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

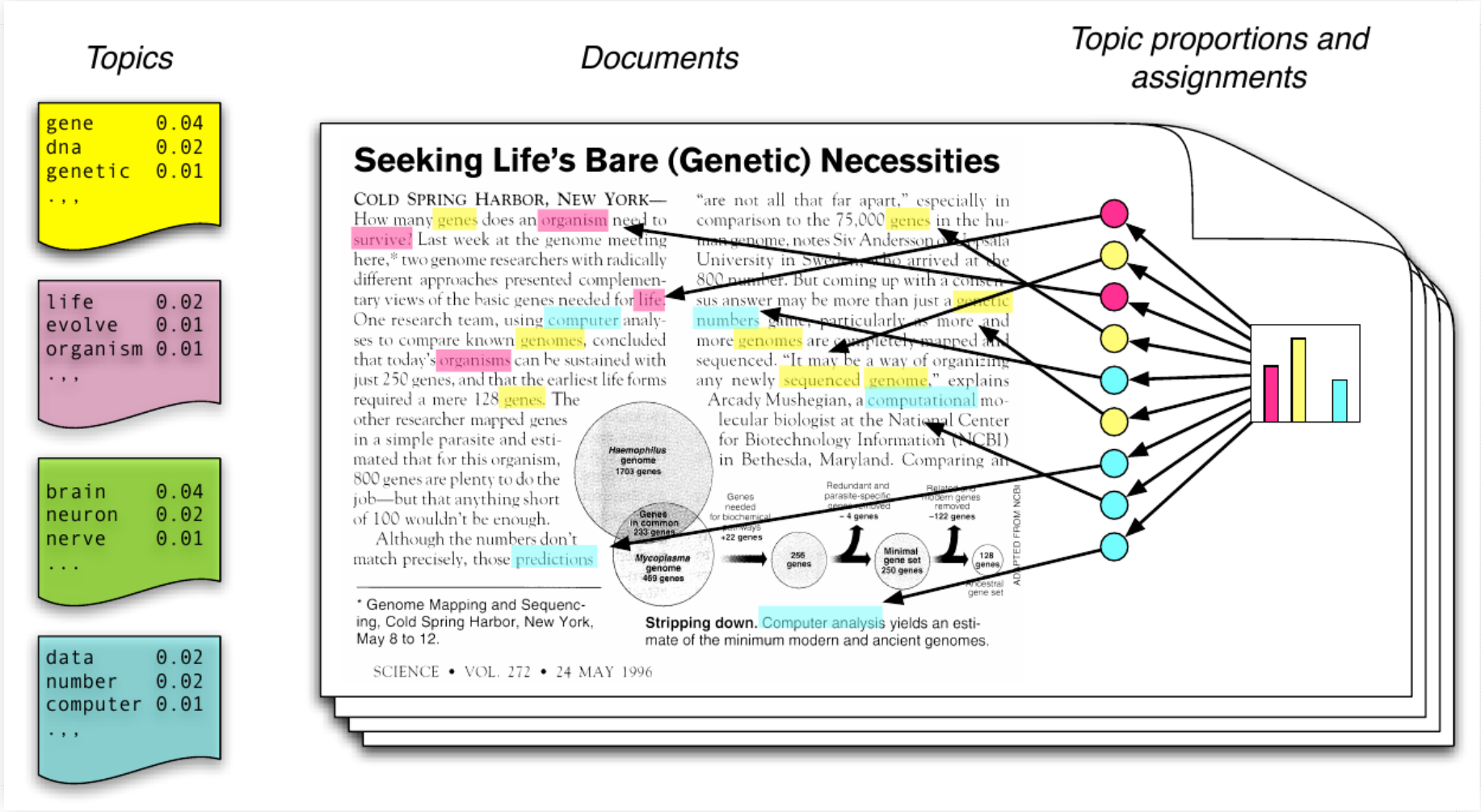
1. Was ist Topic Modeling?

Topic Modeling

Grundidee

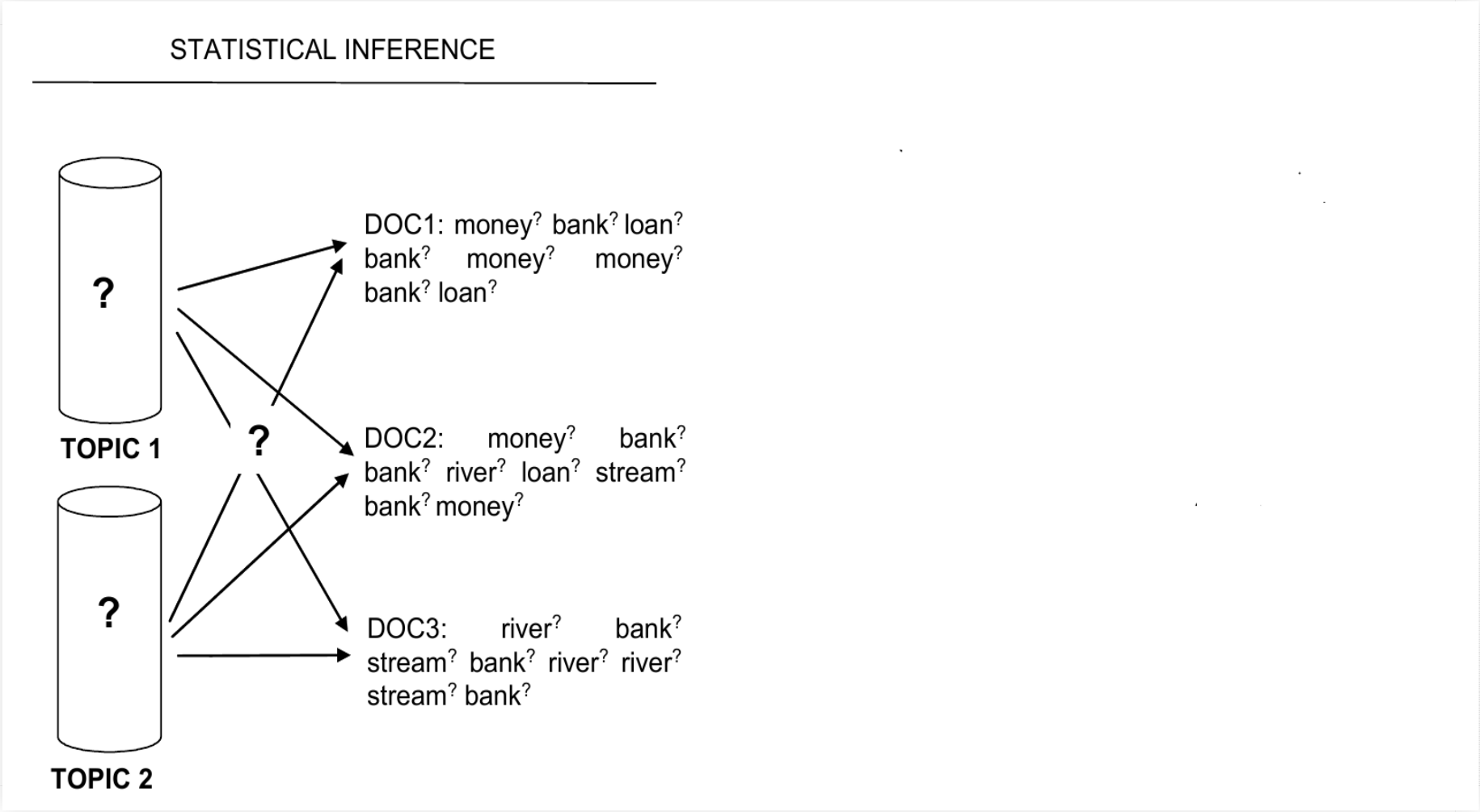
- Entdeckt Wörter, die immer wieder gemeinsam bzw. in ähnlichen Kontexten vorkommen (Topics)
- Berechnet, wie wichtig jeder Topic in jedem Dokument ist
- Wichtig: keinerlei explizites semantisches Wissen fließt ein
  Etwas technischer
- Ein Topic ist eine Verteilung von Wahrscheinlichkeiten von Wörtern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

