Makroperspektive

Können wir Genres und Sprachen auch in ihrer Gesamtheit mit Clustering-Techniken analysieren?

Vorgehen

1. Zusammenfassen aller Texte mit gleichem Label (Sprachen/ Genres)
2. Filtern von zusammengefassten Instanzen die aus nur wenigen Texten bestehen
3. Cluster-Analyse der zusammengefassten Daten
    1. KMeans, BayesianGaussianMixture-Models
       => Anzahl der Cluster = Anzahl an Einzelsprachen
    2. DBSCAN => "Tuning" der *eps* und *min_samples* Parameter
4. Evaluation: durch subjektive Analyse der Cluster & (Davies Bouldin Score + Calinski Harabasz Score)

Schwachstellen dieses Ansatzes

- Die Anzahl der Texte (sowie die Anzahl der Tokens) sind stark schief verteilt.
- Geeignete Evaluationsmetrik(en)?

Feature-Extraction Pipeline

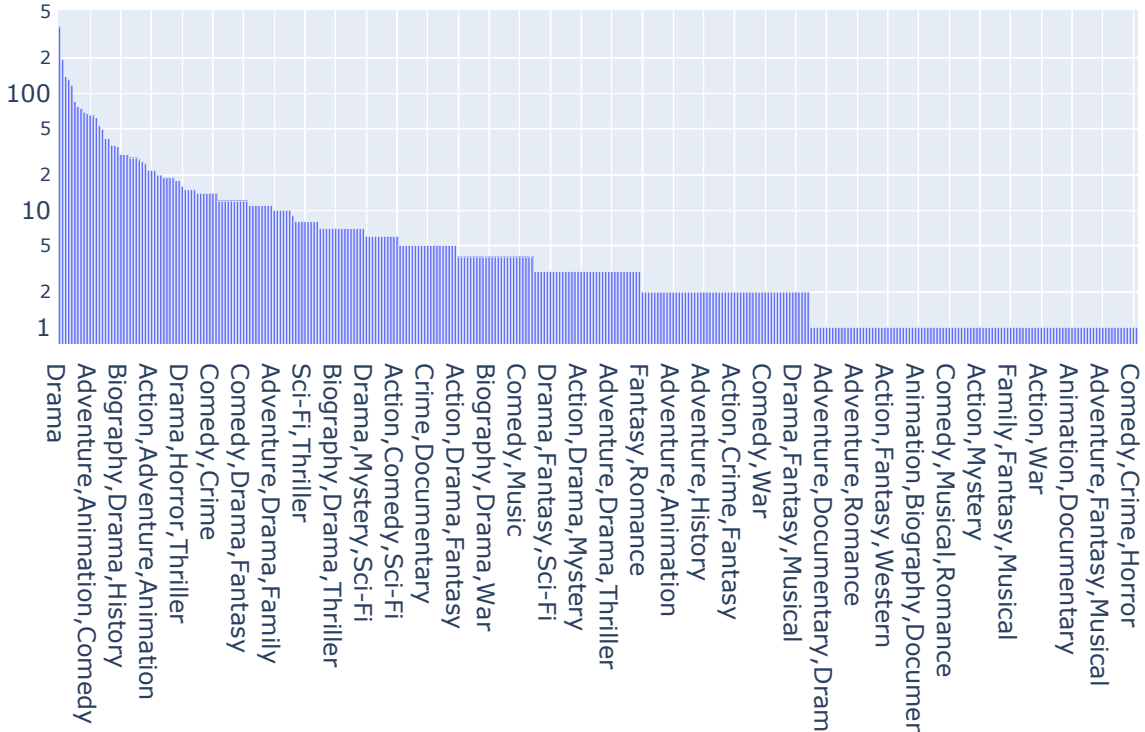
1. `TfidfVectorizer(max_features=[100, 150, 250, 500, 1000, 5000], stop_words=None)`
2. (`UMAP(n_components=[90, 70, 50, 30, 25, 10])` oder `PCA(<n_components von UMAP>)`)
    - => Maximale Anzahl der Components durch Anzahl an Instanzen beschränkt

Besonderheit: Ebenfalls wurde auch eine gefilterte Version des Textes verwendet bei der alle Wörter bis auf die Stopwörter entfernt wurden.

