Untertitel

Clustering

Gliederung

1. Datensatzvorstellung
2. Korpusexploration mittels Topic Modelling
3. Analyse von Sprache und Genre durch Clusterverfahren
    a. Mikroperspektive
    b. Makroperspektive
4. Fazit

Datensatz

OpenSubtitles (http://opus.nlpl.eu/OpenSubtitles-v2018.php)

- total number of files: 3,735,070
- total number of german subtitle files: 81,062
- 62 languages
- often multiple subtitle versions for a movie

Datensatz

Ergänzung des Datensatzes durch zusätzliche Metadaten

1. IMDB Datensätze
2. IMDB API

Was wurde eliminiert:

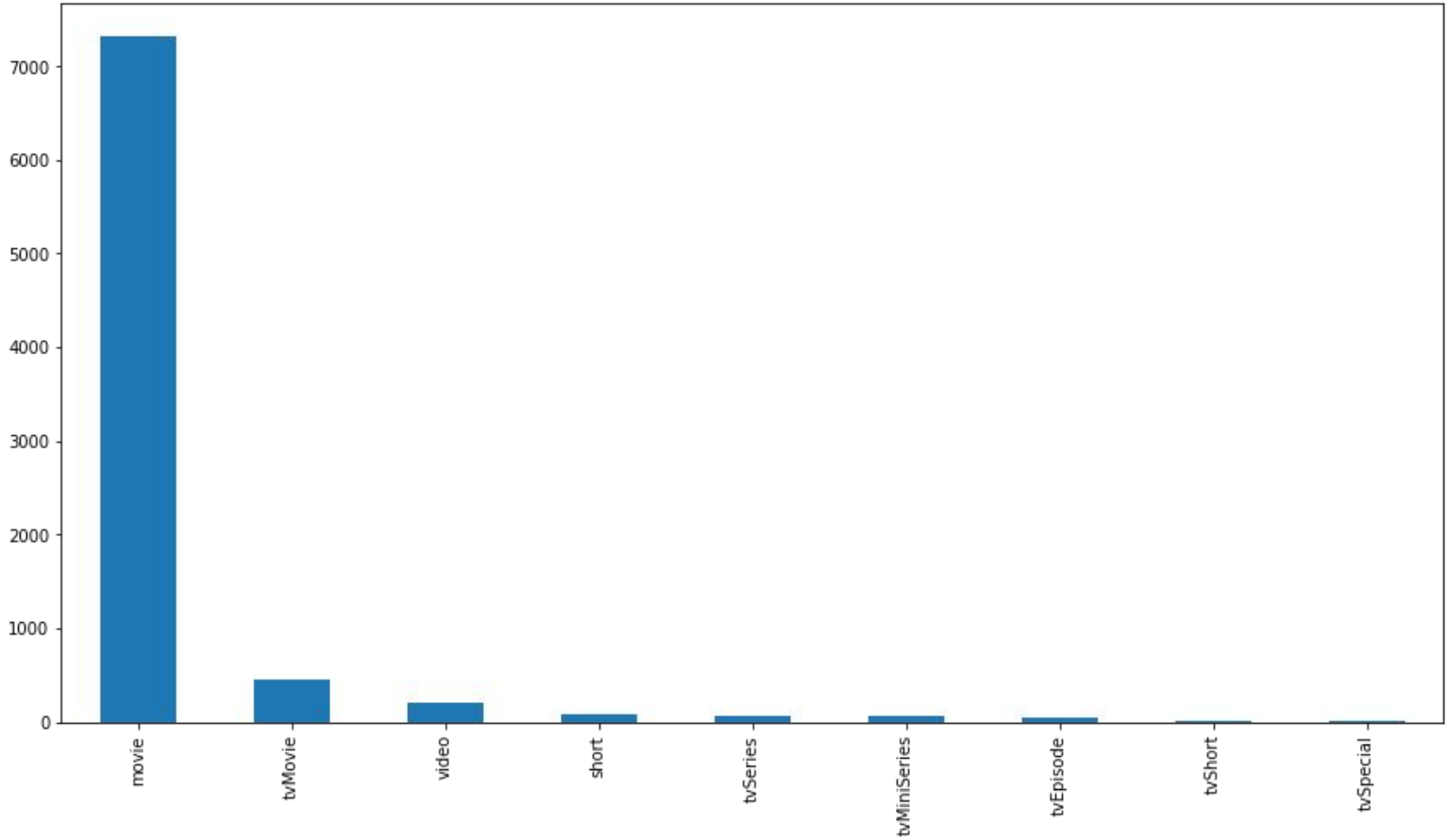
- alles außer Filme
- kurze Texte unter 3000 Zeichen
- Instanzen mit fehlenden relevanten Metadaten
- Duplikate