Wörter, Topics, Dokumente



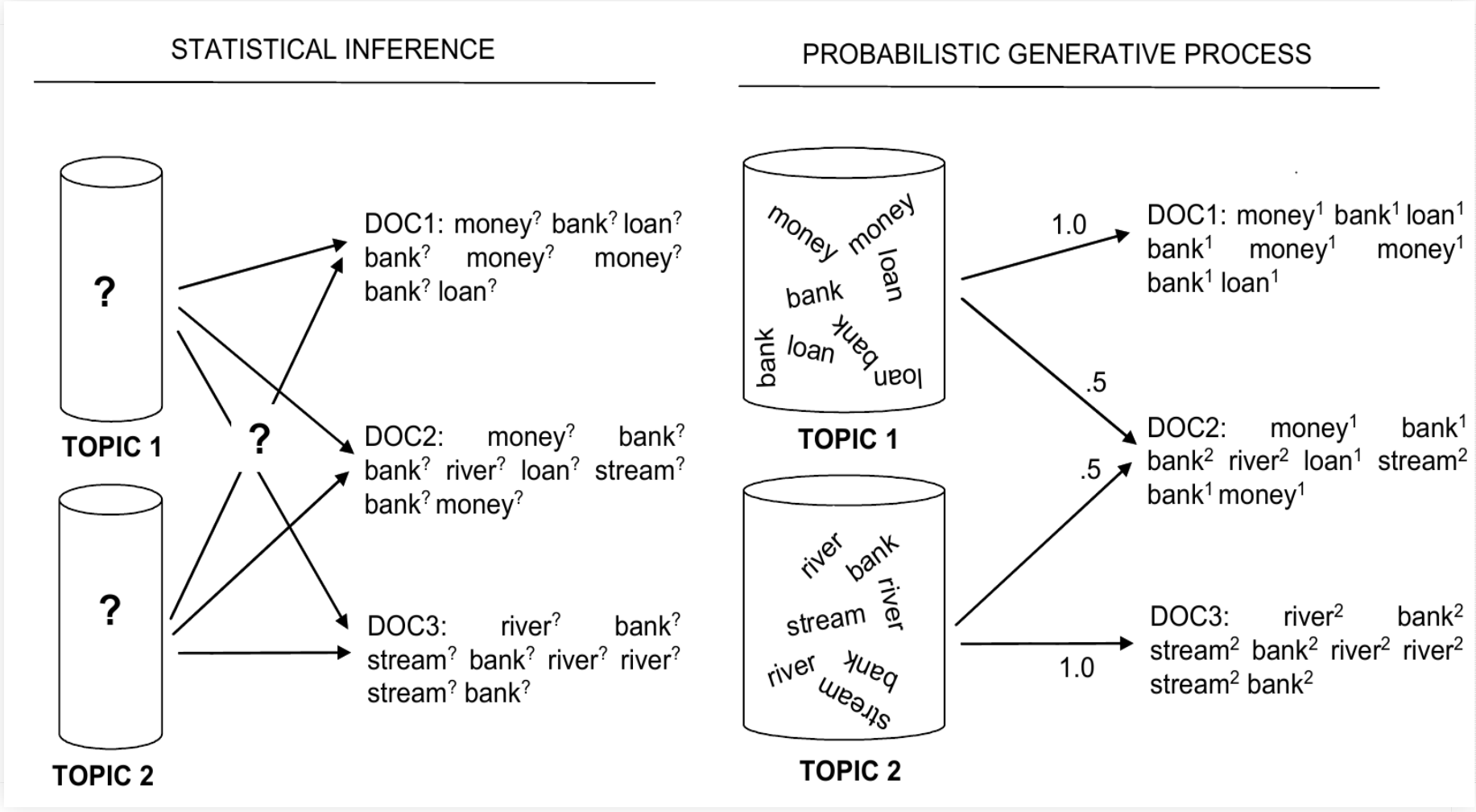
(David Blei, "Probabilistic Topic Models", 2012)