Visualisierungen

`UMAP(n_components=2, n_neigbors=[5,..,10])` auf den Features aus der Pipeline

Genres: Textverteilung

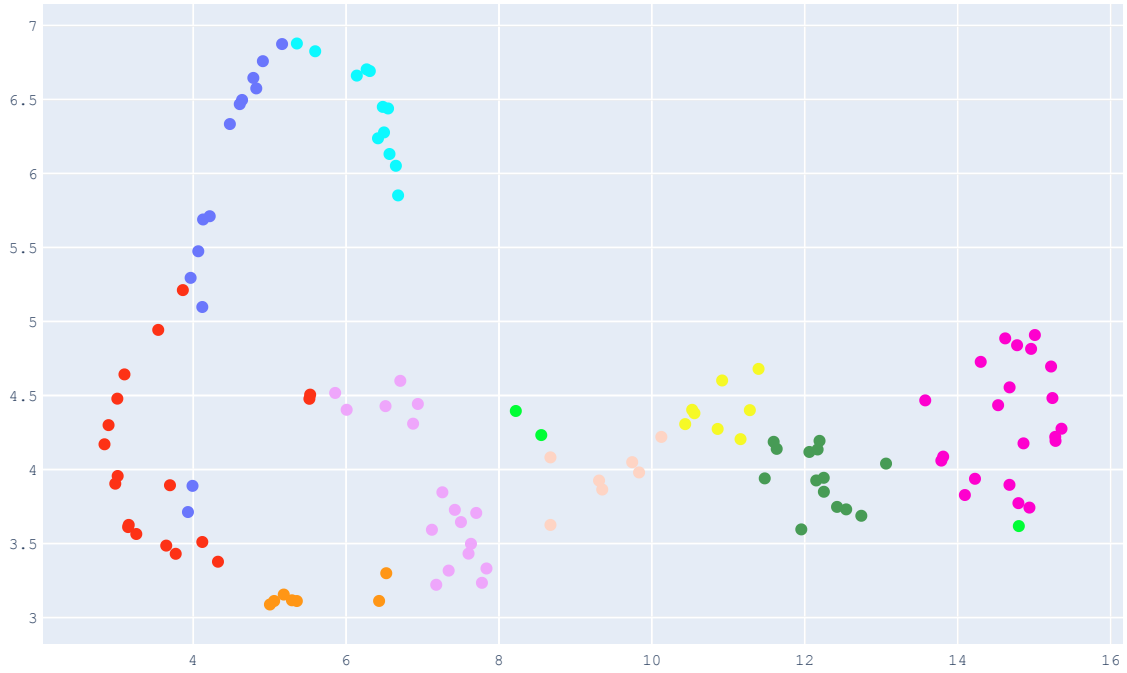


Genres: Beste Feauture-Kombination

```
Pipeline(steps=[('tfidfvectorizer', TfidfVectorizer(max_features=1000)),
                ('densetransformer', DenseTransformer(),
                ('umap', UMAP(n_components=10))])
```

Genres: KMeans

```
[(TfidfVectorizer(max_features=1000), UMAP(n_components=10))]=>KMeans(n_clusters=10)
```

