

# Topic Modeling Tutorial

---

Workshop Digitale Forschungsmethoden in der Romanistik  
Forum Junge Romanistik 2016, Universität Würzburg

---

Christof Schöch  
(CLiGS, Universität Würzburg)



# Überblick

1. Theorie: Was ist Topic Modeling?
2. Anwendung: Der Topic Modeling Workflow (tmw)
3. Ergebnisse: Visualisierungen der Topics
4. Ergebnisse: Visualisierungen der Verteilungen
5. Weiterführendes

# 1. Was ist Topic Modeling?

# Topic Modeling

## Grundidee

- Entdeckt Wörter, die immer wieder gemeinsam bzw. in ähnlichen Kontexten vorkommen (Topics)
- Berechnet, wie wichtig jeder Topic in jedem Dokument ist
- Wichtig: keinerlei explizites semantisches Wissen fließt ein  
Etwas technischer
- Ein Topic ist eine Verteilung von Wahrscheinlichkeiten von Wörtern
- Ein Dokument ist eine Verteilung von Wahrscheinlichkeiten von Topics

# Wörter, Topics, Dokumente

*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

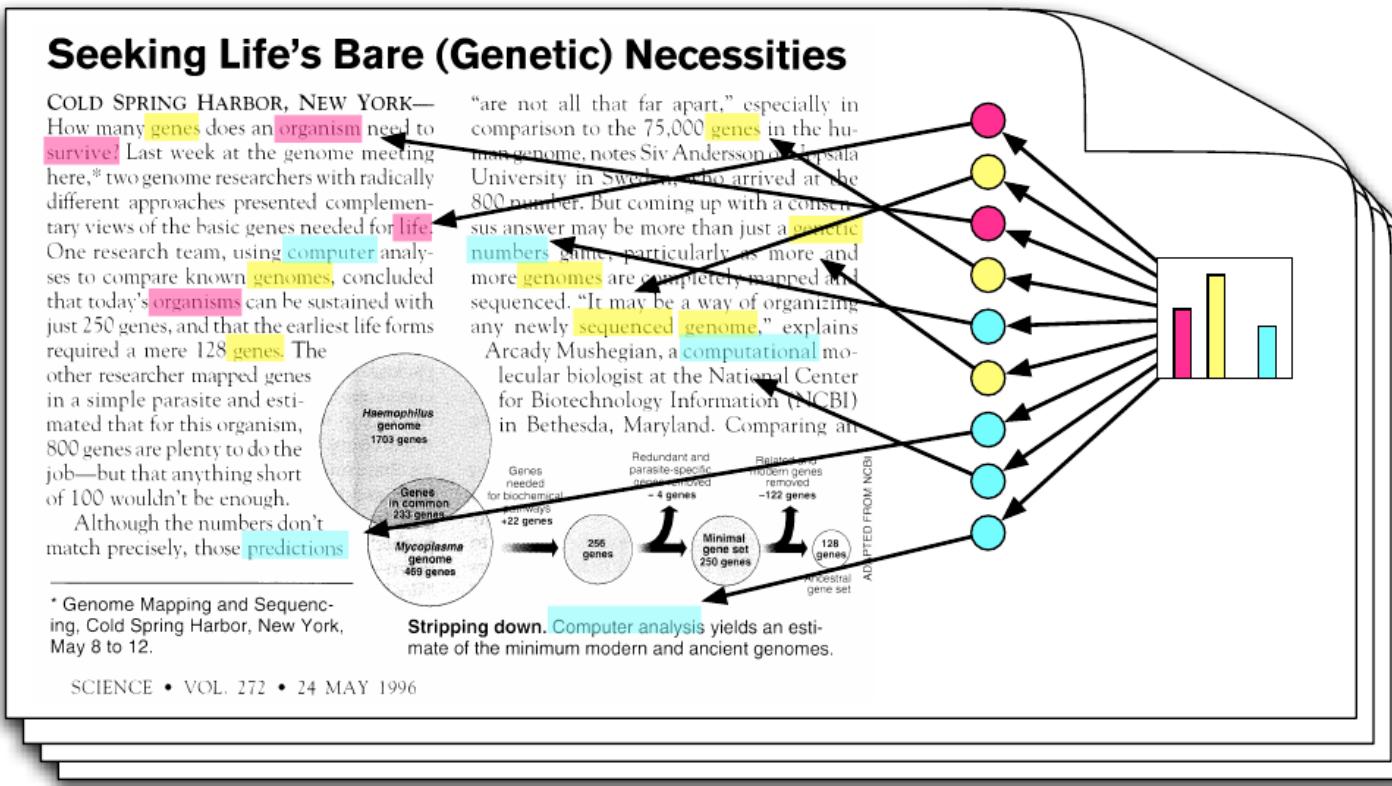
COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,\* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known 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

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

SCIENCE • VOL. 272 • 24 MAY 1996

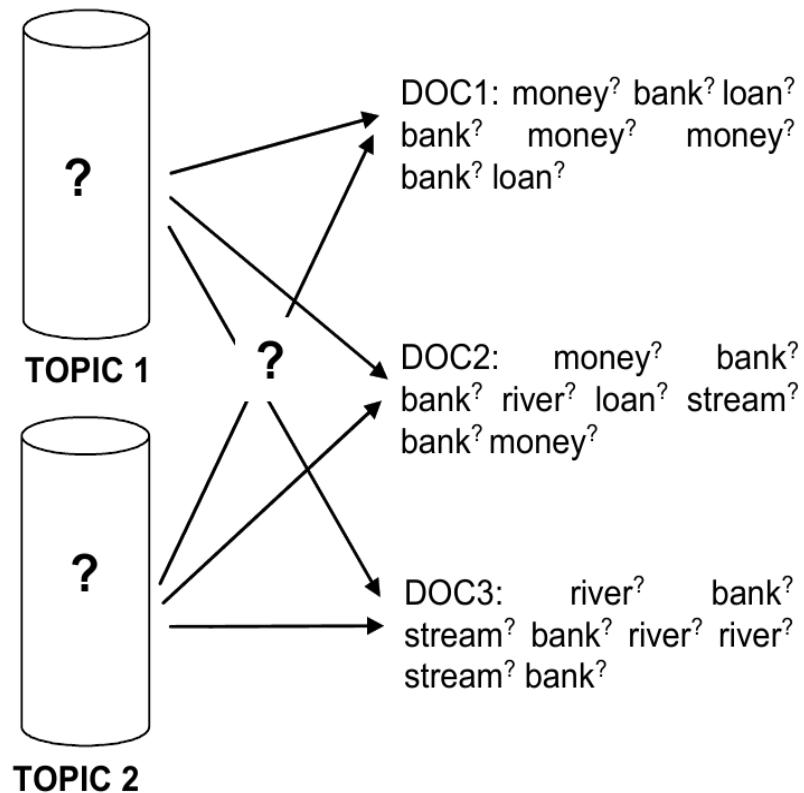
*Topic proportions and assignments*



(David Blei, "Probabilistic Topic Models", 2012)