Generativ, Iterativ



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

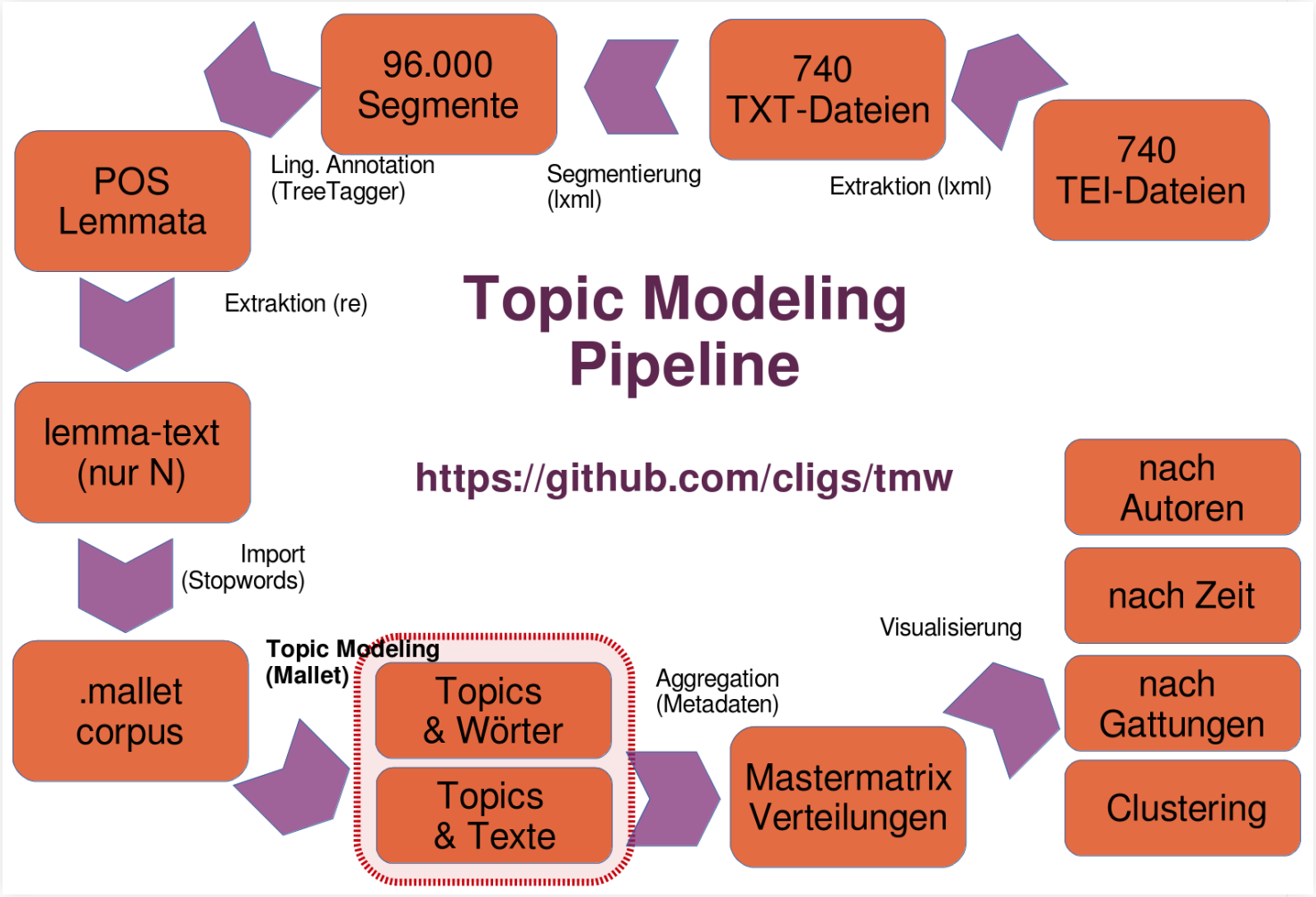
Generativ, Iterativ



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