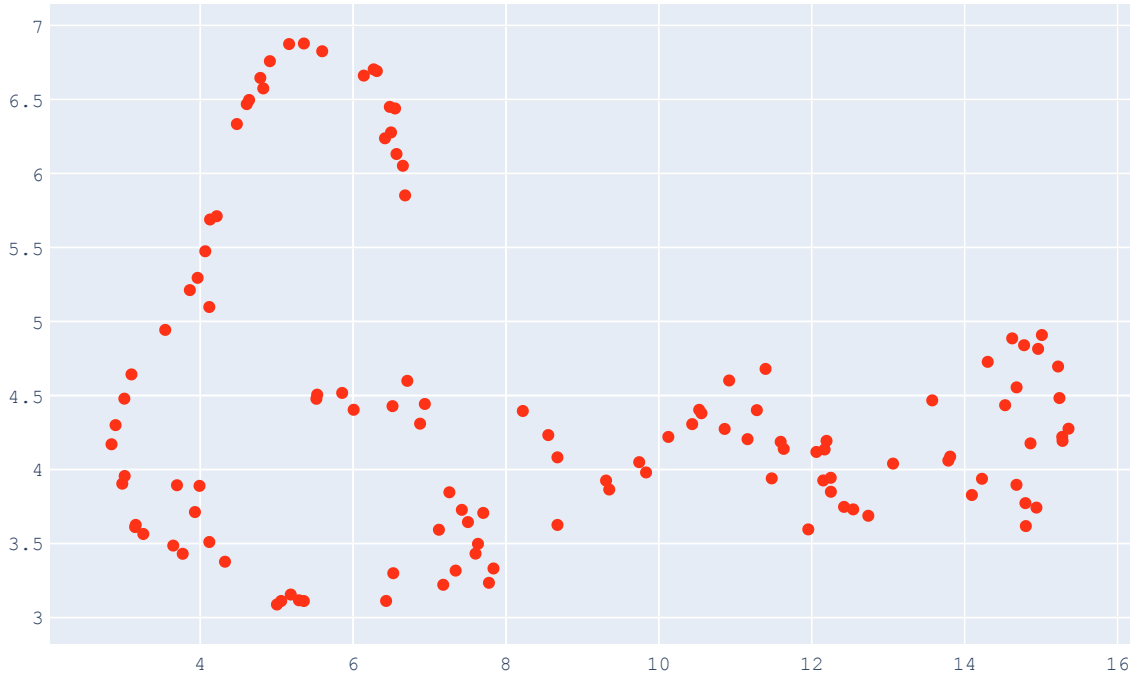


Genres: KMeans

	Cluster	Genre
0	0	Action, Crime, Thriller | Crime, Drama, Mystery, Thriller | Action, Adventure, Sci-Fi | Action, Adventure, Sci-Fi, Thriller | Action, Sci-Fi, Thriller | Action, Drama, Thriller | Action, Adventure, Thriller | Crime, Mystery, Thriller | Mystery, Thriller | Action, Horror, Sci-Fi | Drama, Sci-Fi | Drama, Western | Biography, Crime, Drama | Action, Sci-Fi | Action, Horror, Sci-Fi, Thriller | Biography, Comedy, Drama | Comedy, Drama, Music | Drama, History, Romance, War
1	1	Action, Comedy | Comedy, Crime | Adventure, Comedy, Family
2	2	Documentary | Biography, Drama | Biography, Drama, History | Biography, Drama, Romance | Drama, History, War | Action, Adventure | Documentary, Music | Action, Drama, History, War | Action, Drama, War | Documentary, Biography, Music | Biography, Crime, Drama, Thriller | Drama, History, Romance | Biography, Drama, History, War | Biography, Drama, War
3	3	Action, Comedy, Crime | Drama, Horror, Thriller | Drama, Horror, Mystery, Thriller | Drama, Romance, Thriller | Drama, Mystery | Drama, Horror | Drama, Sci-Fi, Thriller
4	4	Comedy | Comedy, Horror | Animation, Adventure, Comedy, Family | Animation, Adventure, Comedy, Family, Fantasy | Family | Action, Comedy, Horror | Adventure, Family | Comedy, Family | Adventure, Comedy | Adventure, Drama, Family | Adventure, Comedy, Family, Fantasy | Drama, Music | Adventure, Comedy, Drama | Comedy, Drama, Family | Action, Adventure, Comedy, Sci-Fi | Animation, Adventure, Family | Animation, Family | Comedy, Horror, Sci-Fi | Action, Adventure, Drama, Thriller | Comedy, Music | Comedy, Sci-Fi
5	5	Action | Action, Drama | Action, Adventure, Fantasy | Drama, History | Action, Drama, History | Action, Crime, Drama | Action, Adventure, Drama, Fantasy | Adventure, Fantasy | Comedy, Drama, History | Comedy, Fantasy | Western | Biography, Drama, Sport
6	6	Comedy, Drama | Comedy, Drama, Romance | Drama, Romance | Comedy, Romance | Drama, Fantasy, Romance | Drama, Romance, Sci-Fi | Comedy, Horror, Romance | Comedy, Drama, Music, Romance
7	7	Horror, Sci-Fi | Horror, Sci-Fi, Thriller | Action, Horror | Drama, Mystery, Sci-Fi, Thriller | Horror, Mystery | Fantasy, Horror, Thriller | Action, Horror, Thriller
8	8	Drama | Horror | Crime, Drama | Crime, Thriller | Drama, Family | Comedy, Crime, Drama | Adventure, Family, Fantasy | Drama, Fantasy | Drama, Sport | Crime | Action, Crime | Romance | Drama, Romance, War | Action, Comedy, Drama
9	9	Horror, Thriller | Drama, Thriller | Crime, Drama, Thriller | Action, Crime, Drama, Thriller | Action, Thriller | Thriller | Horror, Mystery, Thriller | Drama, Mystery, Thriller | Drama, War | Comedy, Crime, Thriller | Action, Crime, Drama, Mystery, Thriller | Action, Comedy, Crime, Thriller | Sci-Fi, Thriller | Drama, Horror, Sci-Fi, Thriller | Action, Comedy, Horror, Thriller | Crime, Drama, Horror, Thriller | Action, Adventure, Comedy

Genres: DBScan

```
[(TfidfVectorizer(max_features=1000), UMAP(n_components=10))] =>DBSCAN(eps=0.15
```