# Generativ, Iterativ

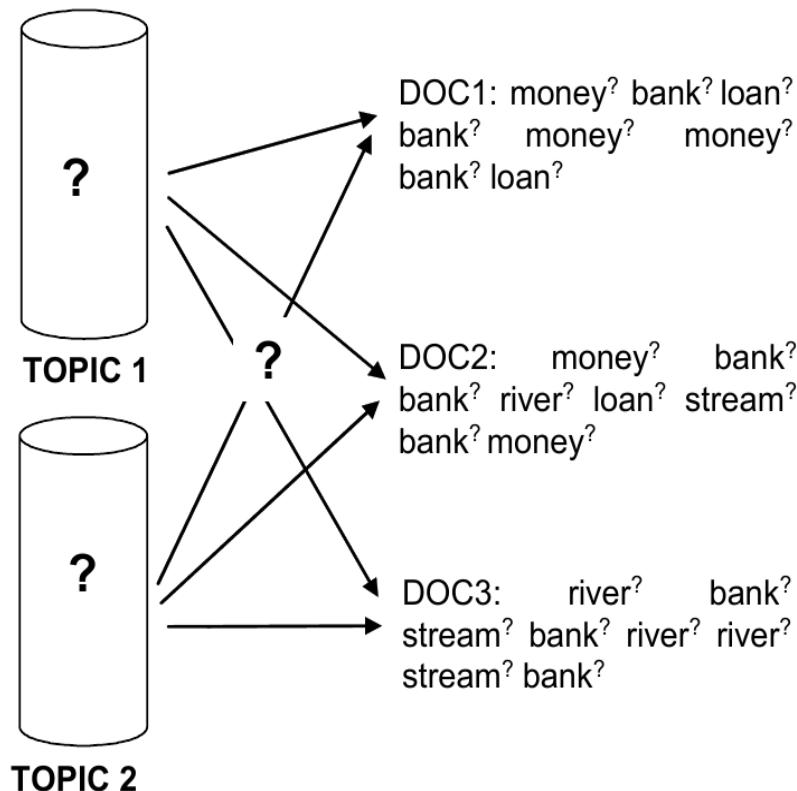
## STATISTICAL INFERENCE



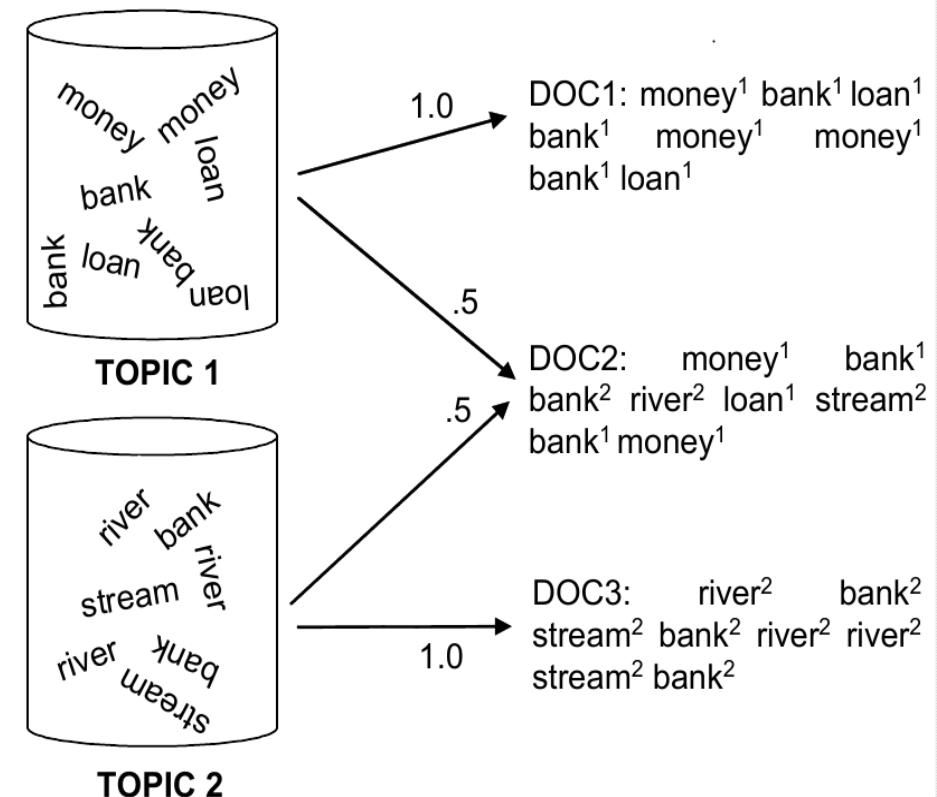
(Steyvers and Griffiths, "Probabilistic Topic Modeling", 2006)