2. Anwendung: Topic Modeling Workflow

Preprocessing, Modeling, Postprocessing

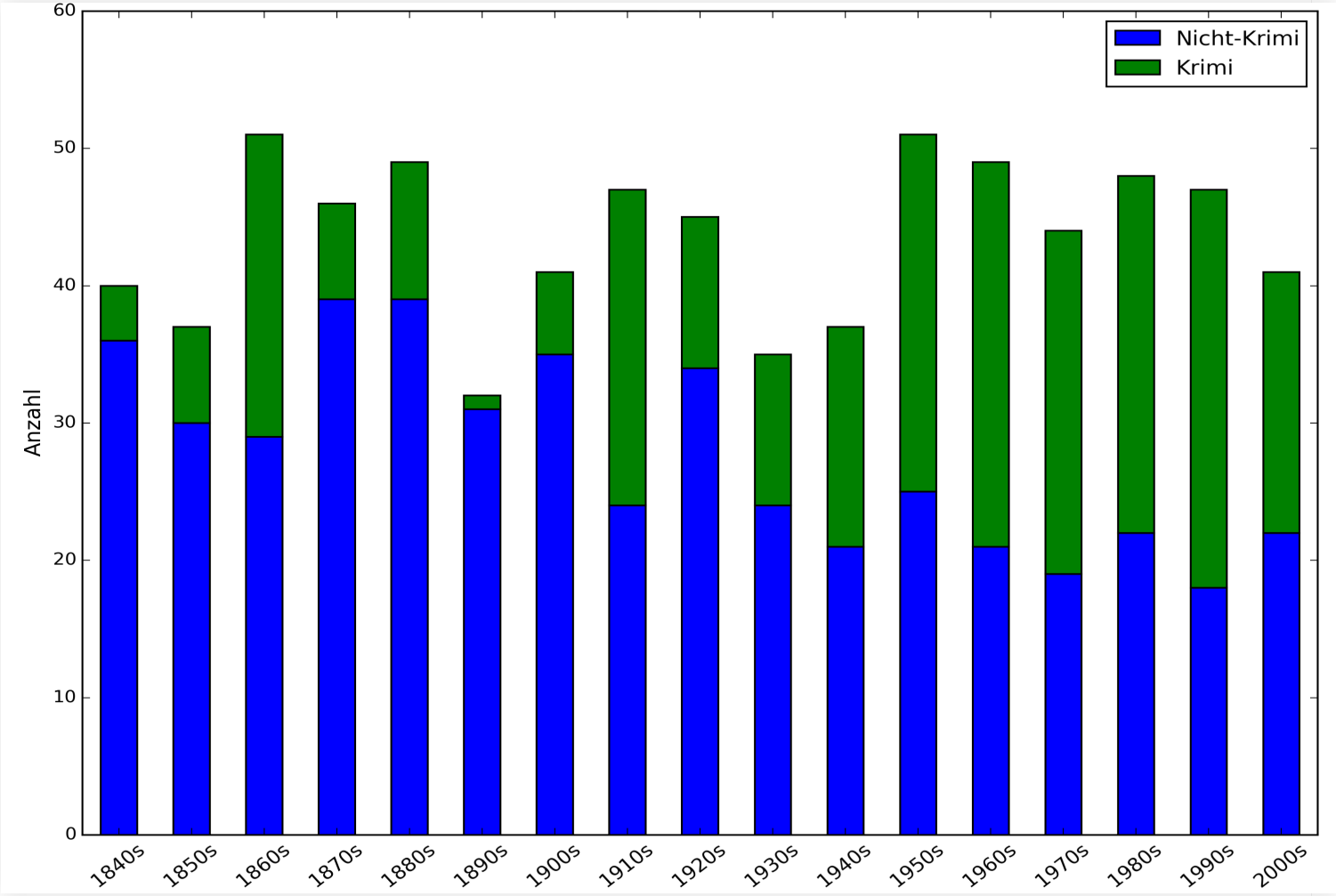


Umsetzung: Mallet und Python