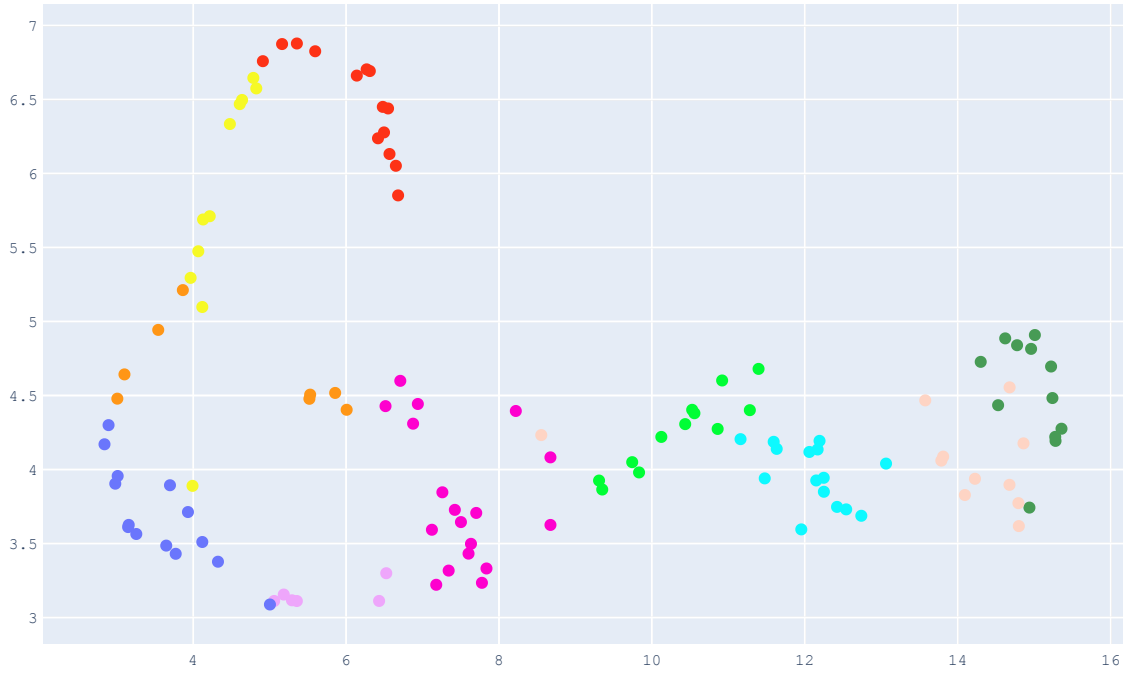


Genres: DBScan

	Cluster	Genre
0	0	Drama | Comedy | Comedy, Drama | Comedy, Drama, Romance | Drama, Romance | Comedy, Romance | Horror, Thriller | Drama, Thriller | Documentary | Horror | Crime, Drama, Thriller | Action, Crime, Drama, Thriller | Action, Crime, Thriller | Action, Thriller | Thriller | Horror, Mystery, Thriller | Crime, Drama | Action | Crime, Drama, Mystery, Thriller | Biography, Drama | Action, Adventure, Sci-Fi | Comedy, Horror | Action, Drama | Drama, Mystery, Thriller | Action, Comedy, Crime | Biography, Drama, History | Animation, Adventure, Comedy, Family | Crime, Thriller | Drama, Horror, Thriller | Drama, Family | Action, Comedy | Biography, Drama, Romance | Action, Adventure, Sci-Fi, Thriller | Action, Sci-Fi, Thriller | Action, Adventure, Fantasy | Comedy, Crime, Drama | Animation, Adventure, Comedy, Family, Fantasy | Action, Drama, Thriller | Action, Adventure, Thriller | Drama, History | Action, Drama, History | Comedy, Crime | Family | Crime, Mystery, Thriller | Action, Comedy, Horror | Drama, Horror, Mystery, Thriller | Adventure, Family, Fantasy | Horror, Sci-Fi | Horror, Sci-Fi, Thriller | Drama, History, War | Adventure, Family | Action, Crime, Drama | Mystery, Thriller | Comedy, Family | Drama, Romance, Thriller | Adventure, Comedy | Drama, Fantasy | Drama, War | Action, Horror | Drama, Mystery | Action, Horror, Sci-Fi | Action, Adventure | Drama, Sport | Adventure, Drama, Family | Drama, Sci-Fi | Adventure, Comedy, Family, Fantasy | Drama, Mystery, Sci-Fi, Thriller | Comedy, Crime, Thriller | Drama, Horror | Action, Crime, Drama, Mystery, Thriller | Crime | Action, Adventure, Drama, Fantasy | Horror, Mystery | Drama, Music | Documentary, Music | Adventure, Comedy, Drama | Comedy, Drama, Family | Action, Drama, History, War | Drama, Sci-Fi, Thriller | Adventure, Fantasy | Drama, Western | Biography, Crime, Drama | Action, Adventure, Comedy, Sci-Fi | Animation, Adventure, Family | Animation, Family | Drama, Fantasy, Romance | Action, Drama, War | Action, Sci-Fi | Action, Horror, Sci-Fi, Thriller | Action, Comedy, Crime, Thriller | Biography, Comedy, Drama | Comedy, Drama, Music | Adventure, Comedy, Family | Comedy, Drama, History | Documentary, Biography, Music | Sci-Fi, Thriller | Fantasy, Horror, Thriller | Action, Crime | Comedy, Horror, Sci-Fi | Action, Adventure, Drama, Thriller | Comedy, Music | Drama, History, Romance, War | Romance | Drama, Horror, Sci-Fi, Thriller | Biography, Crime, Drama, Thriller | Comedy, Fantasy | Drama, Romance, War | Drama, Romance, Sci-Fi | Action, Comedy, Horror, Thriller | Western | Biography, Drama, Sport | Crime, Drama, Horror, Thriller | Comedy, Horror, Romance | Comedy, Drama, Music, Romance | Drama, History, Romance | Action, Horror, Thriller | Biography, Drama, History, War | Comedy, Sci-Fi | Action, Comedy, Drama | Action, Adventure, Comedy | Biography, Drama, War

Genres: Gaussian-Mixture Model

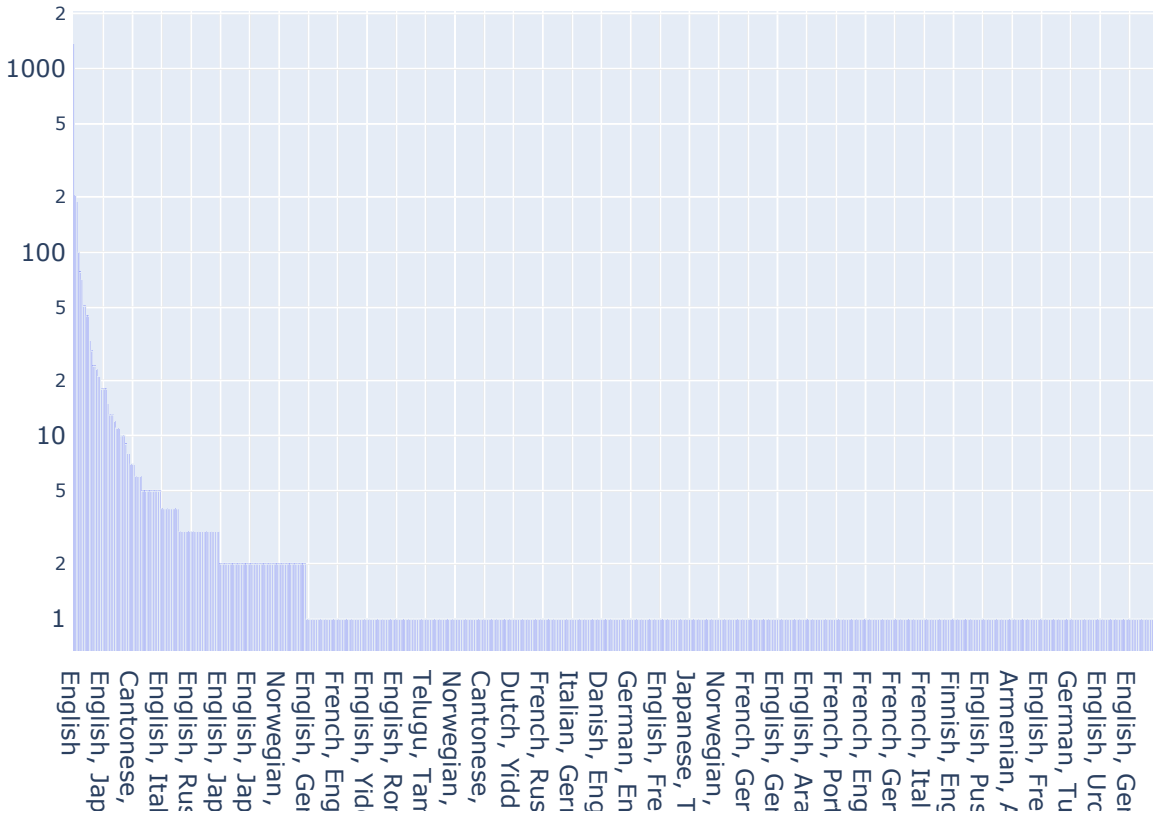
```
[(TfidfVectorizer(max_features=1000), UMAP(n_components=10))] =>BayesianGaussianMixture(n_components=
```