→ final number of files: 3,728

Probleme:

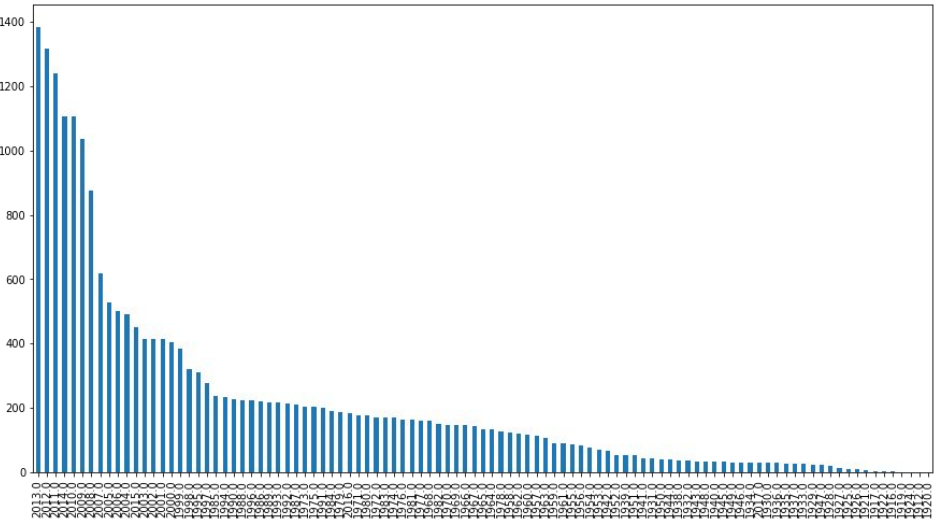
1. Statements in Texten
2. ungleiche Verteilung der Instanzen hinsichtlich der Label

