

Workstattbericht:

Roman-Typen als Topic-Modell-Klassen?

Ulrike Henny-Krahmer
(CLiGS, Würzburg)

Kolloquium "Digital Humanities – Aktuelle Forschungsthemen"
Köln, 6. Juli 2017



Überblick

- Kontext: CLiGS
- Fragestellung: Roman-Typen als Topic-Modell-Klassen?
- Hintergrund
 - Literarische Gattungen, Roman-Untergattungen
 - DH-Methoden: Topic Modeling, Klassifikation, Clustering
- Korpus
 - Hispanoamerikanischer Roman im 19. Jahrhundert
 - Roman-Untergattungen
 - Textsammlung
- Analyse
 - Topic Modeling
 - Klassifikation & Clustering
- Fazit

Kontext: CLiGS

CLiGS

- Computergestützte Literarische Gattungsstilistik
- <http://cligs.hypotheses.org>
- Disziplinäre Verortung: zwischen Literaturwissenschaften, Computerlinguistik und Informatik
- Gegenstand: spanischsprachige und französische Literatur
- 17./18. und 19. bis frühes 20. Jahrhundert

CLiGS

Ziele:

- Einzelne Gattungen beschreiben
- Das Konzept "Gattung" neu denken

Fragestellung:

Roman-Typen als Topic-Modell-Klassen?

Fragestellung:

Roman-Typen als Topic-Modell-Klassen?

- Kann man verschiedene Roman-Typen (oder Roman-Untergattungen) anhand ihrer Themen beschreiben?
- Topic Modeling als Verfahren zum automatischen ermitteln von Themen
- Klassen als ein Konzept der Beschreibung von Roman-Untergattungen

Hintergrund

- Literarische Gattungen, Roman-Untergattungen
- DH-Methoden
 - Topic Modeling
 - Klassifikation & Clustering

Literarische Gattungen, Roman-Untergattungen

Es gibt sie!

[Hierarchies](#)[Sets](#)[Genres](#)[Periods](#)[Denominations](#)[Authors](#)[Login/ Registrat](#)

Customized Search

Type something...

Periods

Genres

Select AllSelect None

core

☒ Prayer

☒ Sermon

☒ Treatise

☒ Controversial Treatise

☒ Catechism

☒ Mimetic Catechism

minor

☒ Religious Biography

associated

Denominations

ResetSearch

Corpus of English Religious Prose

The *Corpus of English Religious Prose* is a diachronic, multi-genre corpus which, in its present outline, covers English religious prose from 1150 to the end of the eighteenth century. It is designed to reflect both continuity and change of English writing in one of its most important domains with emphasis on innovation, transformation, and loss in genres. It is suited to meet the needs of both long-term diachronic studies as well as synchronic studies, particularly from a pragmatic, text-linguistic perspective and with a special interest in the history of individual genres.

Literarische Gattungen, Roman-Untergattungen

Es gibt sie!



Titel, Autor, Stichwort, ISBN



Bücher | eBooks | eReader | Hörbuch | Filme | Musik | Spielwaren | Games | Geschenkkarte | Schule | Wohnen | SALE | Reisen

Bücher

Englische Bücher
Fachbücher
Fantasy & Science Fiction
Jugendbücher
Kinderbücher
Kochen & Backen
Krimis & Thriller
Ratgeber
Reise & Abenteuer
Romane & Erzählungen
Schulbücher & Lernhilfen

Aktuell

Bestseller-Autoren
Neuheiten
SPIEGEL-Bestseller
Bestseller Thalia
Taschenbücher
Beliebte Verlage
Top-Bewertung
Vorbesteller
Kalender
Buch des Monats
SALE



Selfies

von Jussi Adler-Olsen

23,00 €

Literarische Gattungen, Roman-Untergattungen

Es gibt sie!

- > Kinderbücher
- > Kochen & Backen
- > Krimis & Thriller
- > Ratgeber
- > Reise & Abenteuer
- **Romane & Erzählungen**
- > Biografien & Tagebücher
- > Drama
- > Historische Romane
- > Klassiker
- > Kurzgeschichten & Anthologien
- > Liebesromane
- > Lyrik
- > Märchen & Legenden
- > Preisgekrönte Romane
- > Romane & Erzählungen
- > Unterhaltung für Frauen
- > Witz & Unterhaltung
- > Nach Autoren
- > Nach Ländern & Kontinenten
- > Sachbücher
- > Schule & Lernen
- Specials



Die Töchter der Tuchvilla
von Anne Jacobs

★★★★★ (19)
Buch (Taschenbuch)
9,99 €



Das Erbe der Tuchvilla
von Anne Jacobs

★★★★★ (21)
Buch (Taschenbuch)
9,99 €



Die Tuchvilla
von Anne Jacobs

★★★★★ (54)
Buch (Taschenbuch)
9,99 €



Die fremde Königin
von Rebecca Gablé

★★★★★ (52)
Buch (gebundene Ausgabe)
26,00 €



Die Wege der Macht
von Jeffrey Archer

★★★★★ (17)
Buch (Klappenbroschur)
9,99 €



Der Flamme
von Bernard

★★★★★
Buch (Tasch)
10,99 €



Literarische Gattungen, Roman-Untergattungen

Aber was sind sie eigentlich?