Genres: Gaussian-Mixture Model

	Cluster	Genre
0	0	Action | Action, Drama | Action, Adventure, Fantasy | Drama, History | Action, Drama, History | Action, Crime, Drama | Action, Adventure | Action, Adventure, Drama, Fantasy | Action, Drama, History, War | Adventure, Fantasy | Comedy, Drama, History | Comedy, Fantasy | Western | Biography, Drama, Sport
1	1	Comedy, Drama | Comedy, Drama, Romance | Drama, Romance | Comedy, Romance | Drama, Horror, Thriller | Drama, Horror, Mystery, Thriller | Drama, Mystery | Drama, Horror | Drama, Sci-Fi, Thriller | Drama, Fantasy, Romance | Drama, Romance, Sci-Fi | Comedy, Horror, Romance
2	2	Action, Adventure, Sci-Fi | Action, Adventure, Sci-Fi, Thriller | Action, Sci-Fi, Thriller | Action, Drama, Thriller | Action, Adventure, Thriller | Crime, Mystery, Thriller | Action, Horror, Sci-Fi | Drama, Sci-Fi | Drama, Mystery, Sci-Fi, Thriller | Drama, Western | Action, Drama, War | Action, Sci-Fi | Action, Horror, Sci-Fi, Thriller | Drama, History, Romance, War
3	3	Comedy, Horror | Comedy, Crime | Action, Comedy, Horror | Adventure, Comedy | Adventure, Drama, Family | Adventure, Comedy, Drama | Comedy, Drama, Family | Adventure, Comedy, Family | Comedy, Horror, Sci-Fi | Action, Adventure, Drama, Thriller | Comedy, Sci-Fi
4	4	Horror, Thriller | Drama, Thriller | Crime, Drama, Thriller | Action, Crime, Drama, Thriller | Thriller | Horror, Mystery, Thriller | Drama, Mystery, Thriller | Action, Comedy, Crime | Action, Comedy | Drama, Romance, Thriller | Drama, War | Comedy, Crime, Thriller | Action, Comedy, Crime, Thriller | Sci-Fi, Thriller | Drama, Horror, Sci-Fi, Thriller | Action, Comedy, Horror, Thriller | Crime, Drama, Horror, Thriller | Action, Adventure, Comedy
5	5	Drama | Horror | Crime, Drama | Crime, Thriller | Drama, Family | Comedy, Crime, Drama | Adventure, Family, Fantasy | Drama, Fantasy | Drama, Sport | Crime | Action, Crime | Romance | Drama, Romance, War | Comedy, Drama, Music, Romance | Action, Comedy, Drama
6	6	Documentary | Biography, Drama | Biography, Drama, History | Biography, Drama, Romance | Drama, History, War | Documentary, Music | Documentary, Biography, Music | Biography, Crime, Drama, Thriller | Drama, History, Romance | Biography, Drama, History, War | Biography, Drama, War
7	7	Action, Crime, Thriller | Action, Thriller | Crime, Drama, Mystery, Thriller | Mystery, Thriller | Action, Crime, Drama, Mystery, Thriller | Biography, Crime, Drama | Biography, Comedy, Drama | Comedy, Drama, Music
8	8	Comedy | Animation, Adventure, Comedy, Family | Animation, Adventure, Comedy, Family, Fantasy | Family | Adventure, Family | Comedy, Family | Adventure, Comedy, Family, Fantasy | Drama, Music | Action, Adventure, Comedy, Sci-Fi | Animation, Adventure, Family | Animation, Family | Comedy, Music
9	9	Horror, Sci-Fi | Horror, Sci-Fi, Thriller | Action, Horror | Horror, Mystery | Fantasy, Horror, Thriller | Action, Horror, Thriller

Languages: Textverteilung

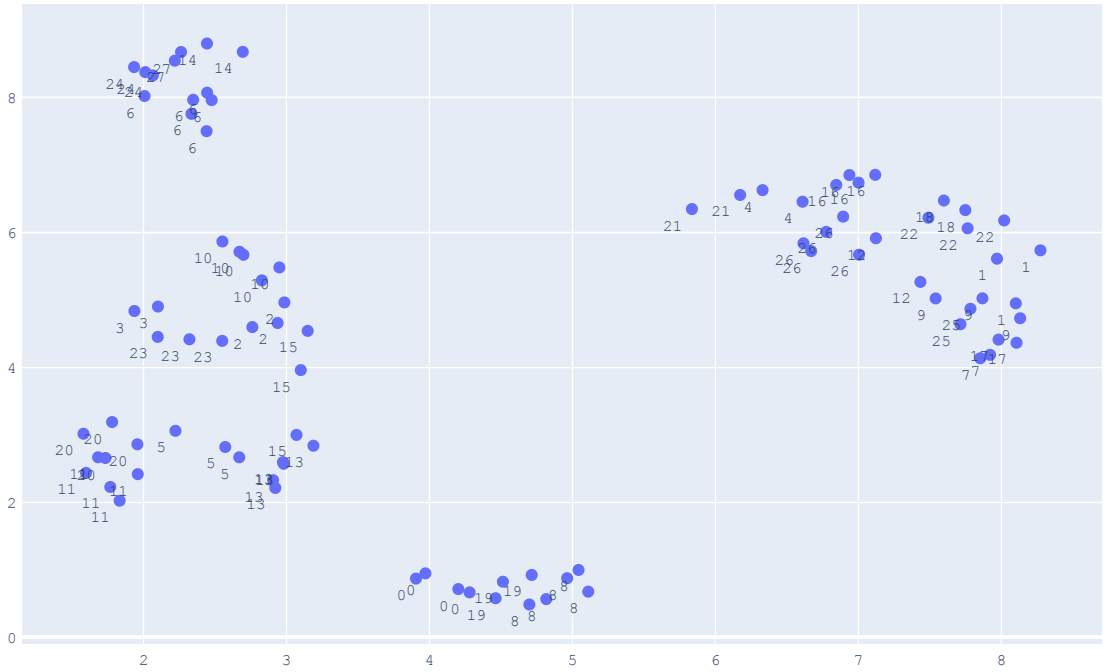


Languages: Beste Feauture-Kombination

```
Pipeline(steps=[('tfidfvectorizer', TfidfVectorizer(max_features=150)),
                ('densetransformer', DenseTransformer(),
                ('umap', UMAP(n_components=40, random_state=42))])
```