Datensatz **Plots**

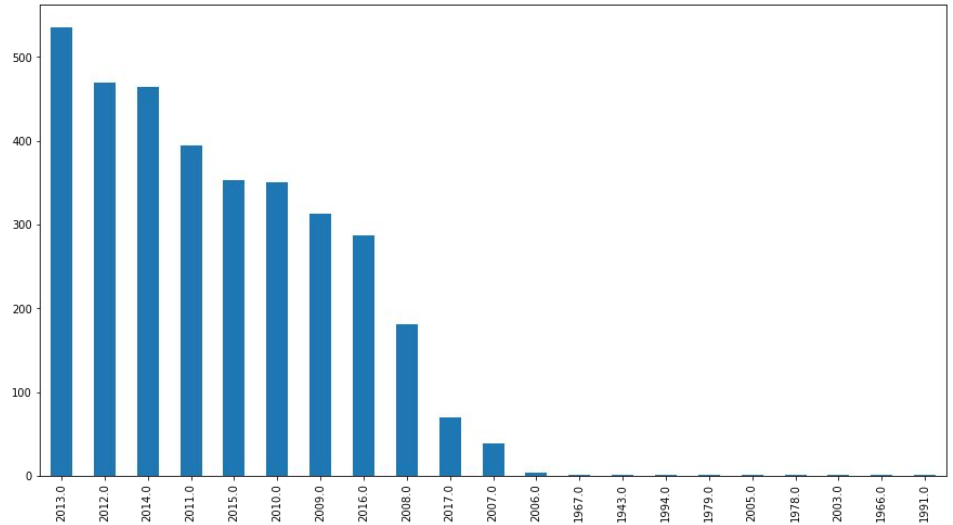


Datensatz **Plots**

Year Original

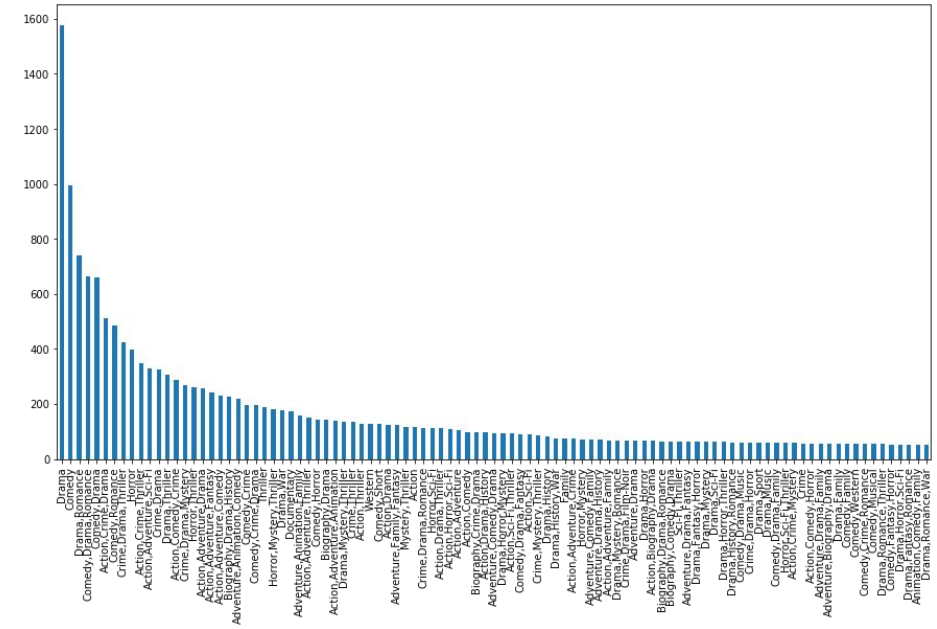


Year Bereinigt

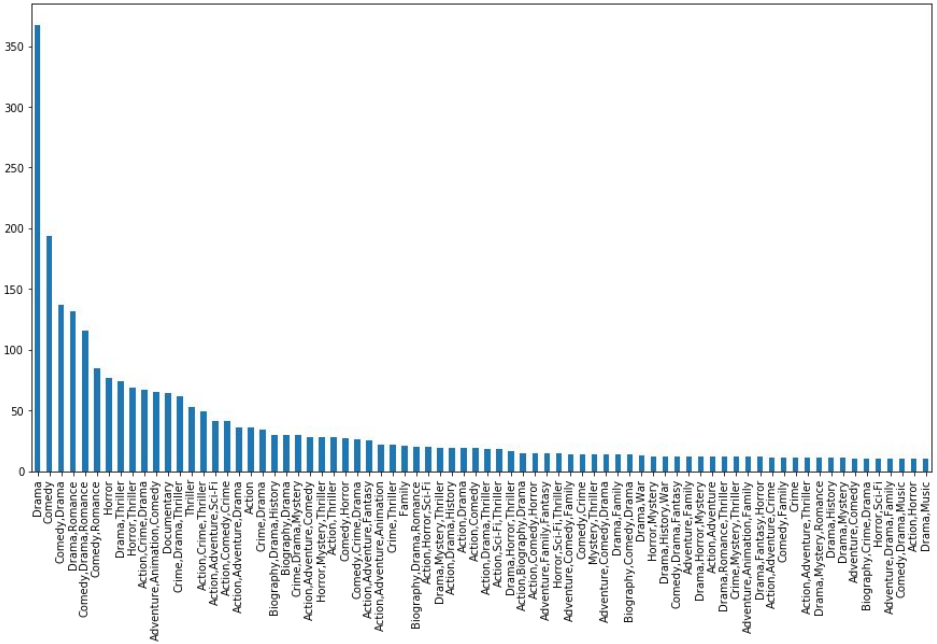