# Generativ, Iterativ

STATISTICAL INFERENCE



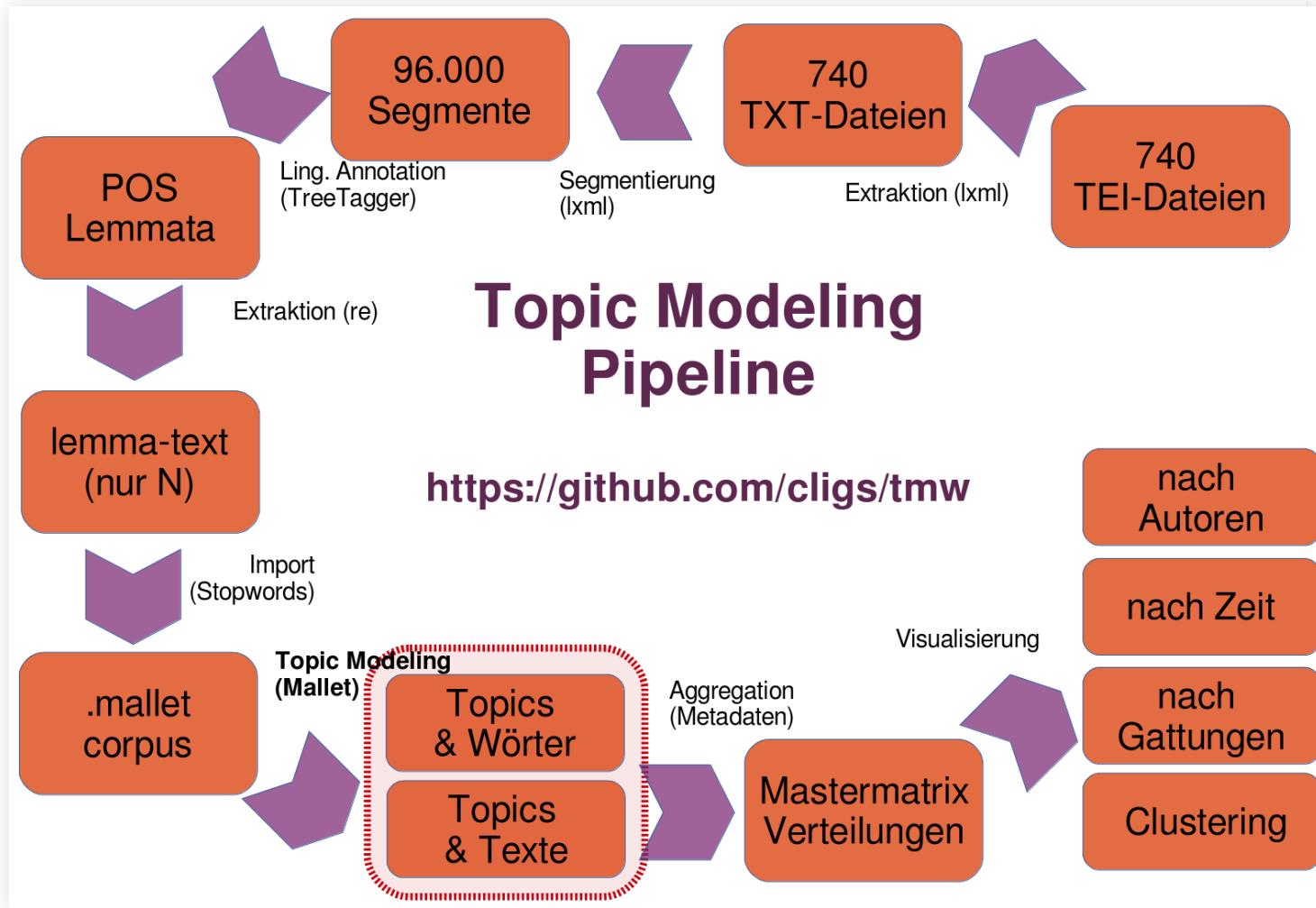
PROBABILISTIC GENERATIVE PROCESS



(Steyvers and Griffiths, "Probabilistic Topic Modeling", 2006)

# 2. Anwendung: Topic Modeling Workflow

# Preprocessing, Modeling, Postprocessing

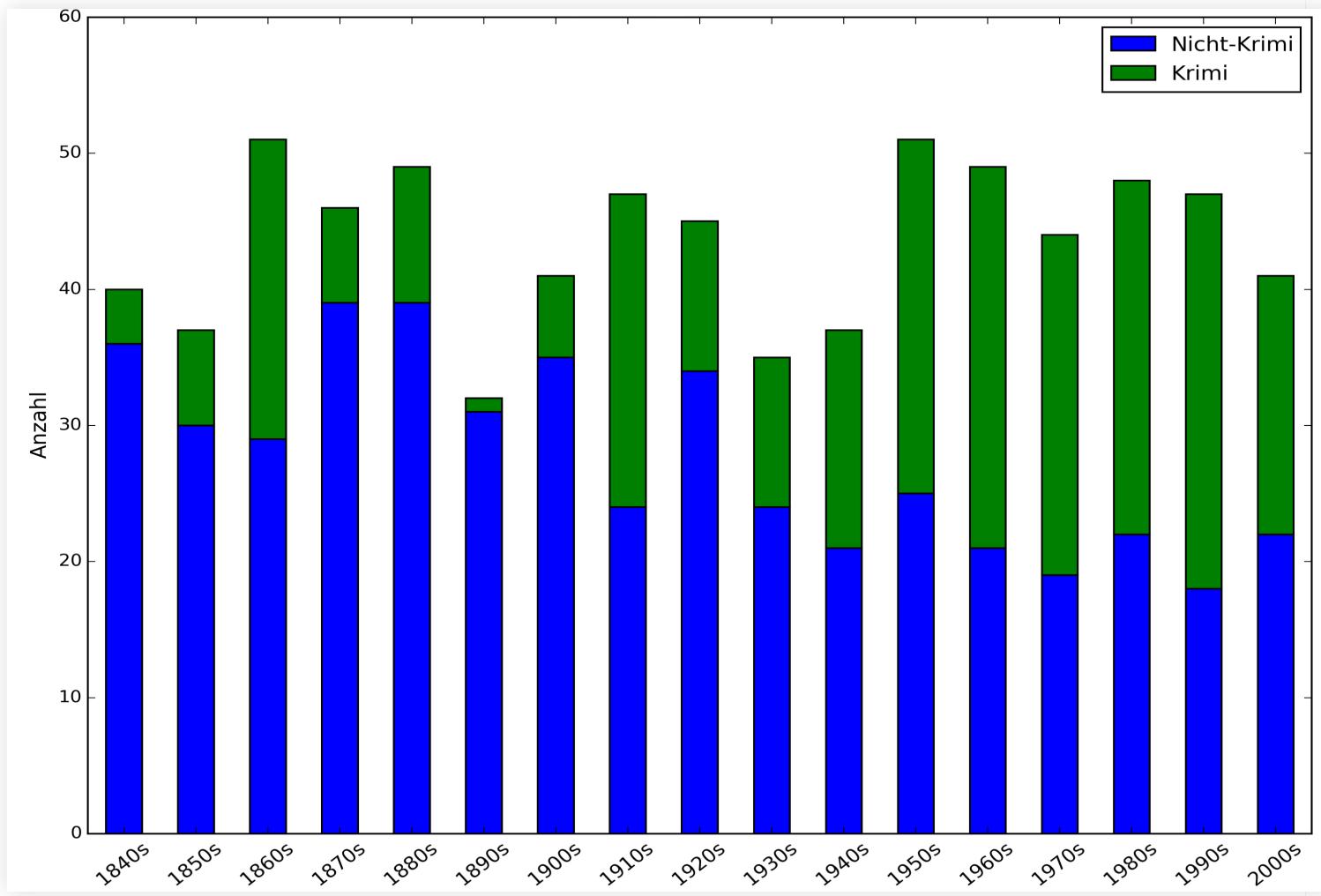


# Umsetzung: Mallet und Python

- MALLET = Machine Learning for Language Toolkit,  
<https://github.com/mimno/Mallet>
- Python = Programmiersprache, <https://www.python.org/>
- tmw = Topic Modeling Workflow, <https://github.com/cligs/tmw>
- tmw = modulare Funktionen + Konfigurationsdatei

# 3. Ergebnisse: Visualisierung der Topics

# Textsammlung: 840 frz. Romane



# Ein Topic als Wordcloud



# Geld-Münze-Franc

topic 28

poche pain louis besoin bourse richesse  
prison pays folie idée mois prix cœur  
paupr<sup>e</sup>re enfant terre aumône fortune  
franc écu travail mendiant trésor  
économie voleur misère pièce monnaie livre

**argent**

# Professor-Wissenschaft-Weiser

# topic 11

# Verbrechen-Tod-Mord

topic 2

justice voleur  
vérité vengeance silence

coupleable nuit  
cadavre innocence témoin criminel

police accusation

vol secret

meurtrier

juge soupçon

drame trace

assassin mort preuve

complice raison assassinat prison victime

# Musik-Piano-Lied

# topic 48

clavier oreille œuvre chœur  
chanson piano  
concert note chanteur instrument  
mélodie leçon mesure disque  
guitare corde refrain  
phrase doigt cœur  
danse morceau flûte rythme  
orchestre accord artiste

# Buch-Bibliothek-Seite

topic 41



# Auto-Straße-Fahrer

topic 61

camion • phare chauffeur  
visage trottoir policier camionnette automobile  
voiture virage conducteur direction pneu garage vitesse arrière chemin  
volant route auto feu  
moteur siège véhicule vitre portière minute kilomètre  
marche roue