- 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

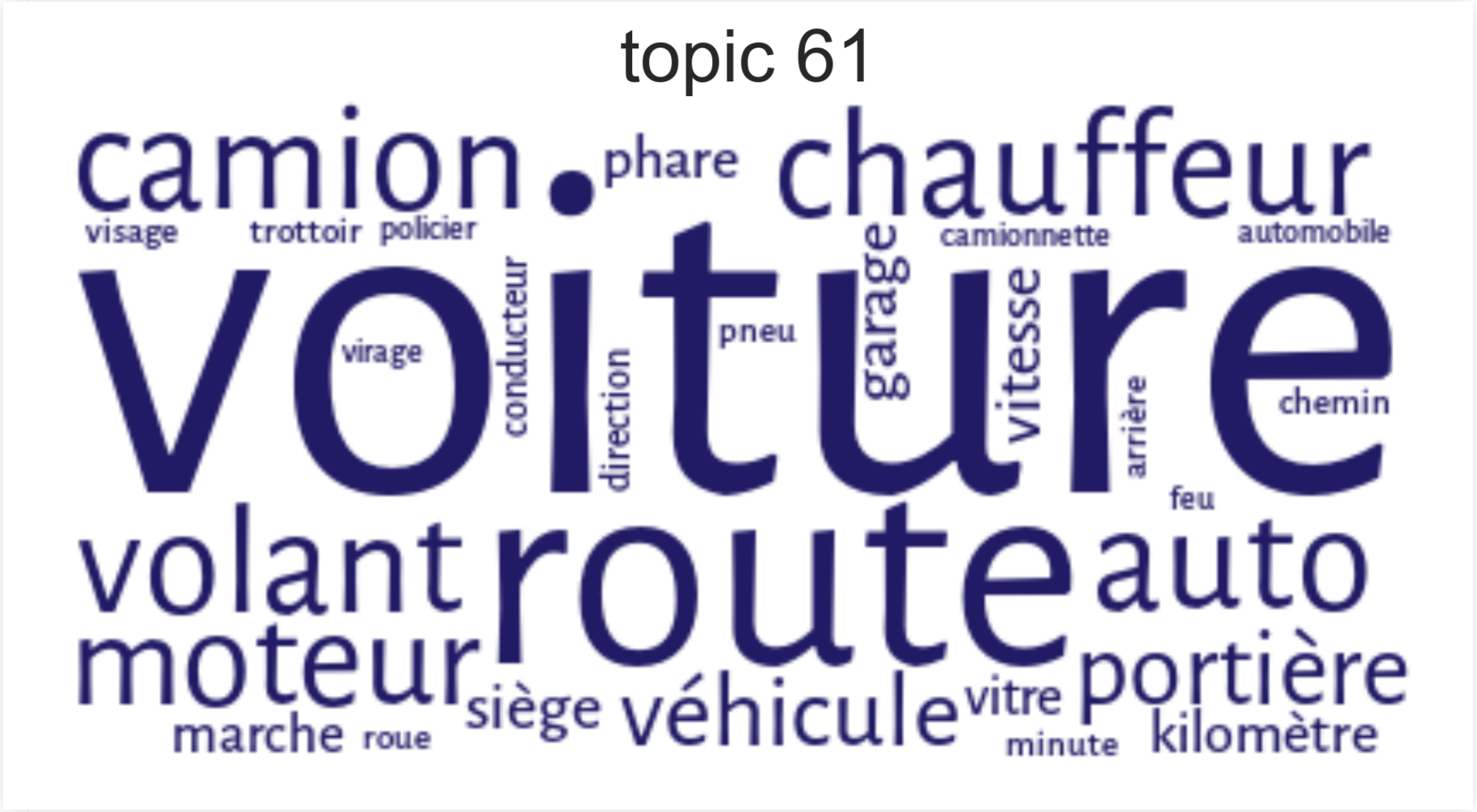
Professor-Wissenschaft-Weiser

Verbrechen-Tod-Mord

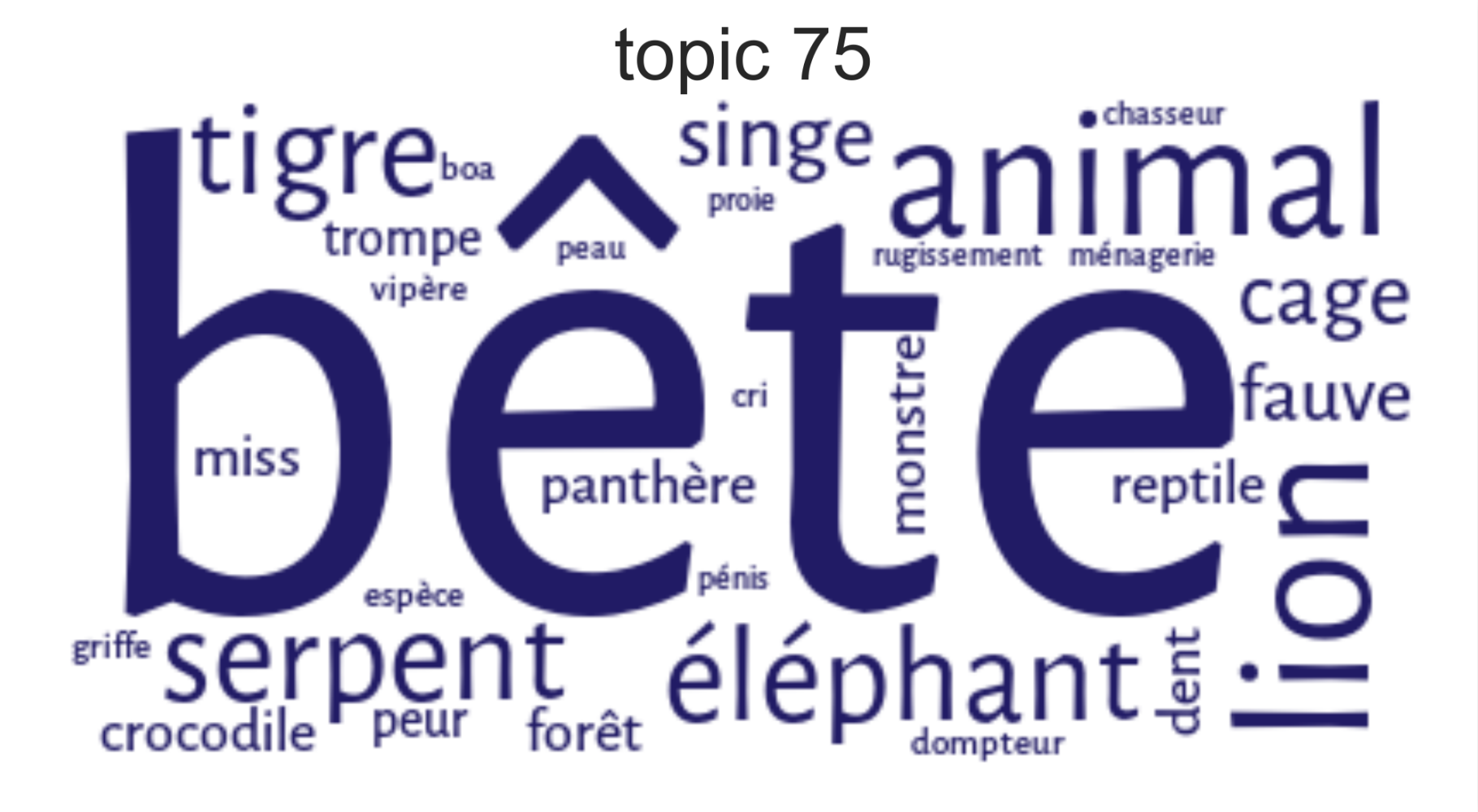