Datensatz **Plots**

Genre Original

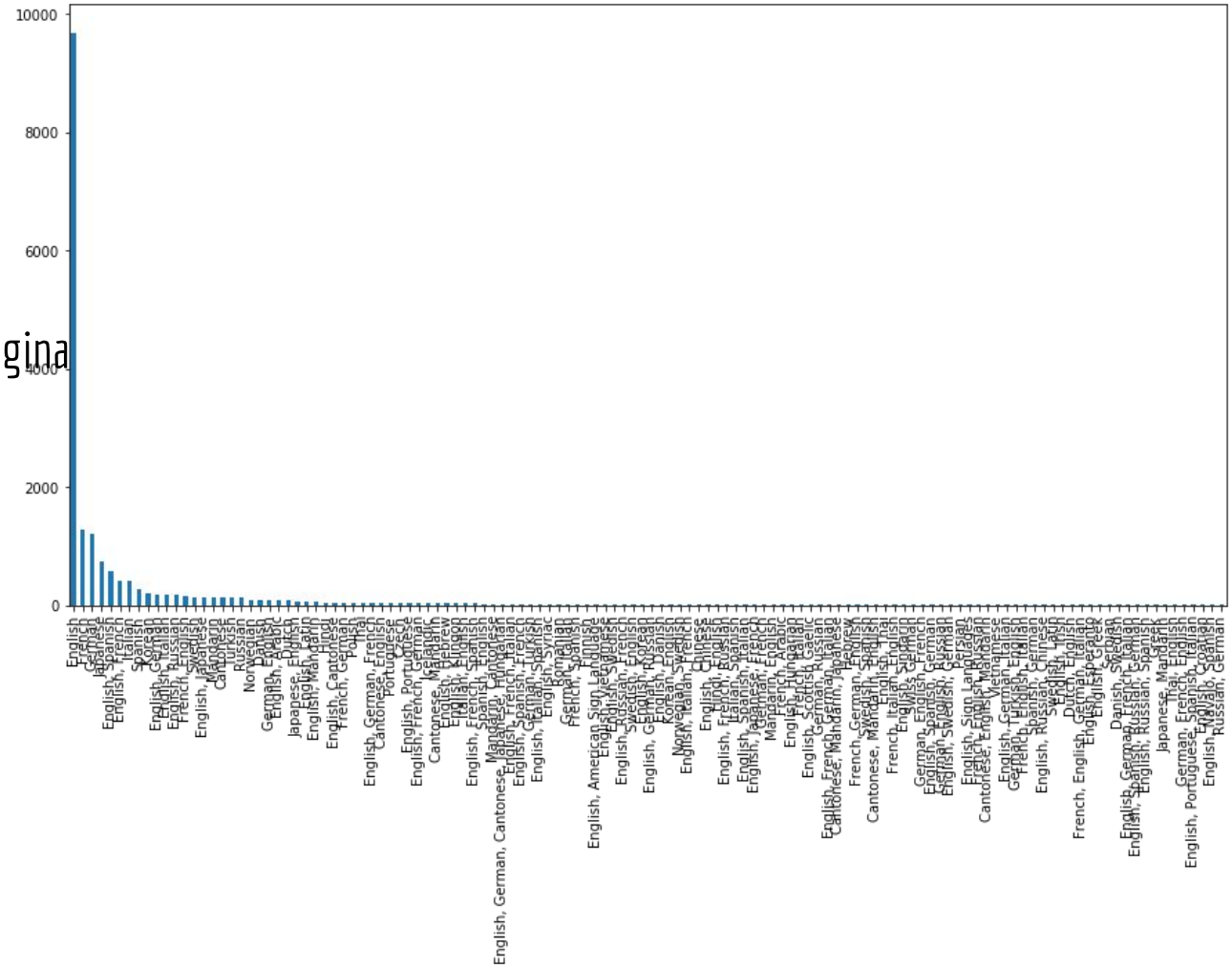


Genre Bereinigt



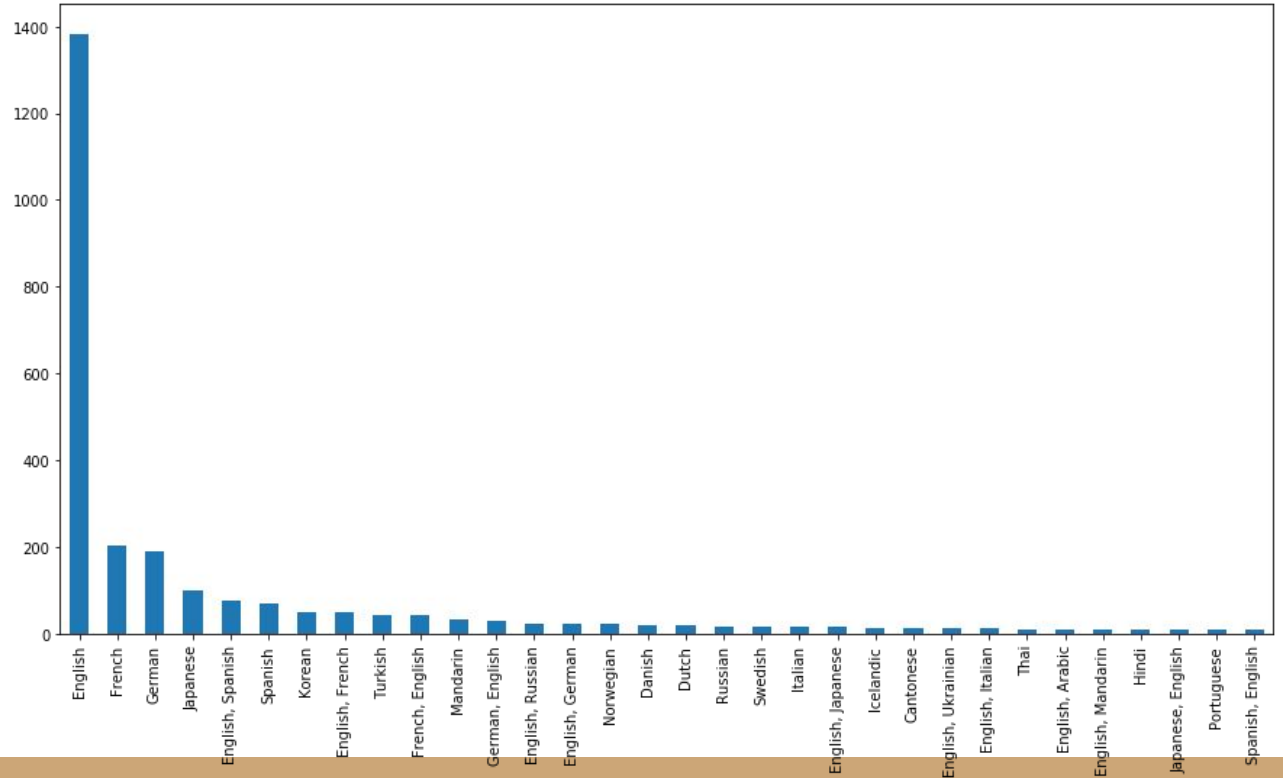
Datensatz **Plots**

Production Region Origina



Datensatz **Plots**