- Gibt es eine Systematik?
- Kann man Unterschiede an den Texten selbst erkennen?
- Mit welchen Konzepten kann man sie beschreiben?
- Mit welchen Methoden kann man sie untersuchen?

DH-Methoden

Topic-Modeling

Was ist Topic Modeling?

- Topic Modeling ist eine Methode der quantitativen Textanalyse
- Ziel: Aufdecken "versteckter" semantischer Strukturen
 - Hintergrund: Distributionelle Semantik
 - J. R. Firth: "You shall know a word by the company it keeps" (1957)
 - Wiederkehrende Themen / Motive / Diskurse werden identifiziert
- Wichtig: automatisch & ohne explizites semantisches Wissen!

Wie funktioniert Topic Modeling?

- gegeben: Dokumente & Wörter
- gesucht: versteckte "Topics"
- Statistik:
 - Welche Wörter kommen innerhalb von Dokumenten und zwischen Dokumenten gemeinsam vor?
 - Ableiten von Topics aus den Wortverteilungen
- Ergebnis: ein Topic-Modell

Topic-Modell: Topics mit Wörtern

Topics

gene	0.04
dna	0.02
genetic	0.01
...	

life	0.02
evolve	0.01
organism	0.01
...	

brain	0.04
neuron	0.02
nerve	0.01
...	

data	0.02
number	0.02
computer	0.01
...	

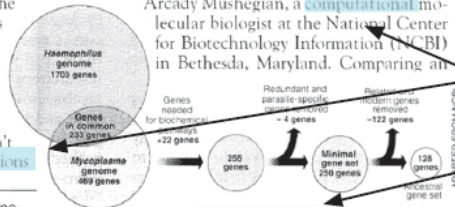
Documents

Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many **genes** does an **organism** need to **survive**? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for **life**. One research team, using **computer** analyses to compare known **genomes**, concluded that today's **organisms** can be sustained with just 250 genes, and that the earliest life forms required a mere 128 **genes**. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those **predictions**

"are not all that far apart," especially in comparison to the 75,000 **genes** in the human genome, notes Siv Andersson, a parasitologist at Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a **genetic** numbers game, particularly as more and more **genomes** are completely mapped and sequenced. "It may be a way of organizing any newly **sequenced genome**," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

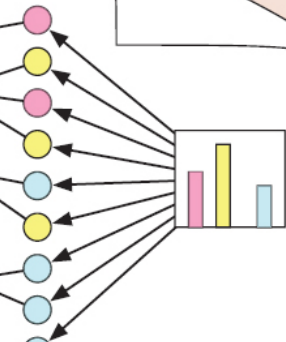


* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

SCIENCE • VOL. 272 • 24 MAY 1996

Topic proportions and assignments







David M. Blei (2012): Probabilistic Topic Models. Communications of the ACM, Vol. 55, No. 4, p. 78.

Topic-Modell: Topics mit Wörtern

0	katholisch religion könig katholik kaiser protestantisch handeln niederland provinz protestantismus
6	erziehung erzieher tugend mensch ideell zögling gesellschaft abhandlung historisch inner
7	grenze nationalität heutig sprachgrenze südlich süddeutsch gebiet inner muttersprache einheitsstaat
14	bahn bahnhof konzession bahnbau projekt geplant annullierung britisch anwenden ausdrücklich
27	virtuose pianoforte talent vater virtuos knabe welt klavierlehrer portrait instrument

Tools

Name			Developer	Language	Link
MALLET	<i>machine learning for language toolkit</i>		David Mimno	Java	http://mallet.cs.umass.edu/topics.php
Gensim	<i>topic modeling for humans</i>		Radim Řehůřek	Python	https://radimrehurek.com/gensim
tmw	<i>topic modeling workflow</i>		Christof Schöch	Python	https://github.com/cligs/tmw
dfr- browser	<i>a simple topic- model browser</i>		Andrew Goldstone	JavaScript	http://agoldst.github.io/dfr-browser/