Languages: KMeans

```
[(TfidfVectorizer(max_features=150), UMAP(n_components=40, random_state=42))]=>KMeans(n_clusters=28)
```

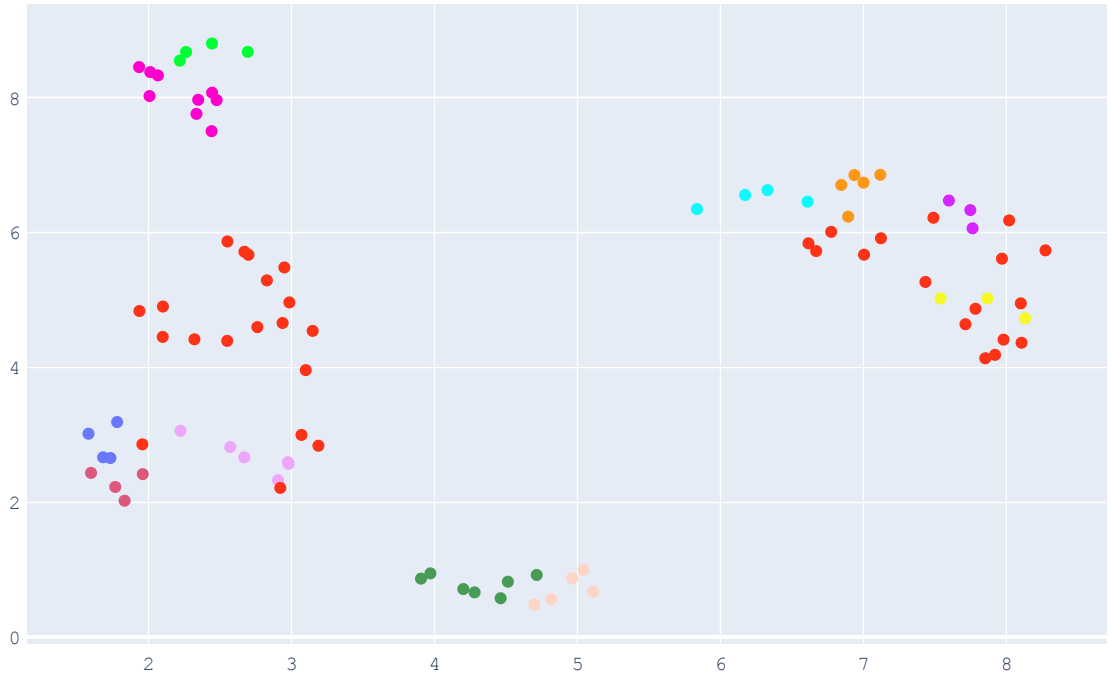


Languages: KMeans

	Cluster	Language
0	0	French, German | French, Italian | English, Russian, French | Cantonese, Mandarin, English
1	1	Danish | Japanese, English | German, Russian, English
2	2	Spanish | English, Arabic, Hebrew | Polish, English
3	3	Norwegian, English | English, Italian, Spanish
4	4	Cantonese | Korean, Mandarin
5	5	English | English, French, Italian | Swedish, English
6	6	Norwegian | Swedish | Icelandic | Swiss German | Norwegian, Swedish | English, Serbo-Croatian
7	7	English, French, German | French, German, English
8	8	English, French | English, German | English, Arabic | English, Mandarin, Cantonese | French, English, Italian
9	9	Cantonese, English | Polish | German, French
10	10	Turkish | Italian | Portuguese | Romanian | Turkish, English
11	11	English, Russian | English, Ukrainian | English, Mandarin | English, Greek | English, Korean
12	12	Mandarin, English | Greek
13	13	French | English, Italian | English, Latin | French, Spanish | Dutch, English
14	14	German | German, English
15	15	French, English | French, Arabic | Arabic, French
16	16	Chinese | Arabic | Mandarin, Cantonese, English | Vietnamese
17	17	English, Japanese | Hebrew
18	18	Russian, English | Hindi, English
19	19	Hindi | English, Cantonese | English, Klingon
20	20	English, Spanish | English, Portuguese | English, Hebrew | English, Thai
21	21	Thai, English | Filipino, Tagalog
22	22	Japanese | Russian | Japanese, Mandarin

Languages: DBScan

```
[(TfidfVectorizer(max_features=150), UMAP(n_components=40, random_state=42))] =
```