Language Bereinigt



angewendete Clustermethoden

Topic Modelling

- Gensim
- pyLDAvis

---

19 Clusters

100 Iterationen

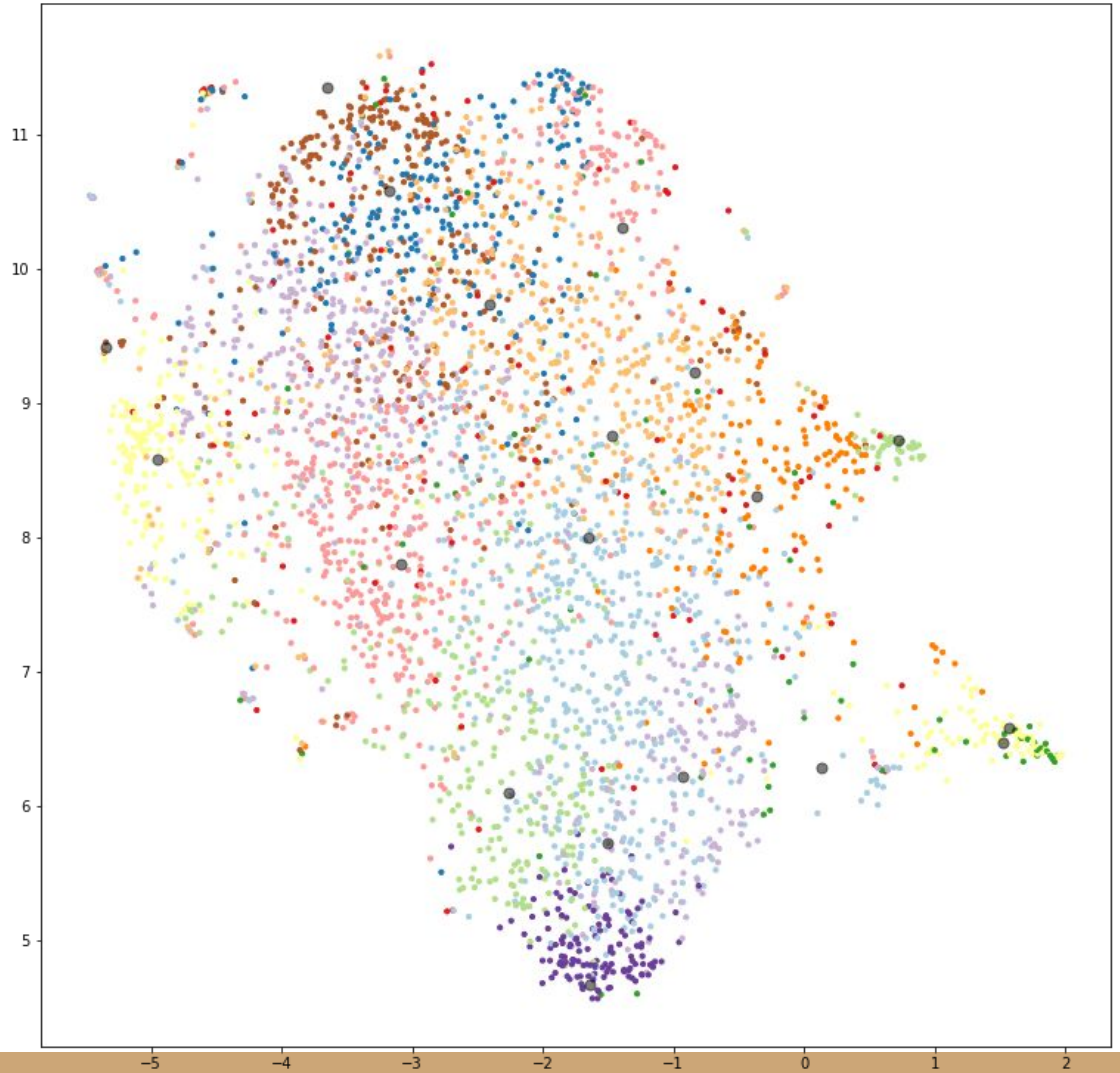
Erste naive Versuche mit **K-Means**:

Genre (Über den kompletten Datensatz)

Anzahl Zentren: **19**,
tf-idf (max_features=5000)

Ergebnisse

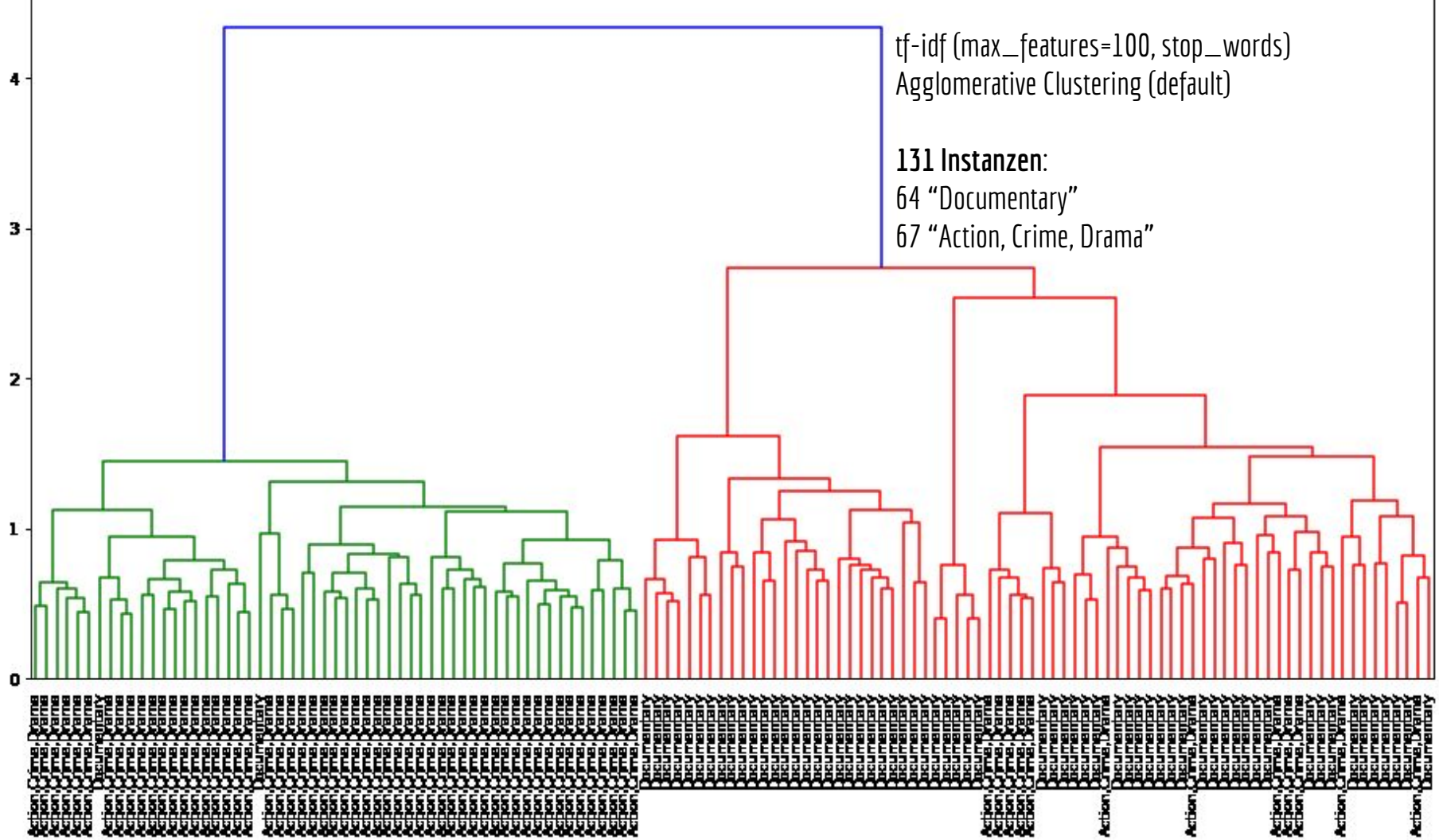
- Daten bilden keine runden Cluster
- zu großer Überlapp zwischen den Clustern, keine scharfen Grenzen
- Datenmenge zu groß um sie sinnvoll zu visualisieren
- Kein Informationsgewinn realisierbar



