0.87	0.81	15
Funk / Soul	0.00	0.00	0.00	2
Jazz	0.63	0.61	0.62	36
Reggae	1.00	0.25	0.40	4
Hip Hop	0.70	0.89	0.79	119
Stage & Screen	0.00	0.00	0.00	1
accuracy			0.70	208
macro avg	0.57	0.32	0.35	208
weighted avg	0.70	0.70	0.65	208

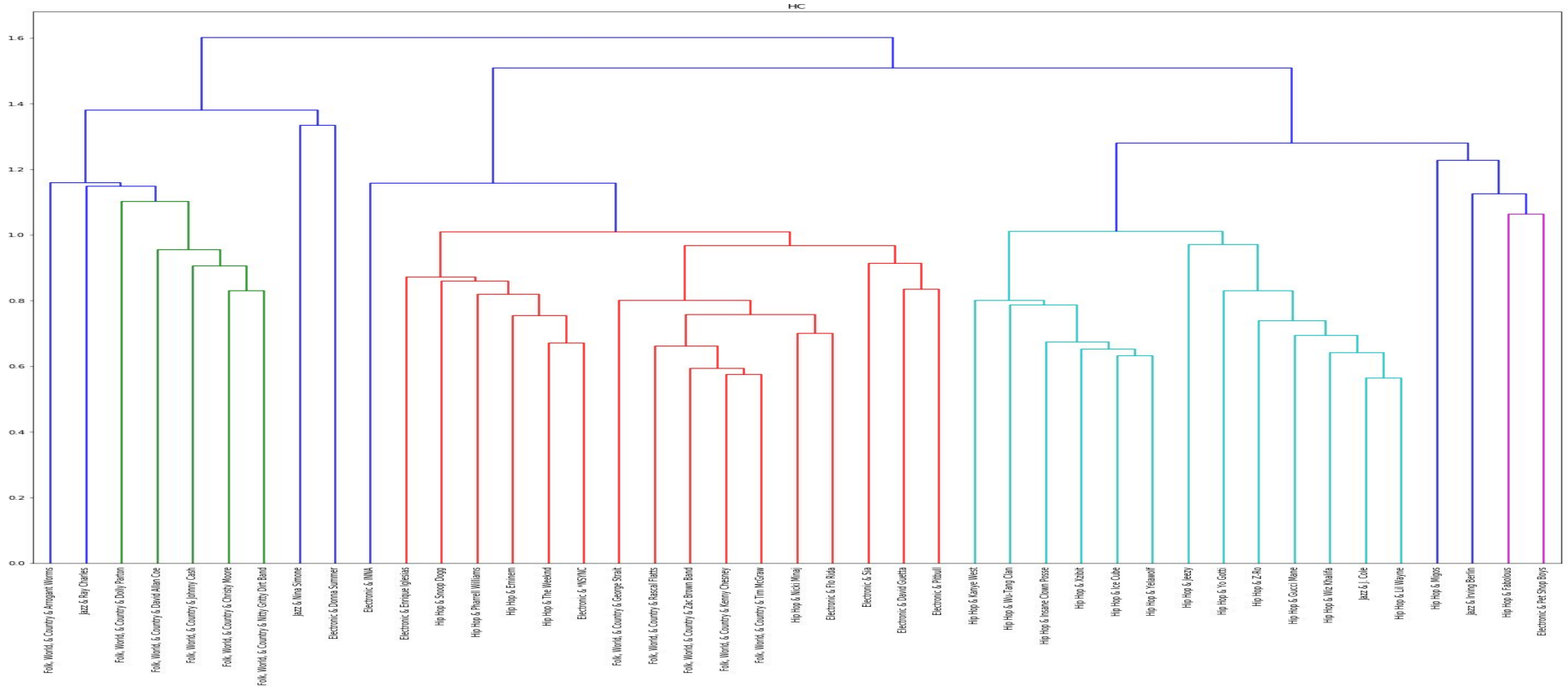
HC – 2000


HC
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
Hip Hop & Yung Joc
Hip Hop & Young Buck
Hip Hop & Ice Cube
Hip Hop & Ying Yang Twins
Hip Hop & Pharrell Williams
Hip Hop & Wyclef Jean
Hip Hop & Eminem
Hip Hop & Xzibit
Hip Hop & LL Cool J
Hip Hop & OutKast
Hip Hop & Fabolous
Hip Hop & Kanye West
Hip Hop & Snoop Dogg
Hip Hop & Wiz Khalifa
Hip Hop & Jeezy
Hip Hop & Lil Wayne
Hip Hop & Wu-Tang Clan
Hip Hop & The Notorious B.I.G.
Hip Hop & Z-Ro
Hip Hop & Insane Clown Posse
Electronic & Flo Rida
Hip Hop & Gucci Mane
Hip Hop & Youngbloodz
Hip Hop & Yukmouth
Hip Hop & Yo Gotti
Hip Hop & Young Dro
Folk, World, & Country & Johnny Cash
Folk, World, & Country & Kenny Rogers
Folk, World, & Country & Kris Kristofferson
Folk, World, & Country & Alison Krauss
Electronic & Rush
Folk, World, & Country & Christy Moore
Folk, World, & Country & Loretta Lynn
Folk, World, & Country & Hank Williams
Folk, World, & Country & Vince Gill
Folk, World, & Country & George Strait
Folk, World, & Country & Kenny Chesney
Folk, World, & Country & Reba McEntire
Folk, World, & Country & Rascal Flatts
Folk, World, & Country & Tim McGraw
Hip Hop & Ms. Lauryn Hill
Electronic & Nine Inch Nails
Electronic & TOOL
Folk, World, & Country & John Prine
Folk, World, & Country & Arrogant Worms
Folk, World, & Country & Zac Brown Band
Jazz & Harry Connick, Jr.
Electronic & Pitbull
Electronic & Modern Talking
Electronic & Sia
Electronic & Kyla
Electronic & *NSYNC
Electronic & Enrique Iglesias
Electronic & David Guetta
Electronic & Pet Shop Boys
Folk, World, & Country & Emmylou Harris
Electronic & Erasure
Hip Hop & Queen Latifah
Electronic & U2
Electronic & Freestyle
Folk, World, & Country & Dolly Parton
Hip Hop & Q-Tip
Electronic & Culture Club
Electronic & Fatboy Slim
Electronic & Donna Summer
Electronic & Underworld
Electronic & Sublime
Electronic & Vanilla Ice