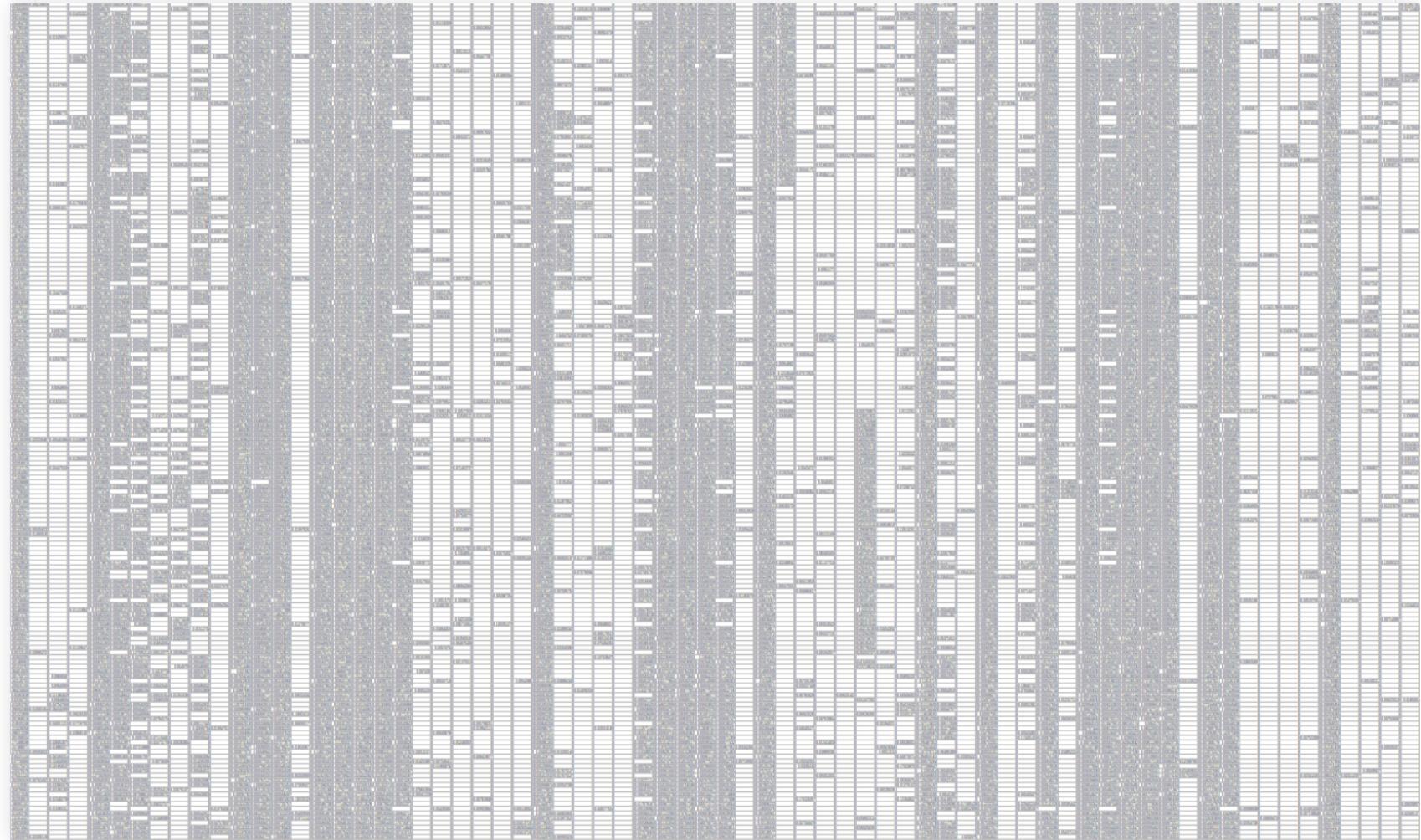
# Tier-Tier-Löwe

topic 75

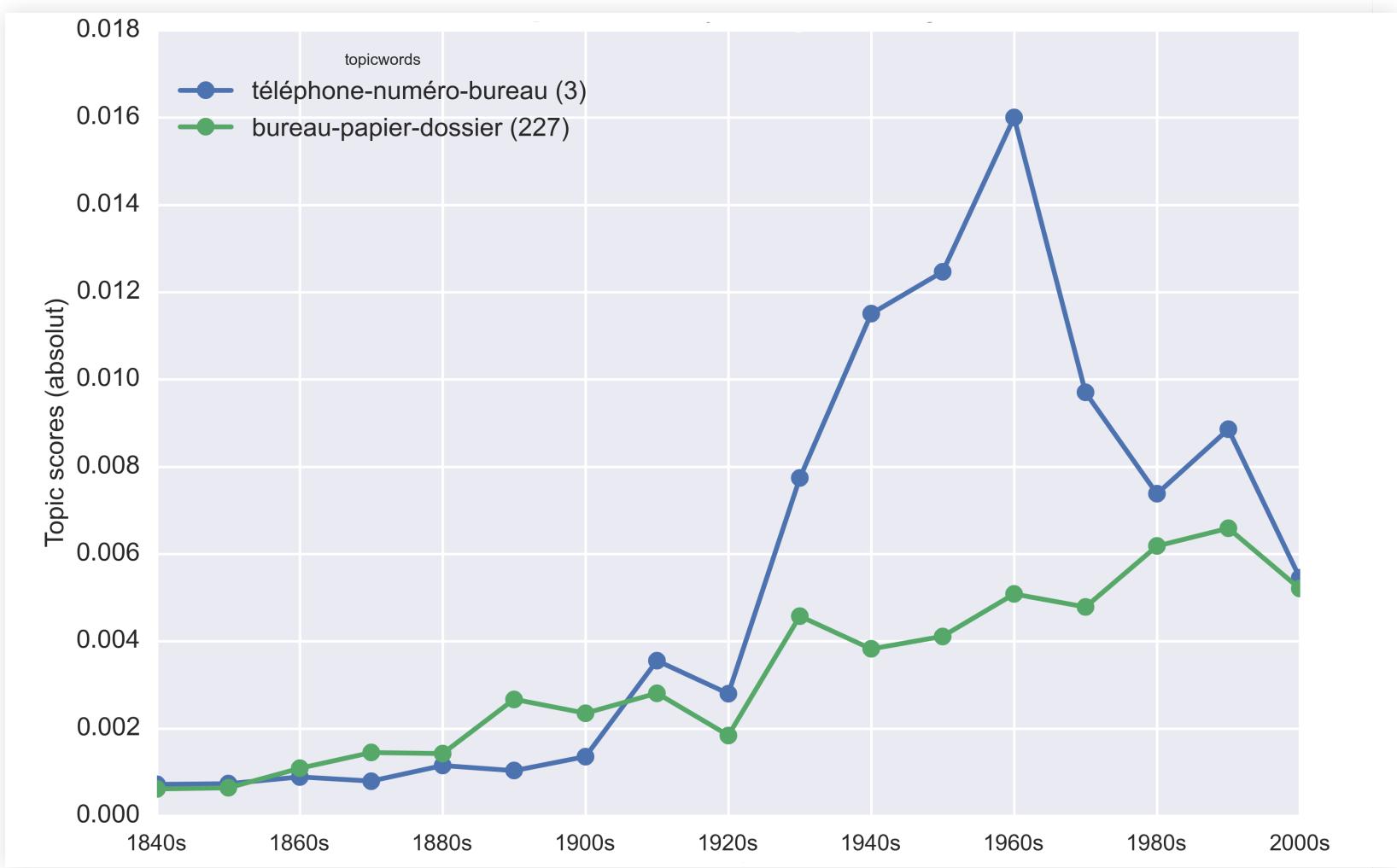


# 4. Ergebnisse: Visualisierung der Verteilungen

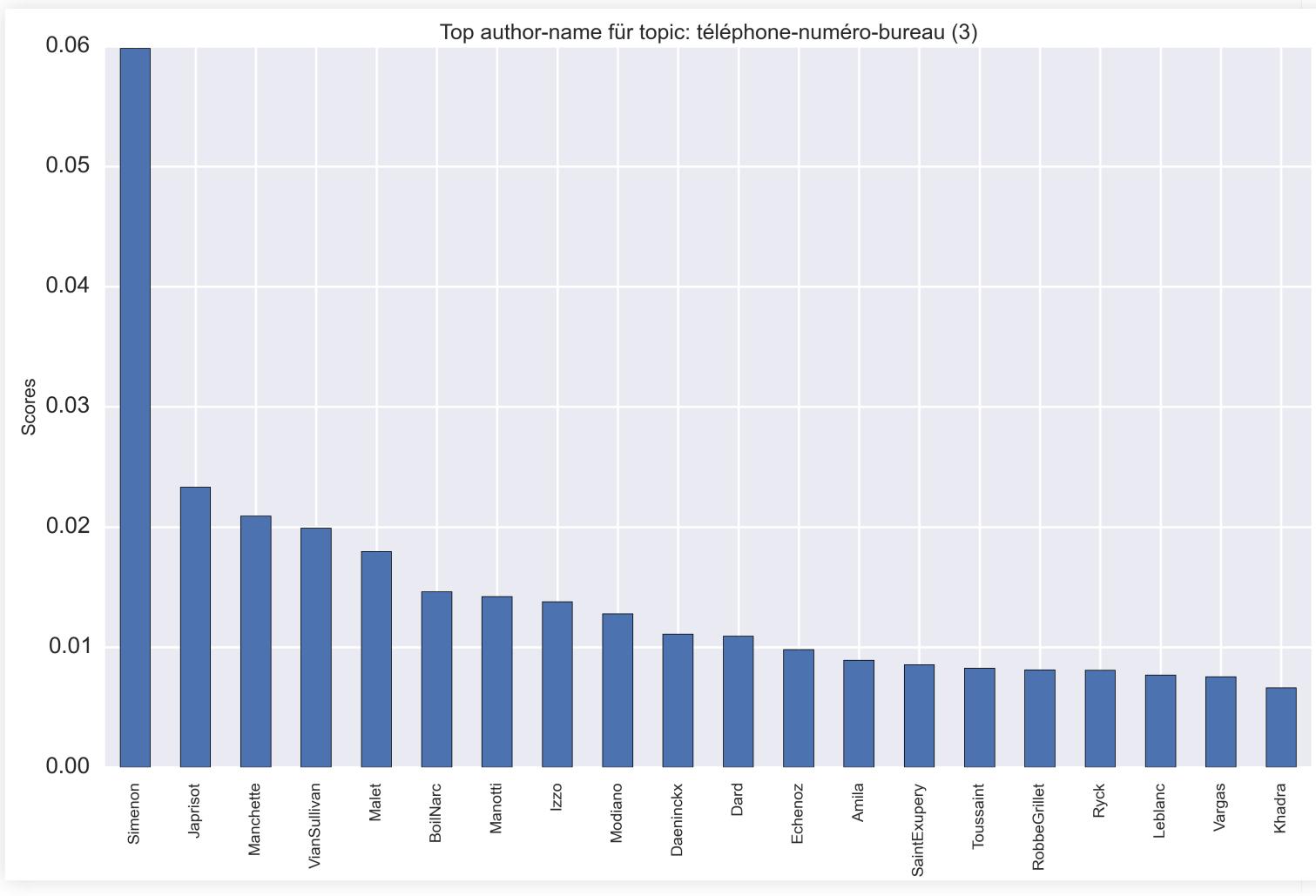
# Topics/Untergattung-Heatmap



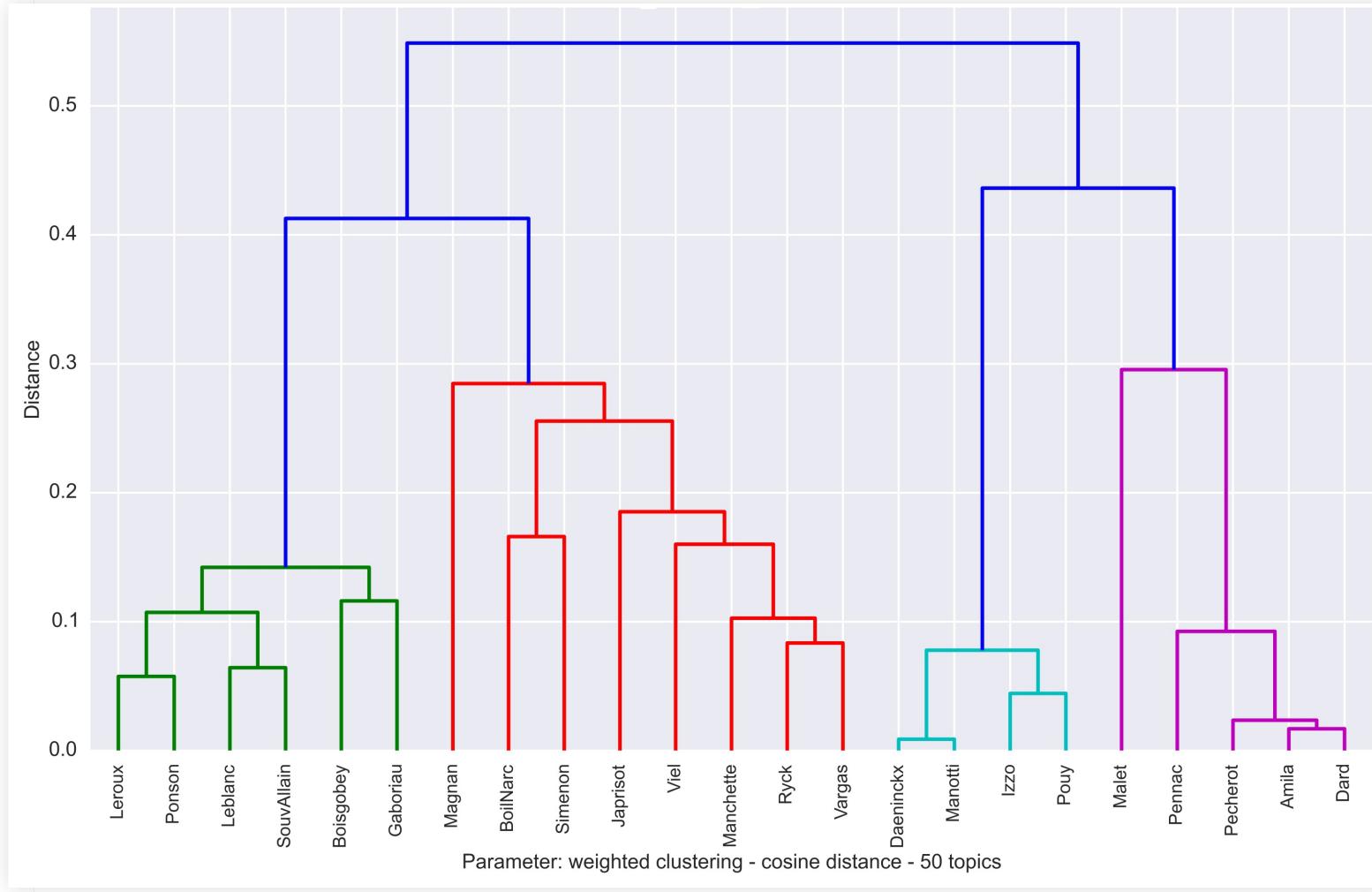
# Topics über die Jahrzehnte



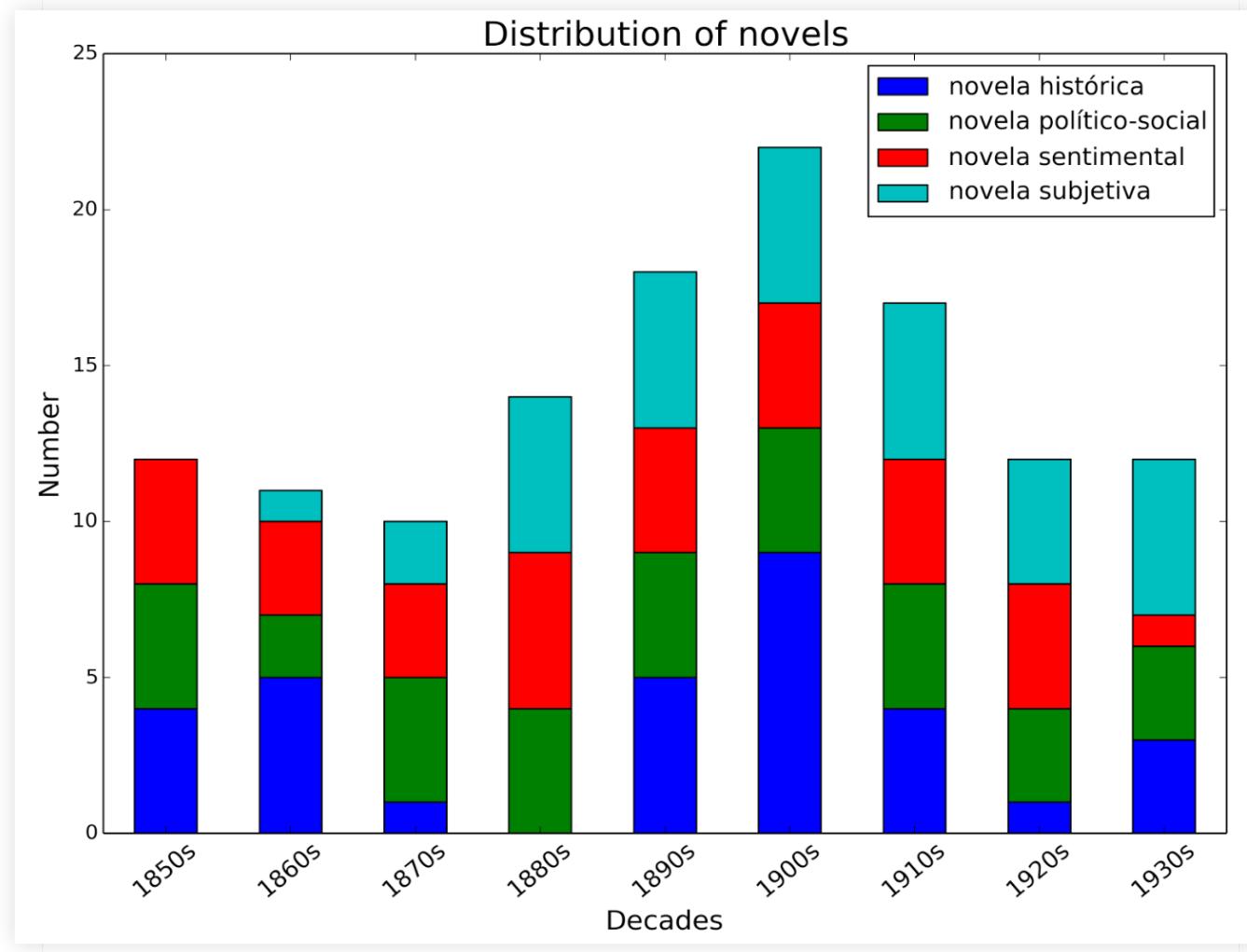
# Topics und Autoren: Einzeltopic



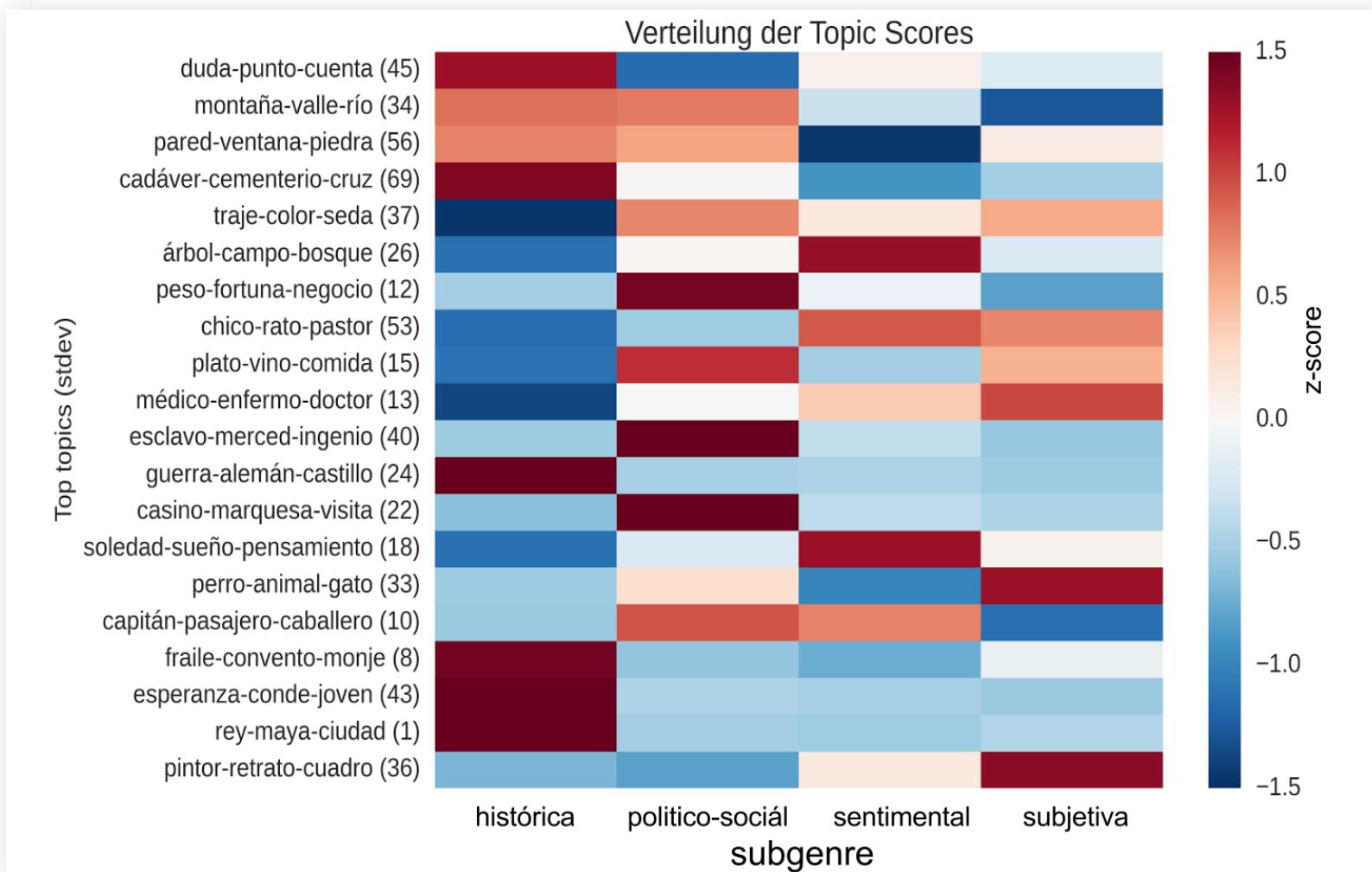
# Topics und Autoren: Clustering



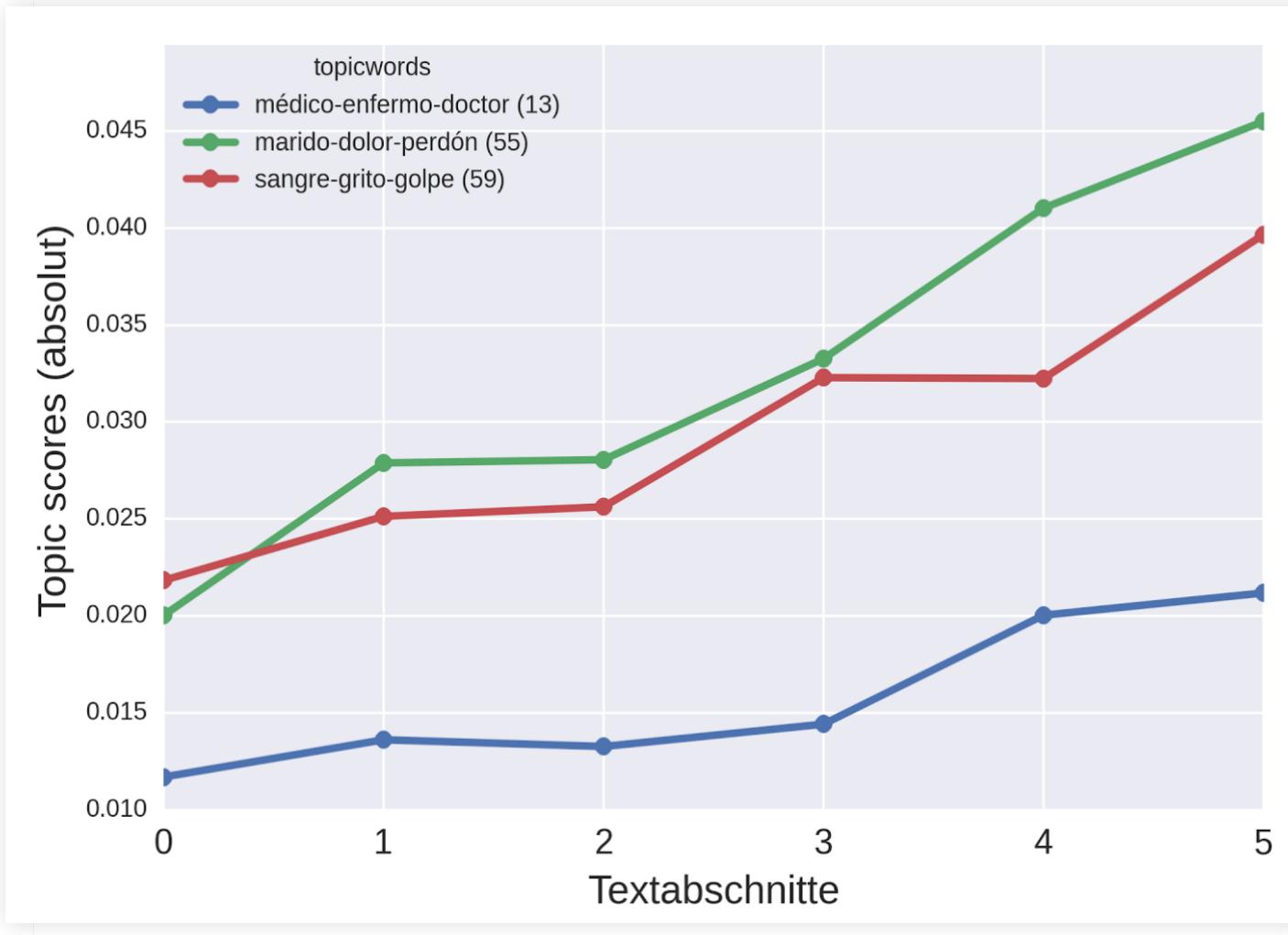