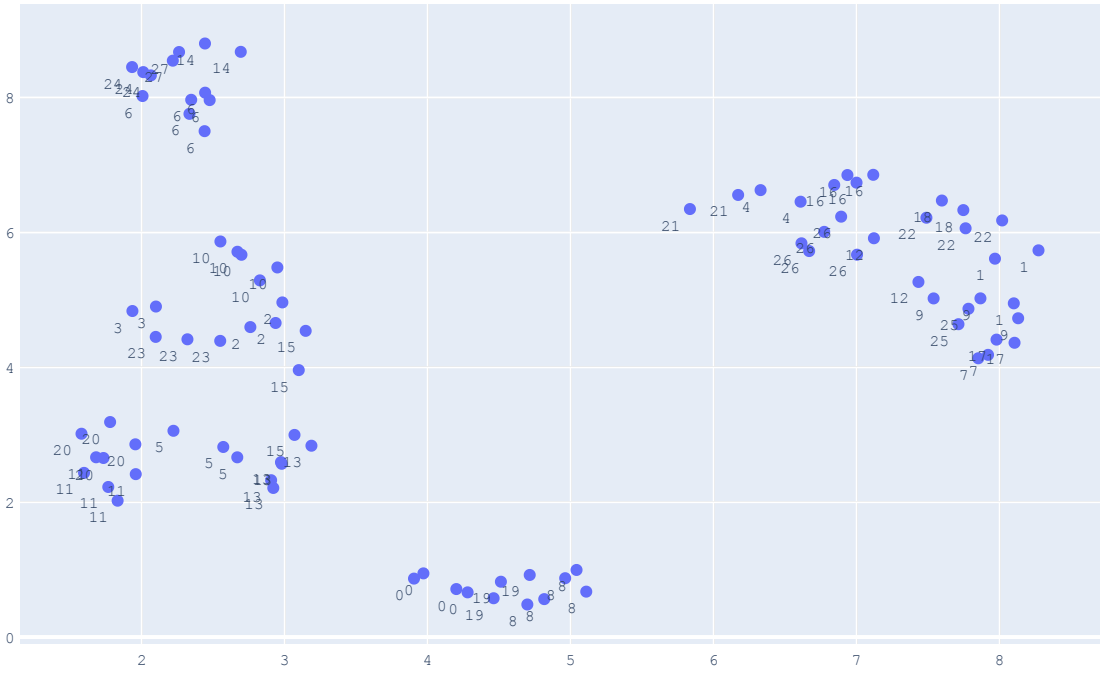


Languages: DBScan

	Cluster	Language
0	-1	French, English | Mandarin | Danish | English, Japanese | Italian | English, Italian | Portuguese | Japanese, English | Spanish, English | English, Portuguese | Mandarin, English | English, Latin | Greek | English, Chinese | Romanian | Cantonese, Mandarin | Hebrew | German, Russian, English | Polish, English | English, Spanish, Russian
1	0	English | French | English, French, Italian | Swedish, English | French, Spanish | Dutch, English
2	1	German | German, English | Dutch | Norwegian, German
3	2	Japanese | Japanese, Mandarin
4	3	English, Spanish | English, Greek | English, Hebrew | English, Thai
5	4	Spanish | English, Arabic, Hebrew
6	5	Korean | Korean, English
7	6	English, French | English, German | English, Arabic | English, Mandarin, Cantonese | French, English, Italian
8	7	Turkish | Turkish, English
9	8	English, Russian | English, Ukrainian | English, Mandarin | English, Korean
10	9	Norwegian | Swedish | Icelandic | German, Turkish | Swiss German | Norwegian, Swedish | Finnish | English, Serbo-Croatian | German, Italian
11	10	Russian | Russian, English | Hindi, English
12	11	Cantonese | Thai, English | Filipino, Tagalog | Korean, Mandarin
13	12	Thai | Chinese | Arabic | Mandarin, Cantonese, English | Vietnamese
14	13	Hindi | French, German | French, Italian | English, Cantonese | English, Russian, French | English, Klingon | Cantonese, Mandarin, English
15	14	Mandarin, Cantonese | French, Arabic, English
16	15	Cantonese, English | Polish | German, French
17	16	French, Arabic | Arabic, French
18	17	Norwegian, English | English, Italian, Spanish
19	18	English, French, German | French, German, English

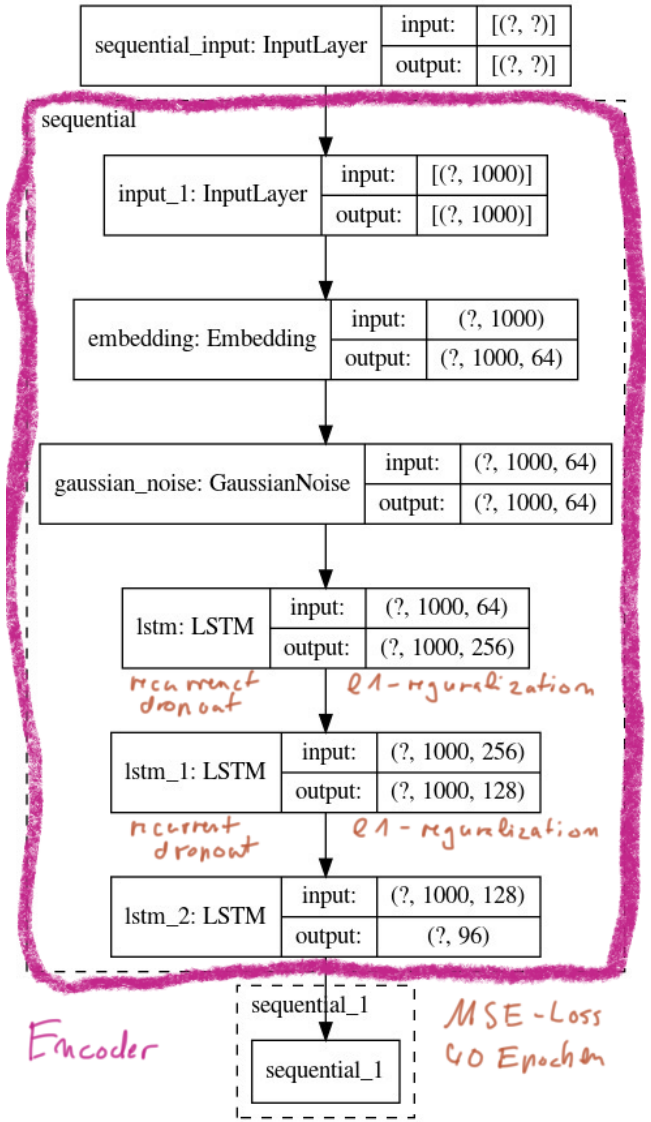
Languages: Gaussian-Mixture Model

```
[(TfidfVectorizer(max_features=150), UMAP(n_components=40, random_state=42))] =>BayesianGaussianMixtu
```