Musik-Piano-Lied

Buch-Bibliothek-Seite

Auto-Straße-Fahrer

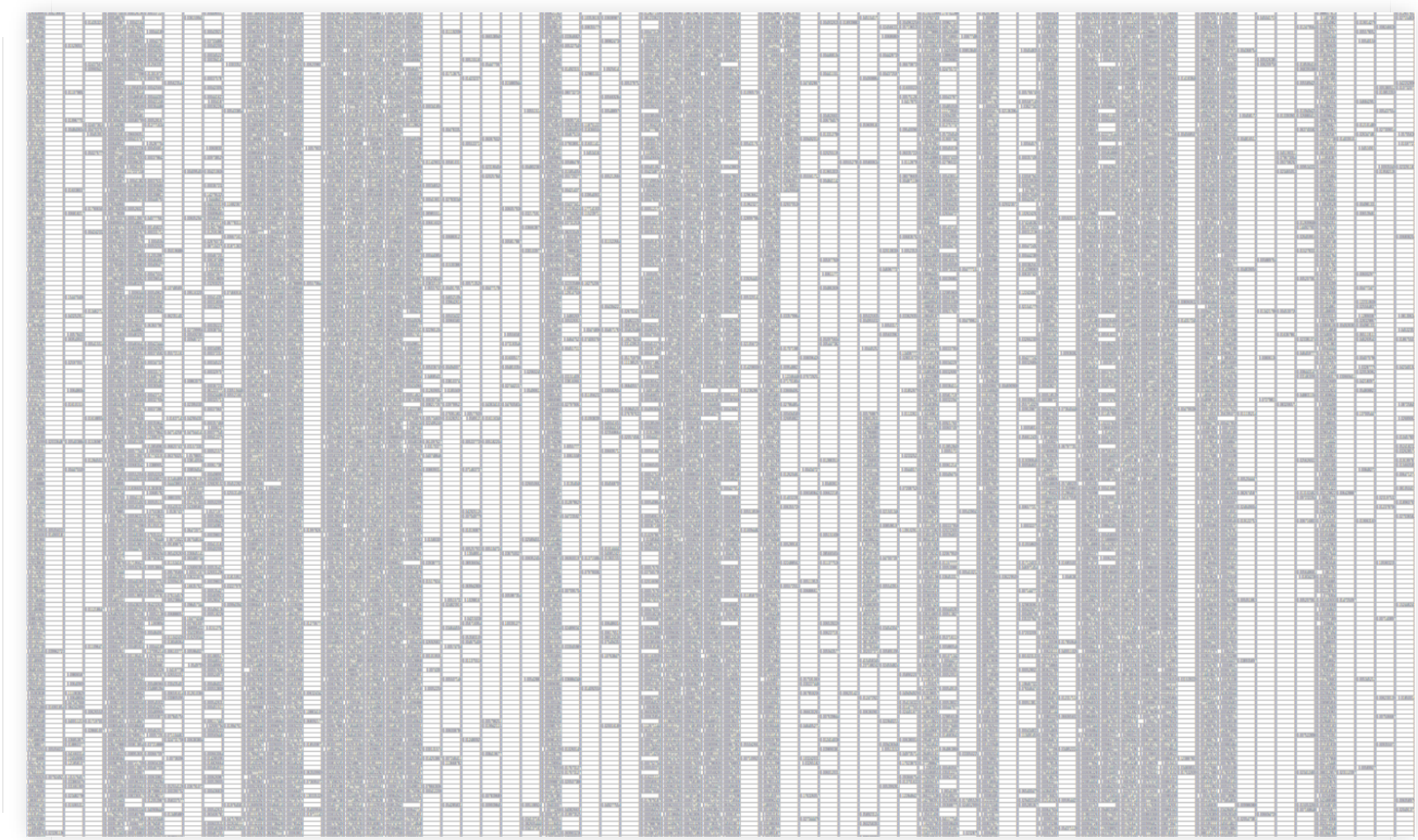


Tier-Tier-Löwe

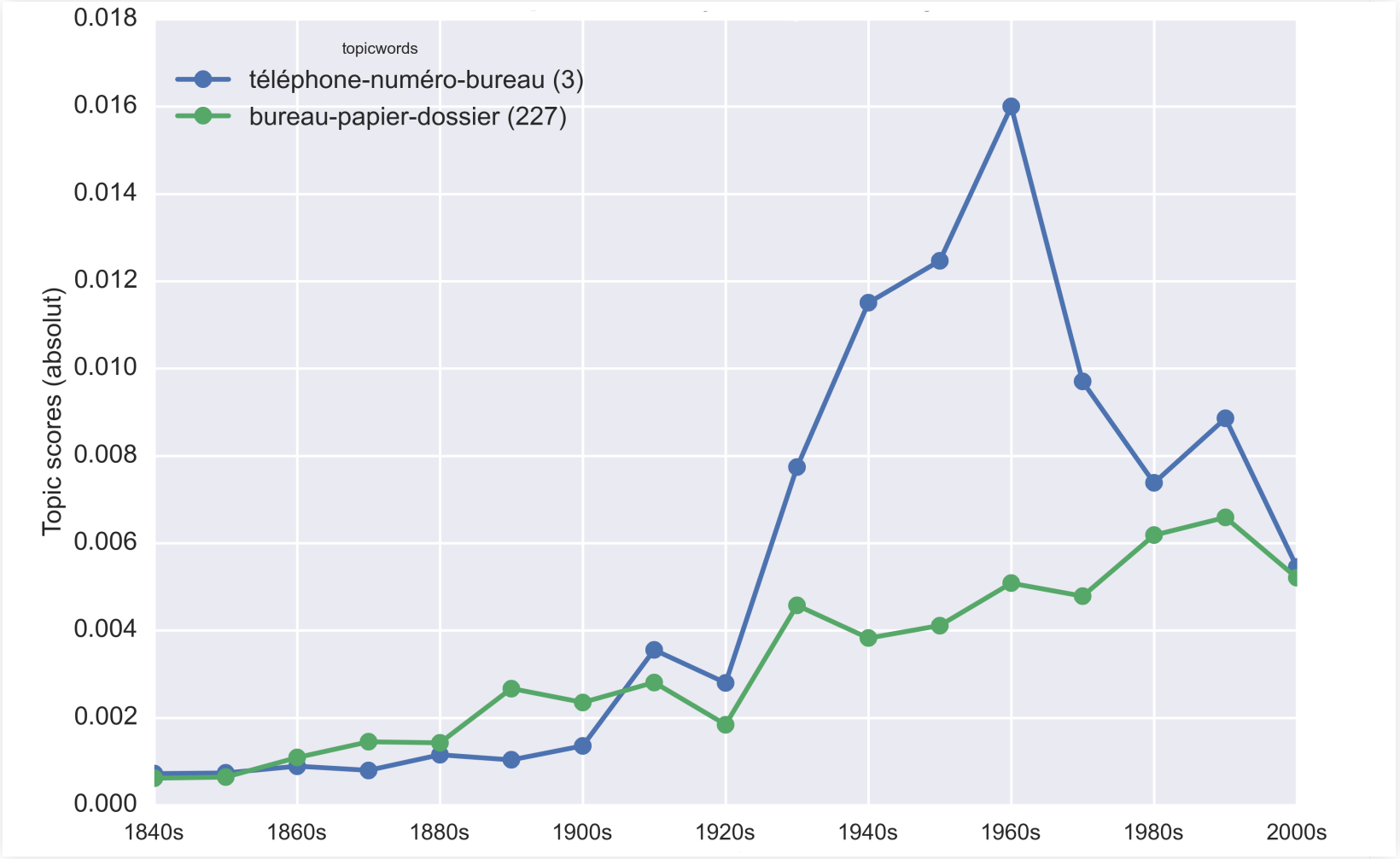