Klassifikation & Clustering

Klassifikation & Clustering

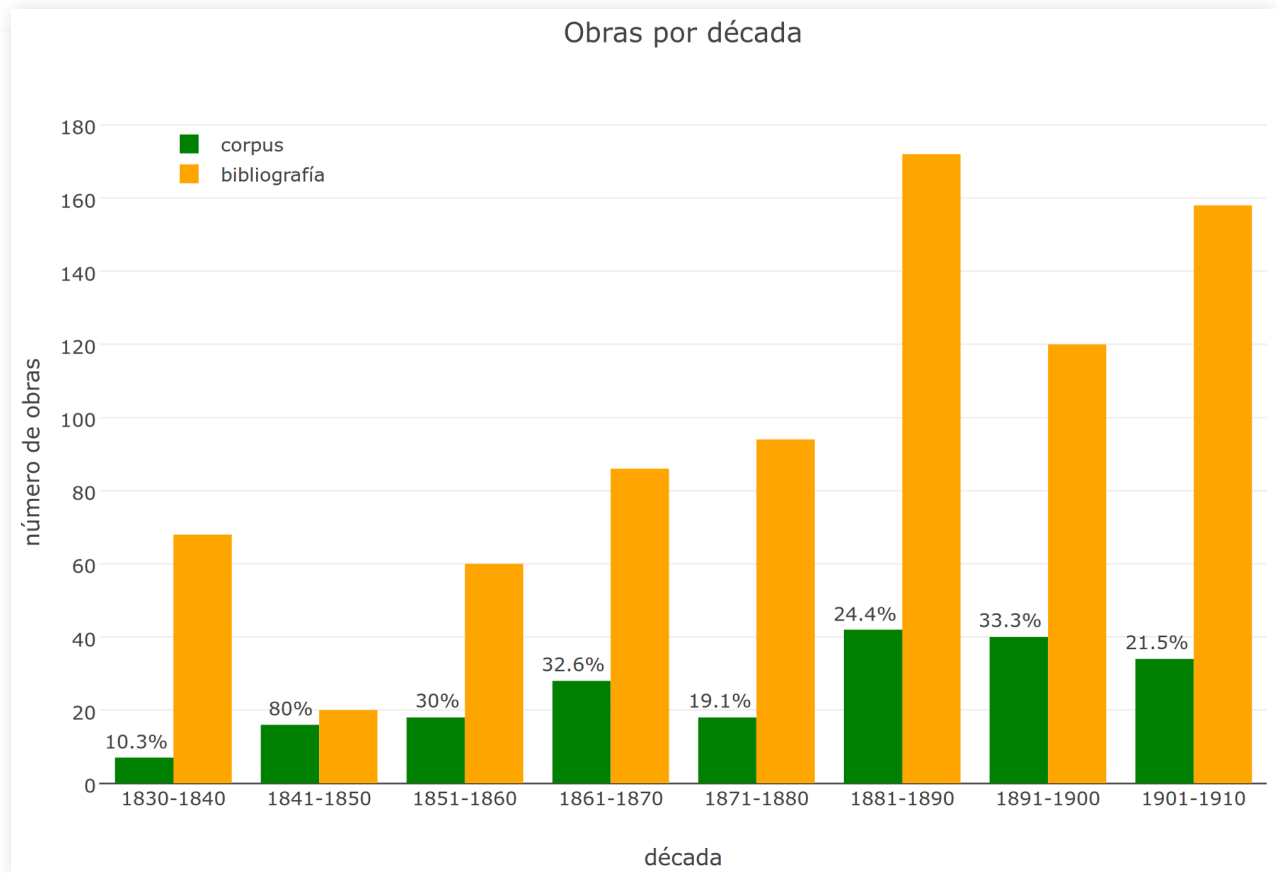
- Lernverfahren
- Automatische Ermittlung von Textgruppen
- Klassifikation:
überwacht; die Gruppen stehen fest (was gehört in welche Gruppe?)
- Clustering:
unüberwacht; die Gruppen stehen nicht fest (was passt zusammen?)
- Eine ganze Reihe verschiedener Algorithmen (SVM, Decision Trees, k-means, ...)

Korpus

Korpus

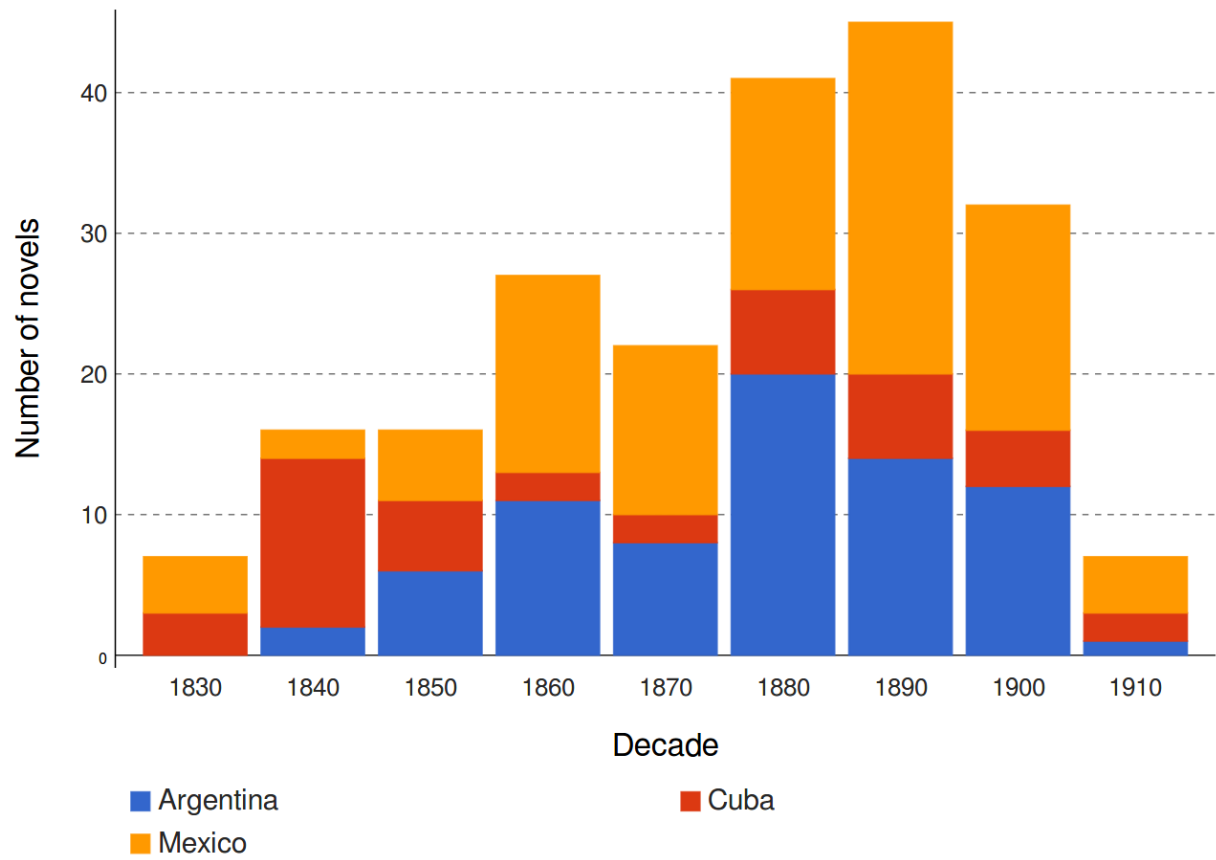
- Hispanoamerikanischer Roman im 19. Jahrhundert
- Roman-Untergattungen
- Textsammlung: rund 200 Romane aus Mexiko, Argentinien, Kuba