Mikroperspektive

Vorgehen:

- Reduktion der Datenmenge durch Auswahl kleiner spezifischer Ausschnitte des Datensatzes
- Clusterings nach Genre und der Sprache des Films
- Genaue Analyse hinsichtlich der Zuordnung der Instanzen zum jeweiligen Cluster



tf-idf (max_features=100, stop_words)
Agglomerative Clustering (default)

131 Instanzen:
64 “Documentary”
67 “Action, Crime, Drama”

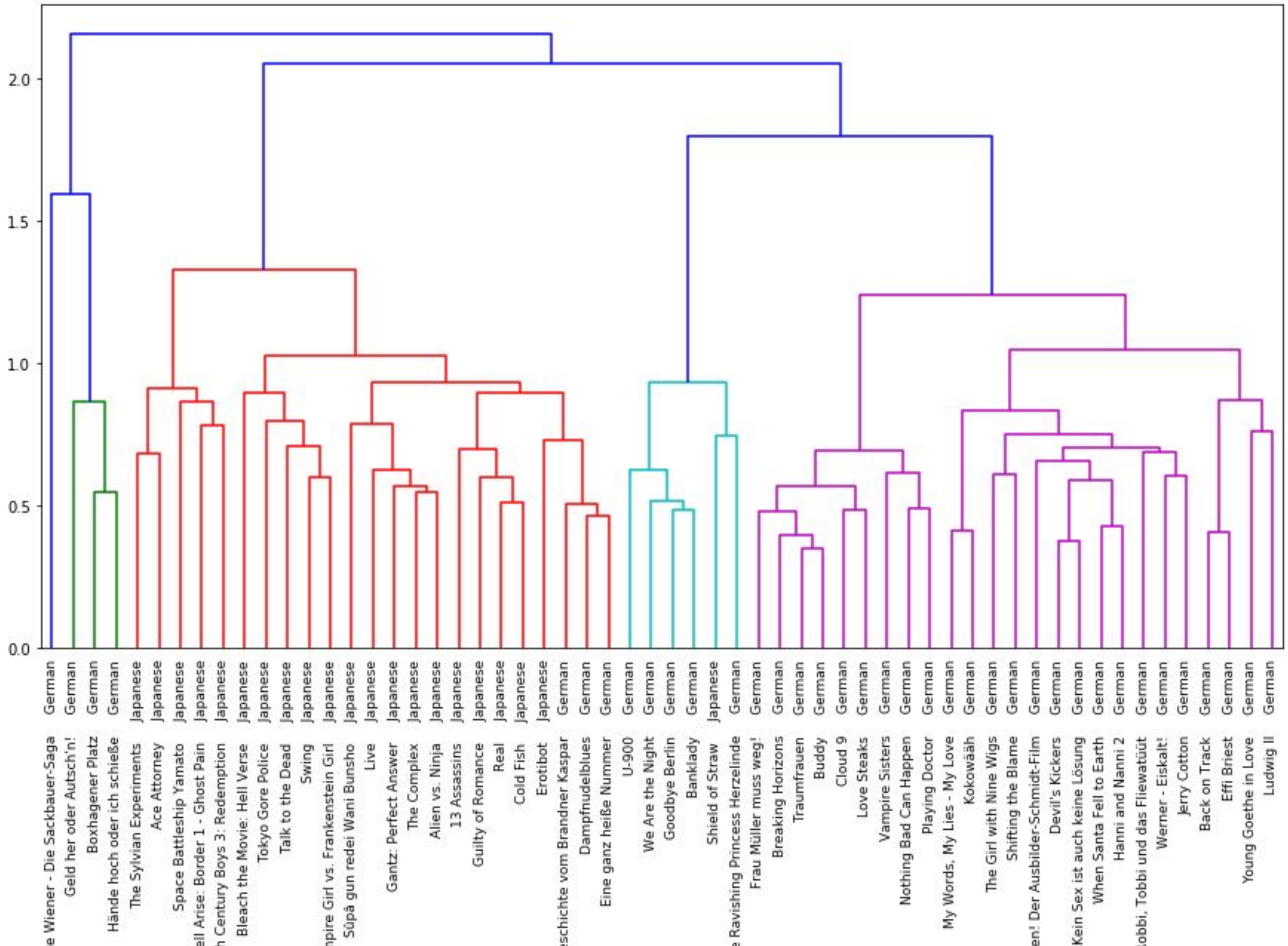
Agglomeratives Cluster: Genre (Documentary vs. Action, Crime, Drama)