Begründung für Ausreißer

- Flo Rida (Hip Hop) mit Electronic in Hip Hop : wird auch dem Genre EDM zugeordnet
- Nine Inch Nails (Electronic) mit Electronic in Country: für Country Award nominiert
- Tool (Electronic) in Country: laut Gruppenmitglied Adam Jones viele Einflüsse aus Country
- Queen Latifah (Hip Hop) mit Hip Hop in Electronic: Album “Nature of sista” house music und electronic

HC –2010


HC
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
Folk, World, & Country & Arrogant Worms
Jazz & Ray Charles
Folk, World, & Country & Dolly Parton
Folk, World, & Country & David Allan Coe
Folk, World, & Country & Johnny Cash
Folk, World, & Country & Chrissy Moore
Folk, World, & Country & Natty Gritty Dirt Band
Jazz & Nina Simone
Electronic & Donna Summer
Electronic & ANNA
Electronic & Enrique Iglesias
Hip Hop & Snoop Dogg
Hip Hop & Pharrell Williams
Hip Hop & Eminem
Hip Hop & The Weeknd
Electronic & *NSYNC
Folk, World, & Country & George Strait
Folk, World, & Country & Rascal Flatts
Folk, World, & Country & Zac Brown Band
Folk, World, & Country & Kenny Chesney
Folk, World, & Country & Tim McGraw
Hip Hop & Nicki Minaj
Electronic & Flo Rida
Electronic & Sia
Electronic & David Guetta
Electronic & Pitbull
Hip Hop & Kanye West
Hip Hop & Wu-Tang Clan
Hip Hop & Insane Clown Posse
Hip Hop & Xzibit
Hip Hop & Ice Cube
Hip Hop & Yelawolf
Hip Hop & Jeezy
Hip Hop & Yo Gotti
Hip Hop & 2Pac
Hip Hop & Gucci Mane
Hip Hop & Wiz Khalifa
Jazz & J. Cole
Hip Hop & Lil Wayne
Hip Hop & Migos
Jazz & Irving Berlin
Hip Hop & Fabolous
Electronic & Pet Shop Boys

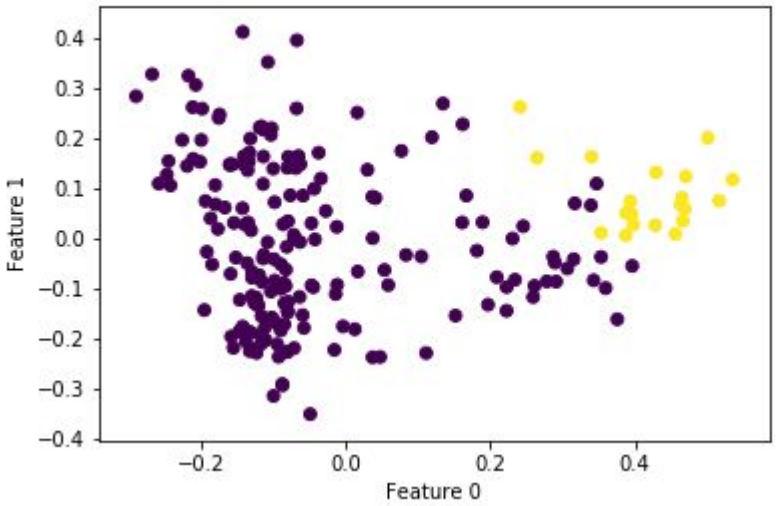
Begründung für Ausreißer

- Enrique Iglesias mit Electronic in Hip Hop: Zusammenarbeit mit Hip Hop- Künstlern/ Hip Hop Reggaeton / Song 2008 Hip Hop und R&B
- Nicki Minaj mit Hip Hop in Electronic: Lieder werden teilweise dem Electro-House Genre zugeordnet und Zusammenarbeit mit David Guetta und Flo Rida (2010)
- J. Cole (Rapper) mit Jazz in Hip Hop: 4. Album hoher Jazz Einfluss
- Pet Shop Boys (Electro Pop) in Hip Hop: 2002-2005 verschieden Genre, unter anderem Hip Hop

Ausreißer nach Wahrscheinlichkeit

artist	genre	decade	0	1	2	3
Nine Inch Nails	Electronic	2000s	0.0	2.706984e-239	1.000000e+00	0.0
George Strait	Folk, World, & Country	2000s	0.0	8.050388e-149	1.000000e+00	0.0
Enrique Iglesias	Electronic	2000s	0.0	1.000000e+00	2.388637e-125	0.0

DBscan



Viktoria Ermisch: 1986364

Timo Günther: 2033581

Julia Jäger: 2124649

Teresa Kaiser: 2353056