Korpus



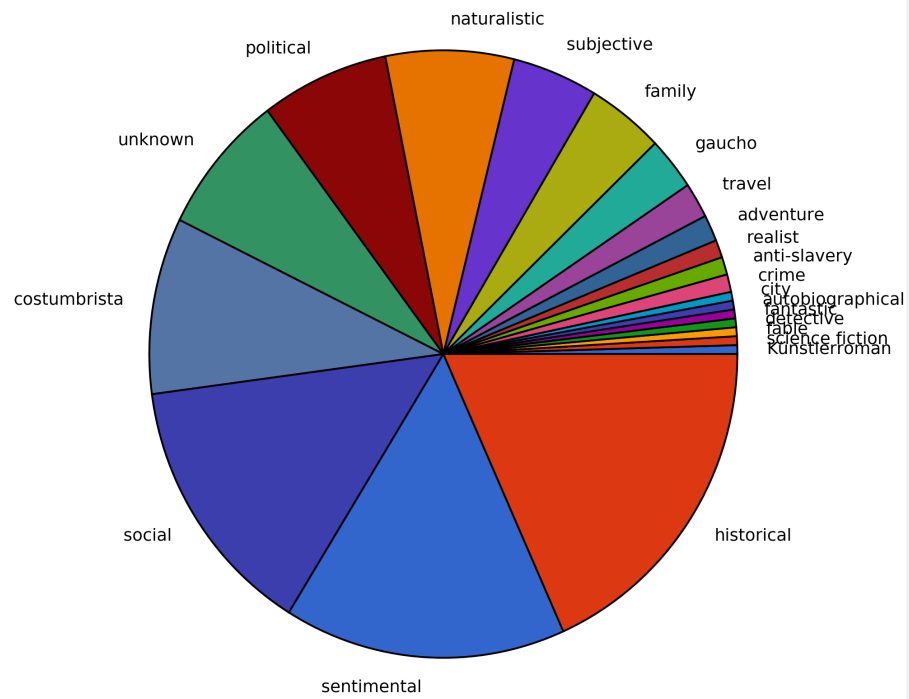
Korpus

Distribution of novels



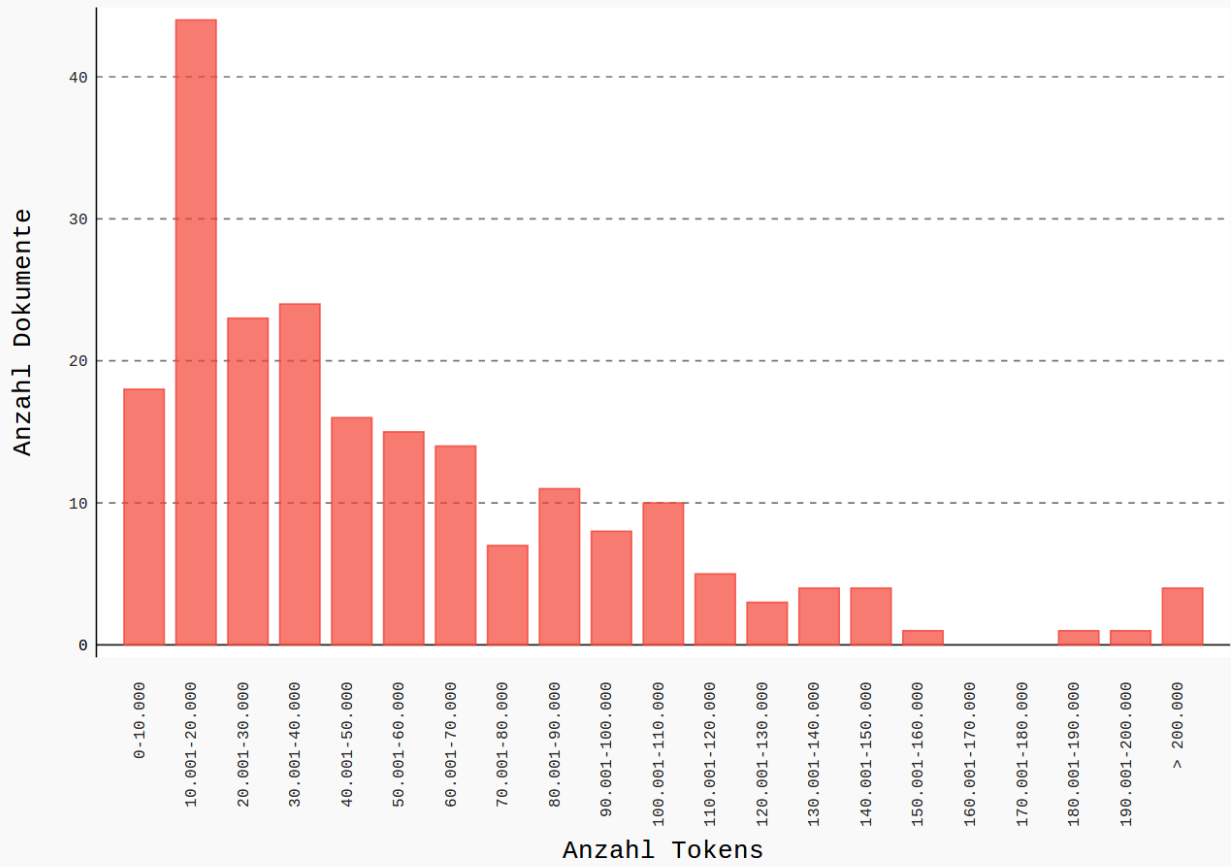
Korpus

Distribution of novels



Korpus

Dokumentlängen hispanoamerikanische Romane



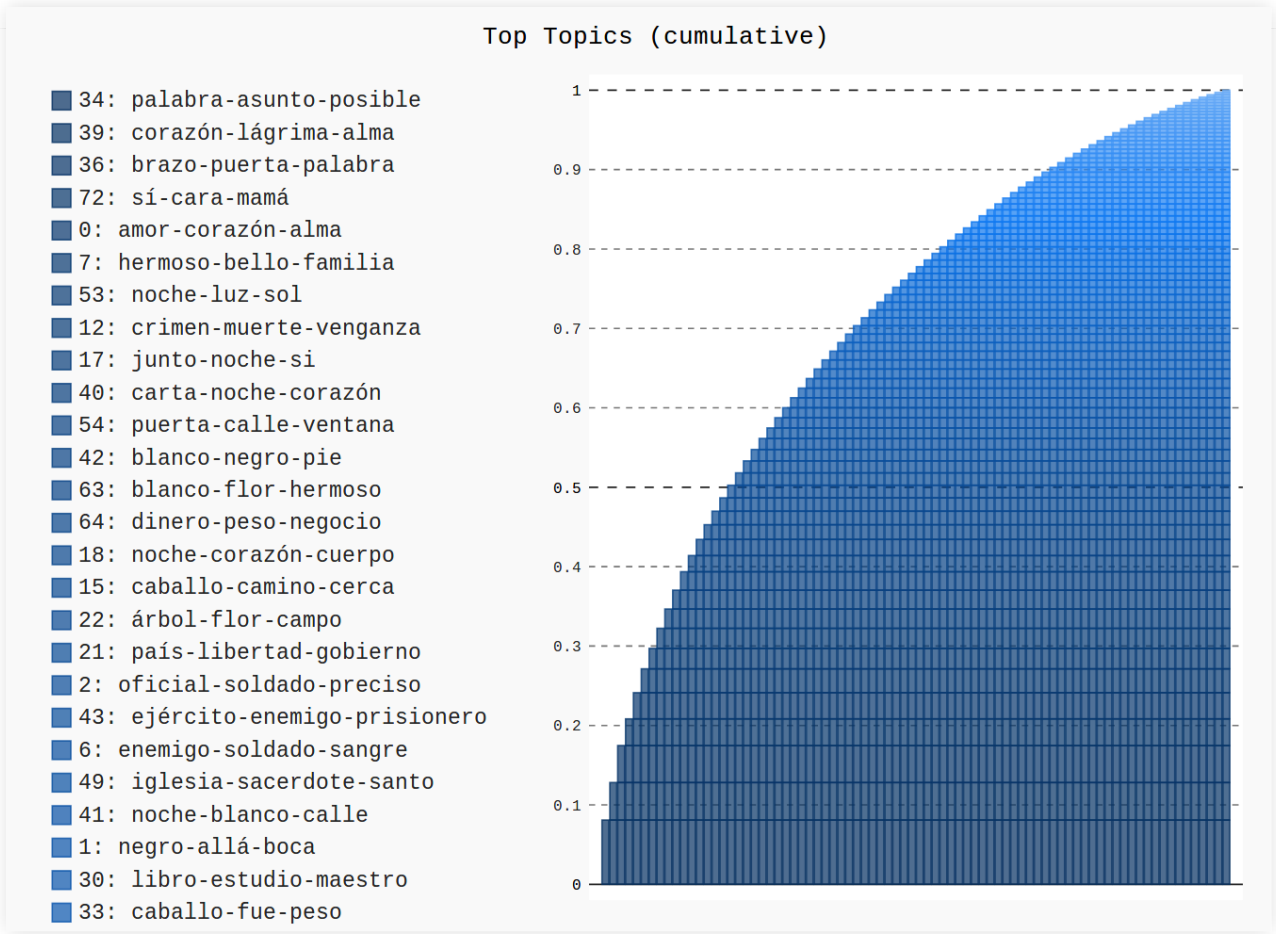
Analyse

- Topic Modeling
- Klassifikation
- Clustering

Topic Modeling

- Vorverarbeitung: Lemmatisierung, Segmentierung
- Tool: Mallet
- Anzahl Topics: 80