20 Häufigsten Wörter im Genre **“Documentary”**

'ja', 1383
'mehr', 1268
'immer', 1067
'gibt', 969
'wurde', 953
'leben', 872
'gut', 871
'menschen', 840
'mal', 773
'leute', 742
'sagte', 717
'geht', 711
'viele', 709
'schon', 650
'sehen', 649
'einfach', 634
'sagen', 618
'weiß', 576
'nein', 568
'wirklich', 550

20 Häufigsten Wörter im Genre **“Action, Crime, Drama”**

'ja', 3096
'hast', 1536
'nein', 1490
'mal', 1380
'schon', 1334
'los', 1326
'gut', 1317
's', 1177)
'hey', 927
'geht', 895
'mann', 879
'wer', 848
'komm', 836
'weiß', 827
'okay', 827
'bitte', 766
'mehr', 726
'scheiße', 649
'gehen', 638
'immer', 635



tf-idf (max_features=100, stop_words)

Agglomerative Clustering (default)

58 Instanzen:
37 “German”
21 “Japanese”

Agglomeratives Cluster: **Language** (German vs. Japanese)

20 Häufigsten Wörter im der Sprache **“German”**

'ja', 2750
'mal', 1490
'hast', 886
's', 870
'schon', 853
'musik', 770
'nein', 747
'gut', 683
'komm', 677
'mann', 672
'na', 574
'geht', 558
'los', 554
'bitte', 540
'mehr', 477
'ganz', 469
'immer', 468
'weiß', 448
'frau', 414
'danke', 410

'ja', 632
'schon', 327
'hast', 324
'mal', 297
's', 294
'gut', 263
'bitte', 246
'nein', 237
'los', 195
'mehr', 191
'freund', 188
'immer', 174
'wer', 170
'welt', 166
'geht', 150
'herr', 150
'wurde', 149
'leben', 138
'weiß', 133
'gehen', 130

20 Häufigsten Wörter im der Sprache **“Japanese”**

Mikroperspektive

Erkenntnis:

- Kleine Datenmengen besser interpretierbar
- Clustering nach Herkunftssprache und Genre möglich
- Rauschen vorhanden, Texte haben keine eindeutige Signale, die den zugeordneten Labeln entsprechen
- “Falsche” Zuordnungen sind erklärbar

Allgemein:

- kürzere Vektoren/weniger Features beeinflussen das Clustering positiv, unabhängig von Stop-Wörtern

...next to be found in another interactive Graphics!