Languages: Gaussian-Mixture Model

	Cluster	Language
0	0	English, French | English, German | English, Arabic | French, English, Italian
1	1	Japanese | Danish | Japanese, English | Japanese, Mandarin
2	2	Turkish | Italian | Portuguese | Romanian | Turkish, English
3	3	English, Spanish | English, Thai
4	4	German, Turkish | Finnish | Norwegian, German | German, Italian
5	5	French, English | English, Latin | Arabic, French
6	6	Cantonese, English | Greek | Polish | German, Russian, English | German, French
7	7	Thai | Arabic
8	8	Cantonese | Thai, English | Filipino, Tagalog | Korean, Mandarin
9	9	French, German | French, Italian | English, Russian, French | Cantonese, Mandarin, English
10	10	Spanish | Spanish, English | French, Arabic | English, Arabic, Hebrew | Polish, English
11	11	English, French, German | French, German, English
12	12	English, Russian | English, Ukrainian | English, Mandarin | English, Korean
13	13	German | German, English | Dutch
14	14	Norwegian, English | English, Italian, Spanish
15	15	Icelandic | Swiss German
16	16	Chinese | Mandarin, Cantonese, English | Vietnamese
17	17	French | French, Spanish | Dutch, English
18	18	Mandarin, Cantonese | French, Arabic, English
19	19	English, Portuguese | English, Greek | English, Hebrew
20	20	Russian | Russian, English | Hindi, English
21	21	English, Japanese | Hebrew
22	22	Mandarin | Mandarin, English | Cantonese, Mandarin



Autoencoder

Architektur

Training:

- num_words: 2500
- Maximale Sequenzlänge: 1000
- Epochen: 40

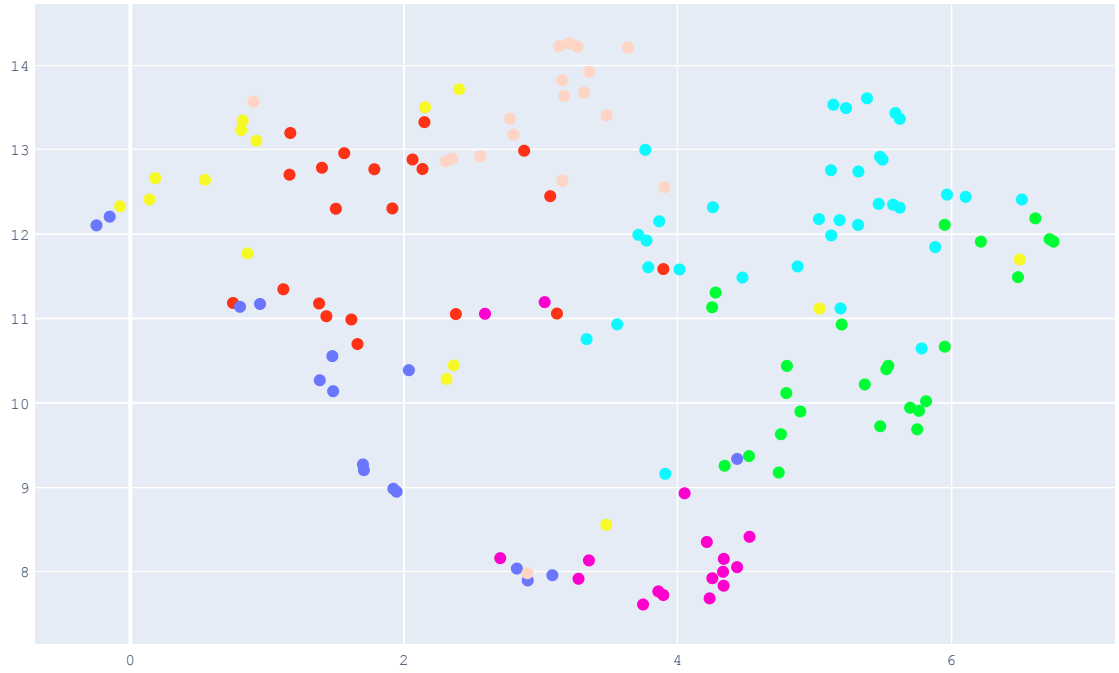
Trainingsdaten:

- Maximale Sequenzlänge ist ein Beschränkung bei rekurrenten Netzen
  - => Bei praktikablen Sequenzlänge würden wir nur den ersten Text betrachten
- Deshalb wurden 250 Sätze pro Instanz zufällig gesampelt und konkanteniert

Autoencoder

Ergebnisse

```
[Autoencoder => KMeans(n_clusters=7)
```



Fazit

1. Inbesondere die Erkennung der Originalsprachen anhand ihrer deutschen Übersetzungen scheint ein *stilometrisches* Problem zu sein, da der Fokus auf wenige häufige (Stop-)Wörter die Ergebnisse verbessert
2. UMAP und PCA als Techniken zur Dimensionreduktion auf den Tfidf-Feature scheinen ebenfalls einen positiven Einfluss auf die Ergebnisse zu haben
3. DBSCAN ist extrem instabil und funktioniert eher nicht auf Textdaten