- Auswertung mit TMW und eigenen Skripten

Topic Modeling



Topic Modeling

topic 34 (1/80)

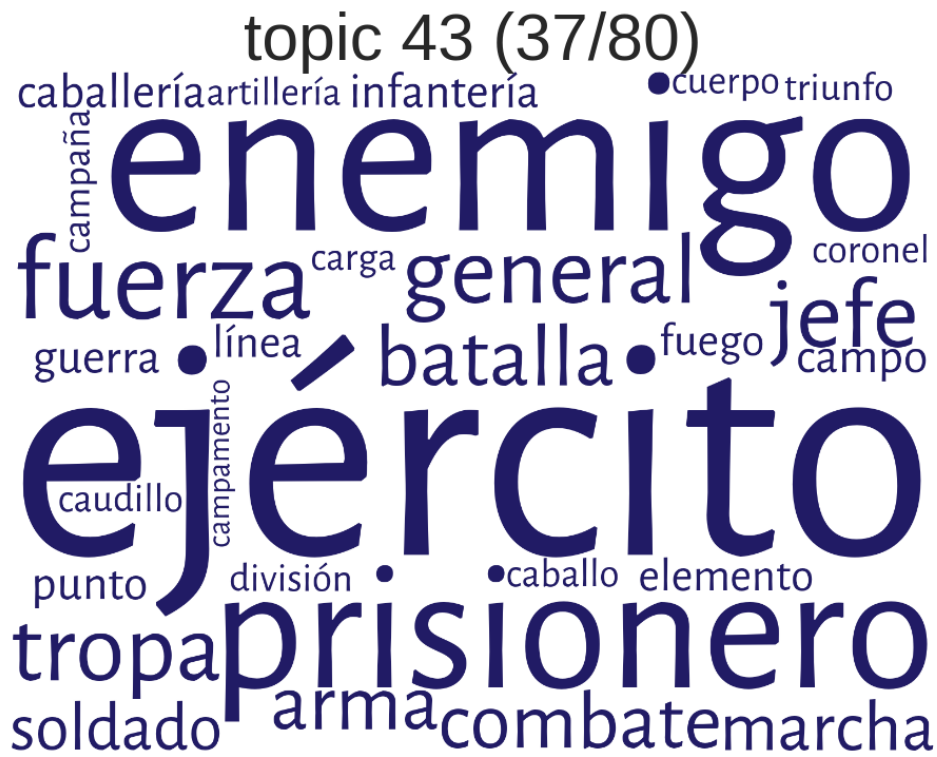
A word cloud visualization for topic 34 (1/80). The words are in Spanish. The most prominent words, shown in the largest font, are 'palabra', 'razón', 'asunto', and 'ideapossible'. Other visible words include 'seguro', 'interés', 'efecto', 'situación', 'carácter', 'atención', 'motivo', 'punto', 'preciso', 'familia', 'cuenta', 'camino', 'confianza', 'lugar', 'claro', 'cierto', 'deseo', 'duda', 'vista', 'bastante', 'hecho', 'creo', 'grave', 'secreto', and 'noche'.

ideapossible
seguro
deseo
palabra
asunto
motivo
punto
preciso
familia
cuenta
camino
confianza
lugar
claro
cierto
carácter
atención
situación
efecto
interés
seguro
vista
bastante
hecho
creo
grave
secreto
noche