4. Ergebnisse: Visualisierung der Verteilungen

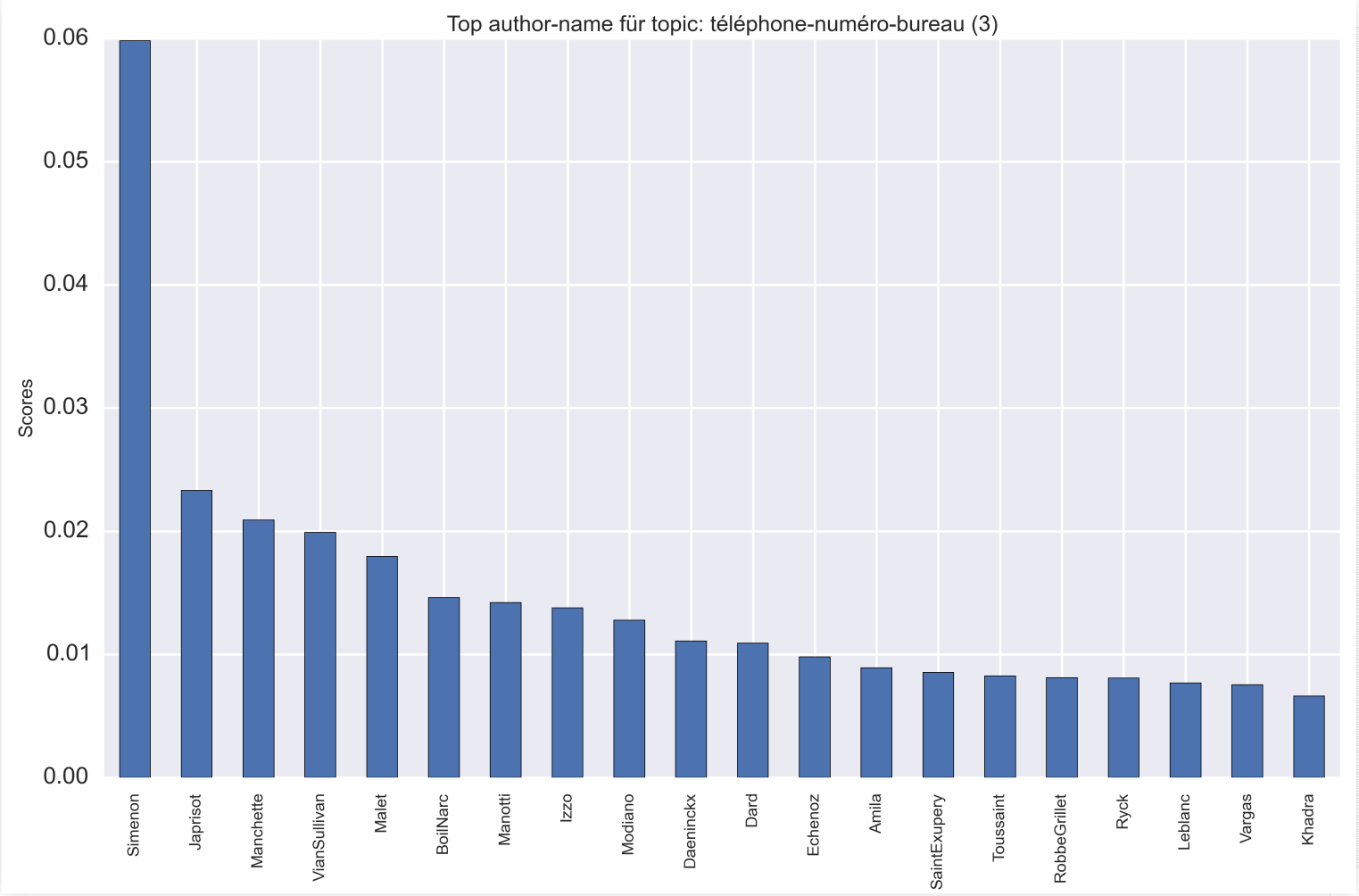
Topics/Untergattung-Heatmap



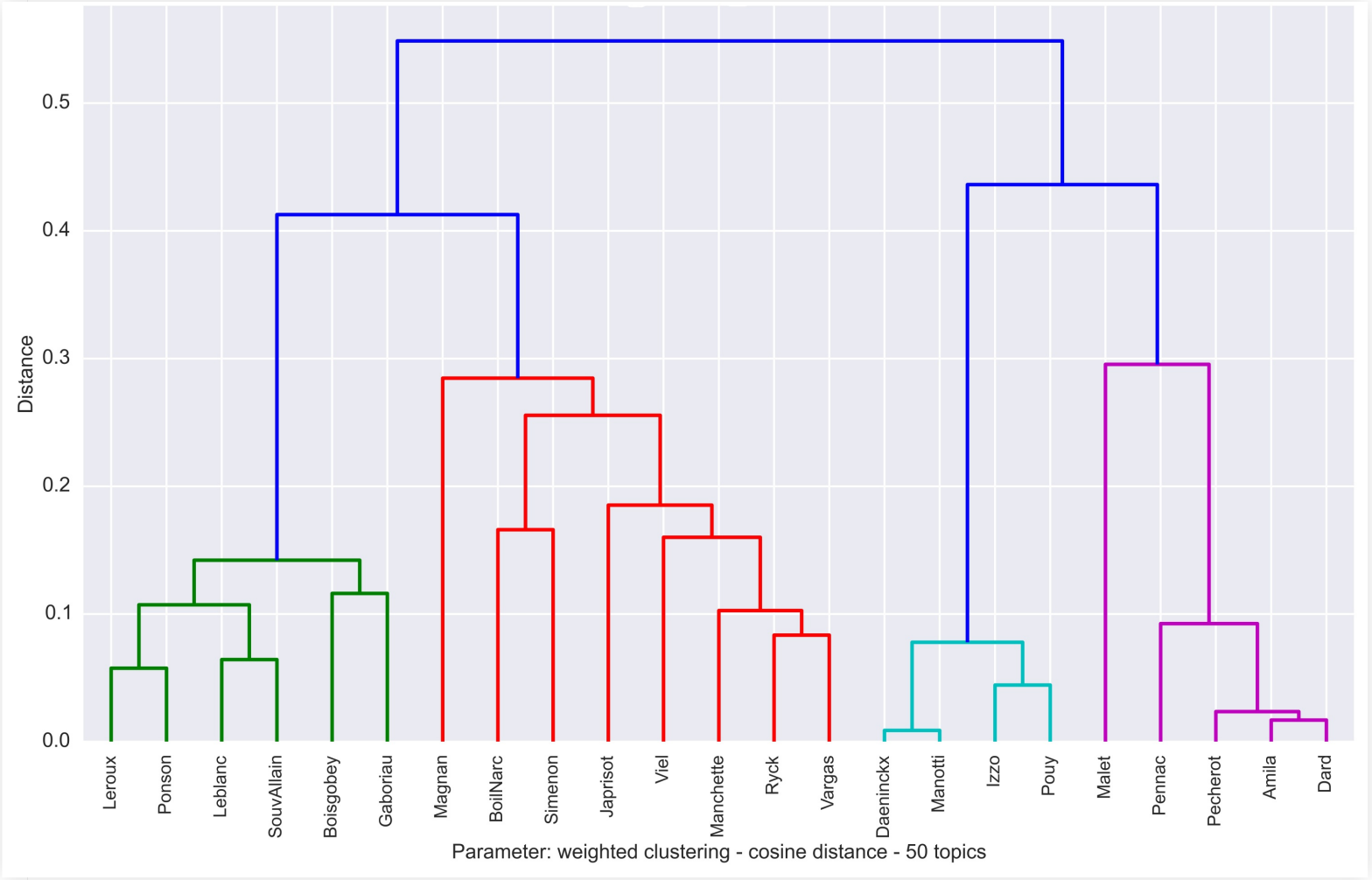