# Textsammlung: 127 spanische Romane



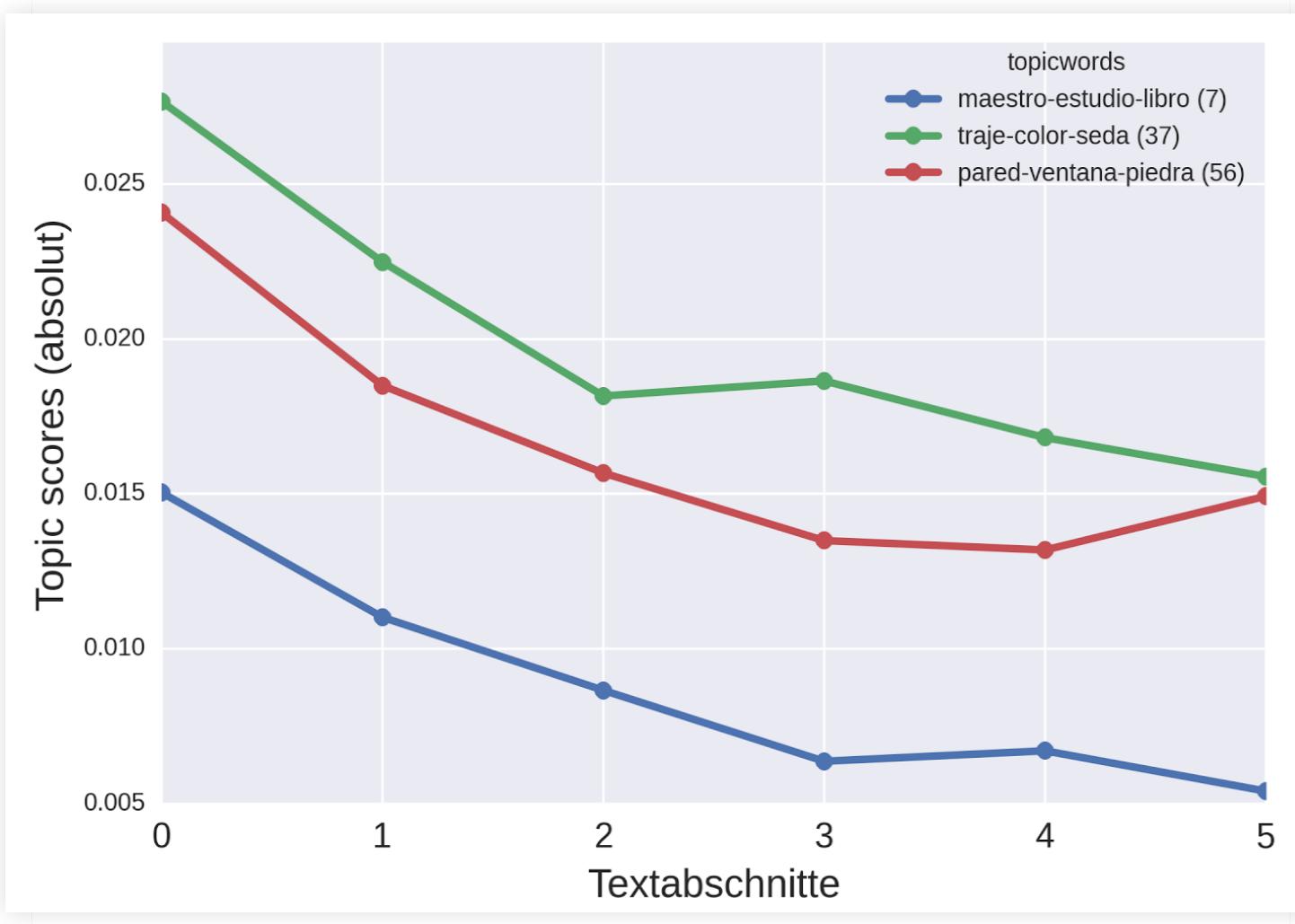
# Topic/Dokument-Matrix



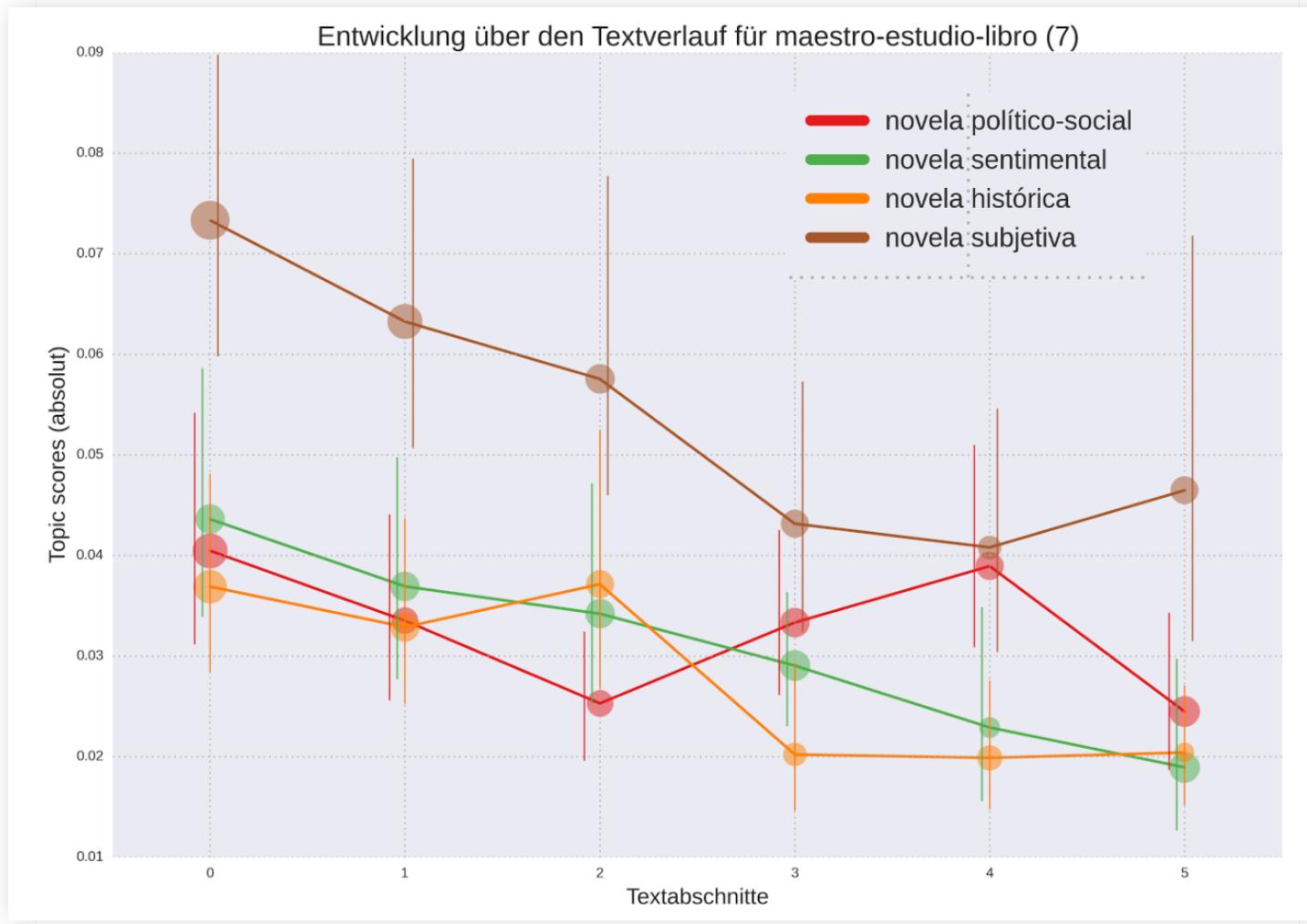
# Topic/Dokument-Matrix



# Topic/Dokument-Matrix



# Textverlauf und Untergattung



# 5. Weiterführendes

# Lektürehinweise: Theorie und Technik

- Blei, D. M. (2012). "Probabilistic topic models". In: *Communications of the ACM*, 55(4): 77–84.  
<http://www.cs.princeton.edu/~blei/papers/Blei2012.pdf>