Topic Modeling



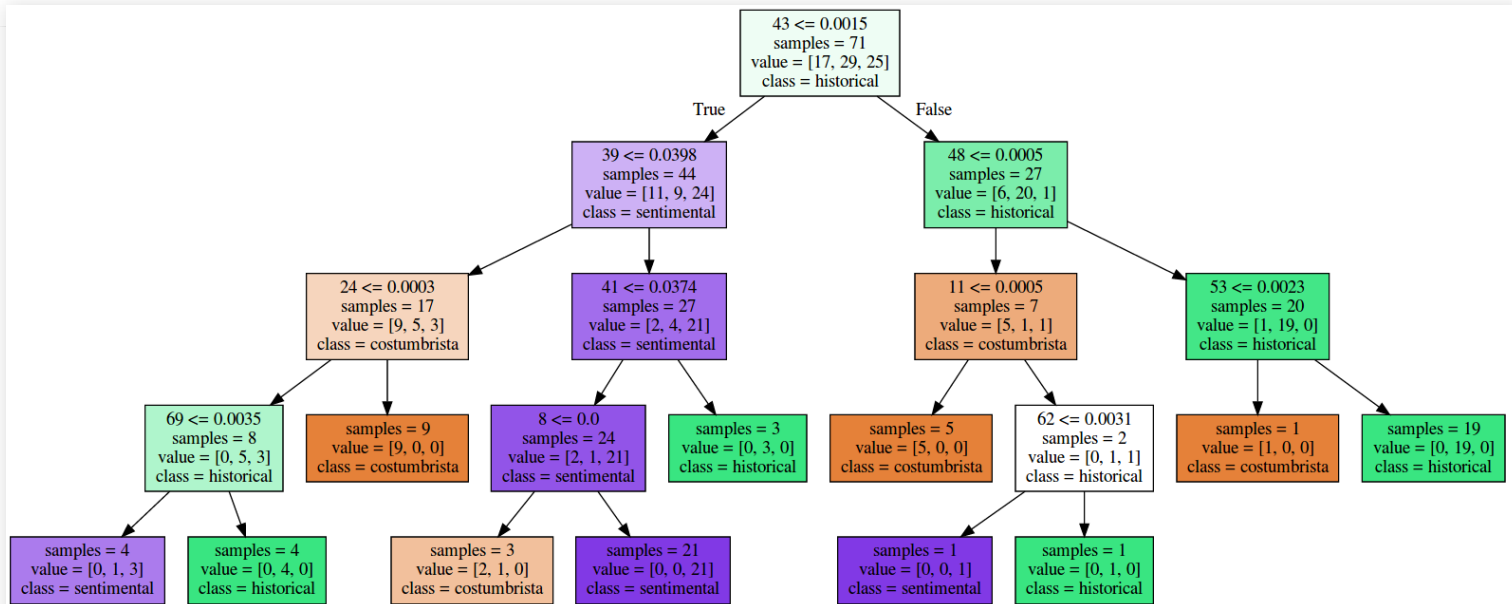
Topic Modeling



Klassifikation

- hier verwendet: Decision Trees
- mit Python und der Bibliothek *sklearn*

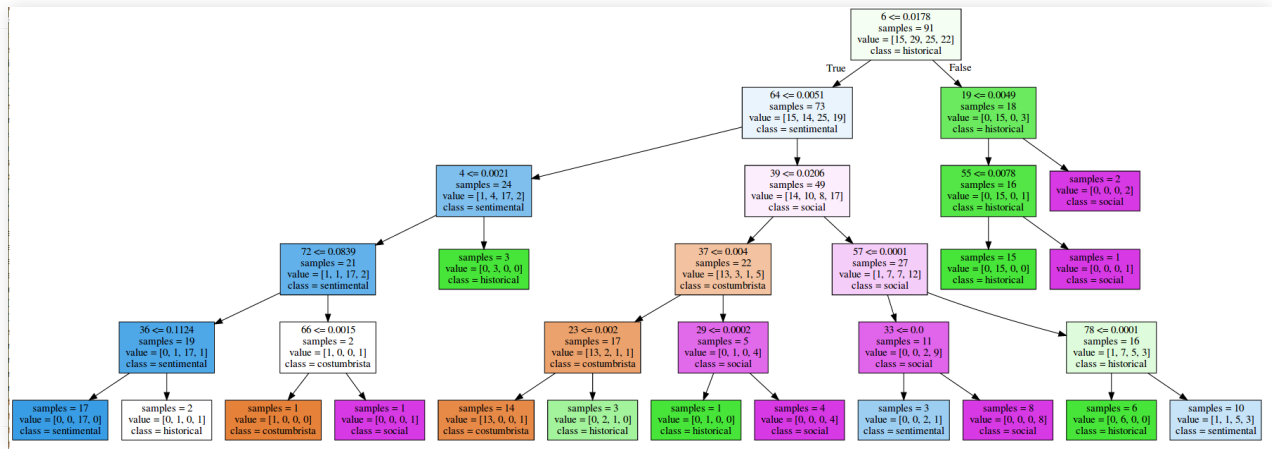
Klassifikation



Klassifikation

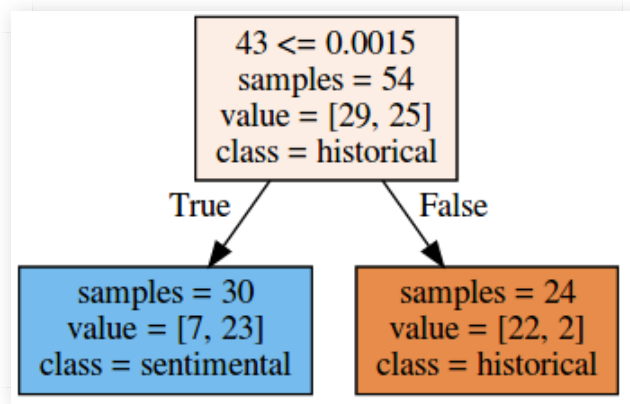


Klassifikation



Klassifikation

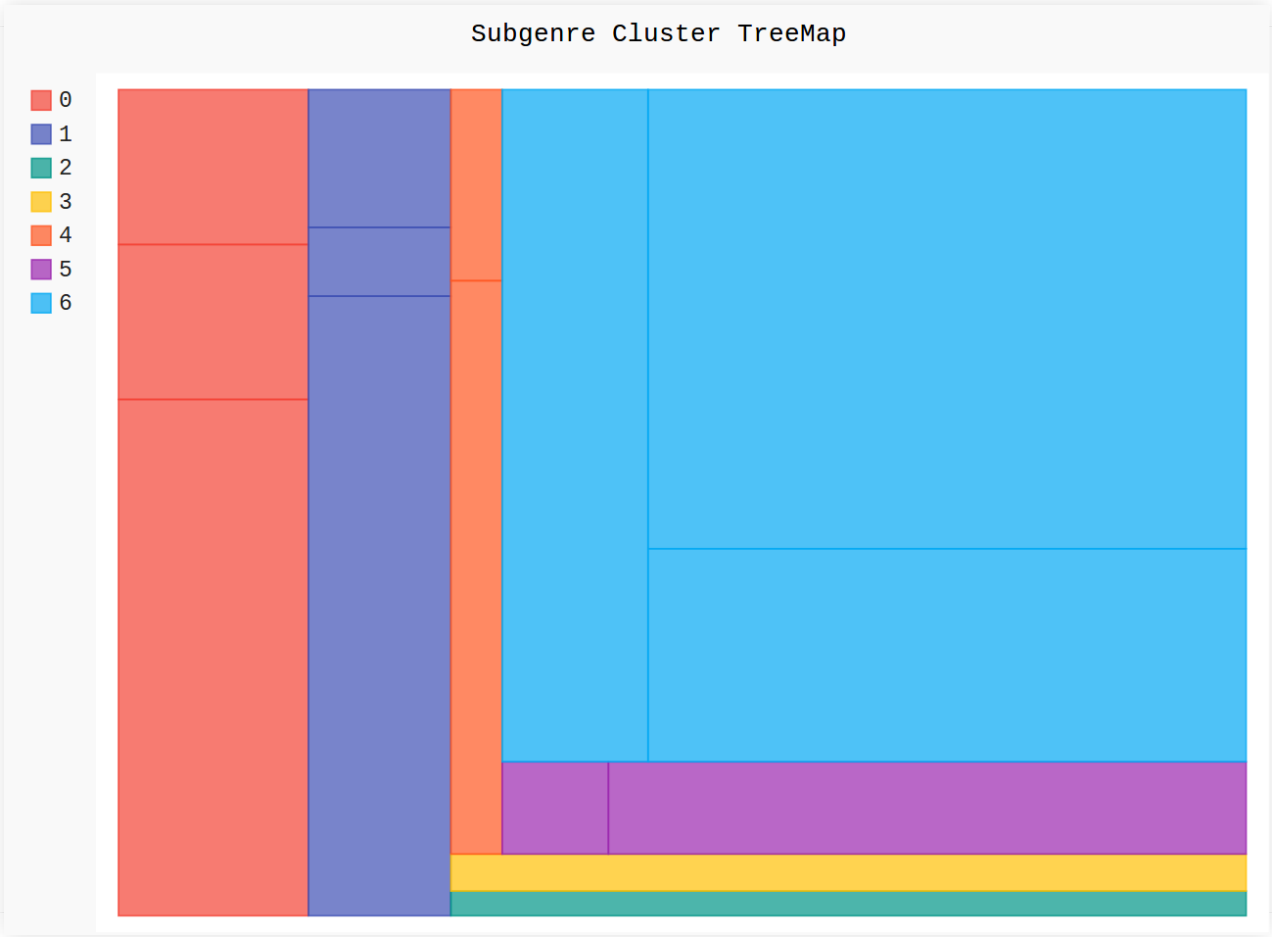
Klassen: historical, sentimental; opt. Baumtiefe: 1; Accuracy training: 0.833; Accuracy test: 0.833



Clustering

- hier verwendet: k-means
- mit Python und der Bibliothek *sklearn*

Clustering



Fazit

Fazit

- Mit Hilfe von Topic-Modeling lassen sich verschiedene Roman-Typen ermitteln.
- Geht man von bekannten Roman-Untergattungen aus, lassen sich manche von ihnen gut als Klassen beschreiben, z.B. historische Romane vs. sentimentale Romane.
- Bei anderen ist es viel schwieriger, z.B. bei kostumbristischen Romane und Gesellschaftsromanen.
- Verwendet man statt (geschlossenen) Klassen (offene) Cluster, ist die Zuordnung immer noch schwierig.

Fazit

- Das könnte bedeuten, dass die Zuordnung einzelner Romane zu Untergattungen überprüft werden muss.
- Evtl. spielen zumindest bestimmte Untergattungen eine geringere Rolle als angenommen.
- Oder es gibt zwar Text-Typen, die näher untersucht werden können, die aber nicht unbedingt mit Untergattungen gleichzusetzen sind.

Vielen Dank!