- Steyvers, M. and Griffiths, T. (2006). "Probabilistic Topic Models". In: Landauer, T. et al. (eds), *Latent Semantic Analysis: A Road to Meaning*. Laurence Erlbaum.
- Weingart, S. (2012). "Topic Modeling for Humanists: A Guided Tour". In: *The Scottbot Irregular*.  
<http://www.scottbot.net/HIAL/?p=19113>

# Lektürehinweise: Anwendungsbeispiele

- Blevins, C. (2010). "Topic Modeling Martha Ballard's Diary". In: *Historying*. <http://historying.org/2010/04/01/topic-modeling-martha-ballards-diary/>
- Jockers, M. L. (2013). *Macroanalysis - Digital Methods and Literary History*. Champaign, IL: University of Illinois Press.
- Rhody, L. M. (2012). "Topic Modeling and Figurative Language". In: *Journal of Digital Humanities*, 2(1) <http://journalofdigitalhumanities.org/2-1/topic-modeling-and-figurative-language-by-lisa-m-rhody/>
- Schöch, C. (2016). "Topic Modeling Genre: An Exploration of French Classical and Enlightenment Drama". In: *Digital Humanities Quarterly*. <http://digitalhumanities.org/dhq/>
- Underwood, T. and Goldstone, A. (2012). "What can topic models of PMLA teach us about the history of literary scholarship?" In: *The Stone and the Shell*. <http://tedunderwood.com/2012/12/14/what-can-topic-models-of-pmla-teach-us-about-the-history-of-literary-scholarship/>

Autor: Christof Schöch, 2016

[christof-schoech.de](http://christof-schoech.de)

---

Lizenz: Creative Commons Attribution 4.0 International

[creativecommons.org/licenses/by/4.0/](http://creativecommons.org/licenses/by/4.0/)