Topics über die Jahrzehnte



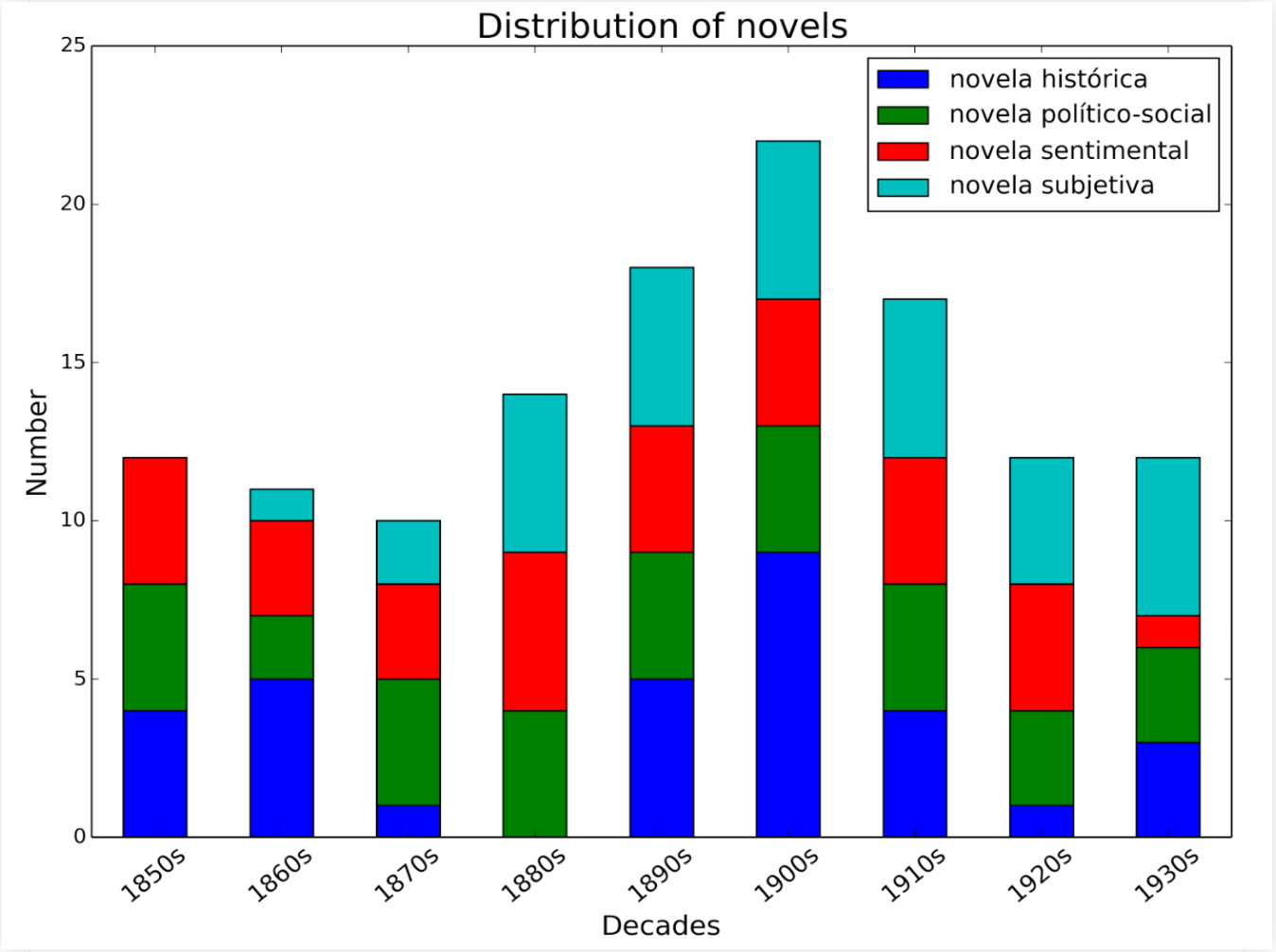
Topics und Autoren: Einzeltopic



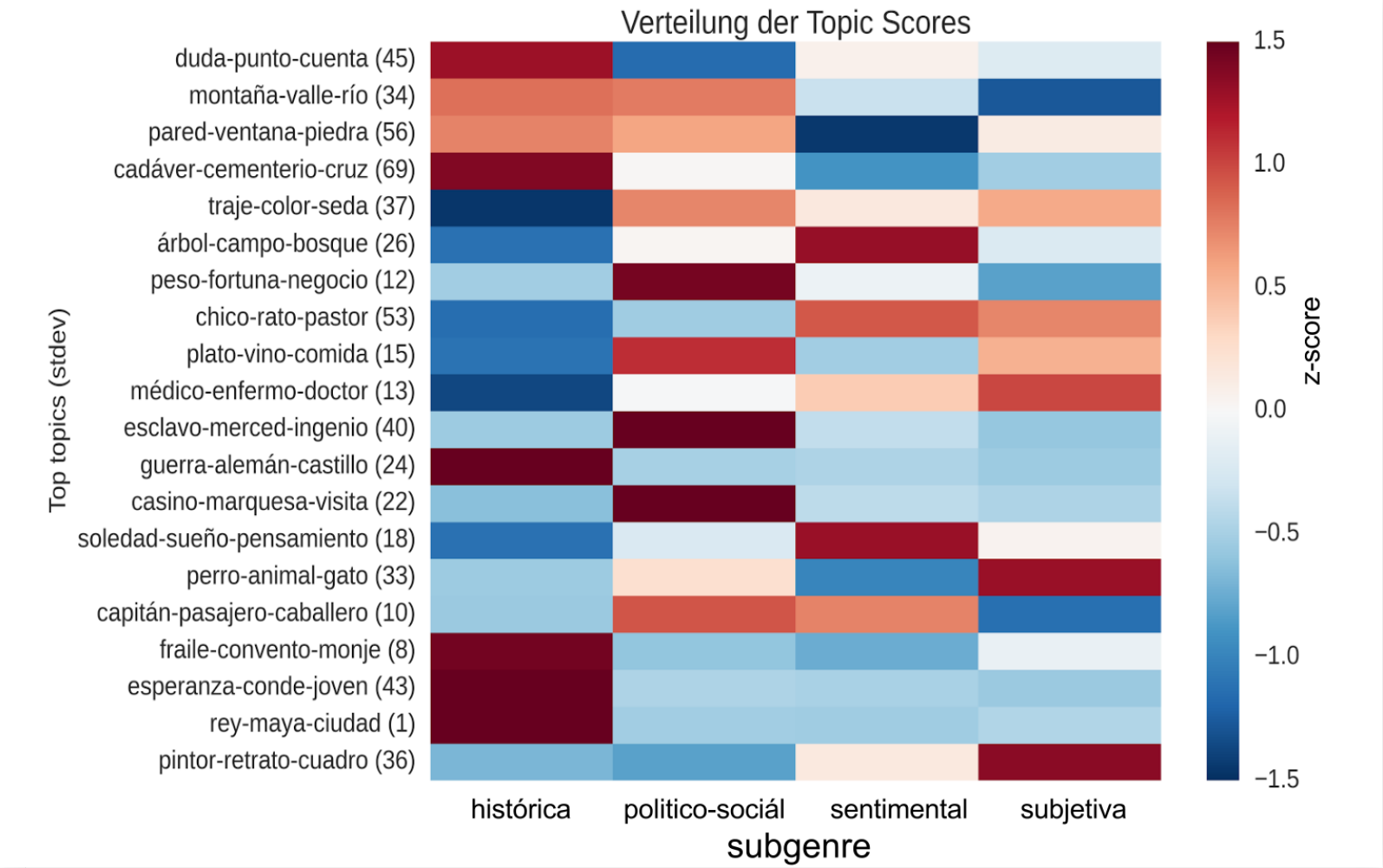
Topics und Autoren: Clustering



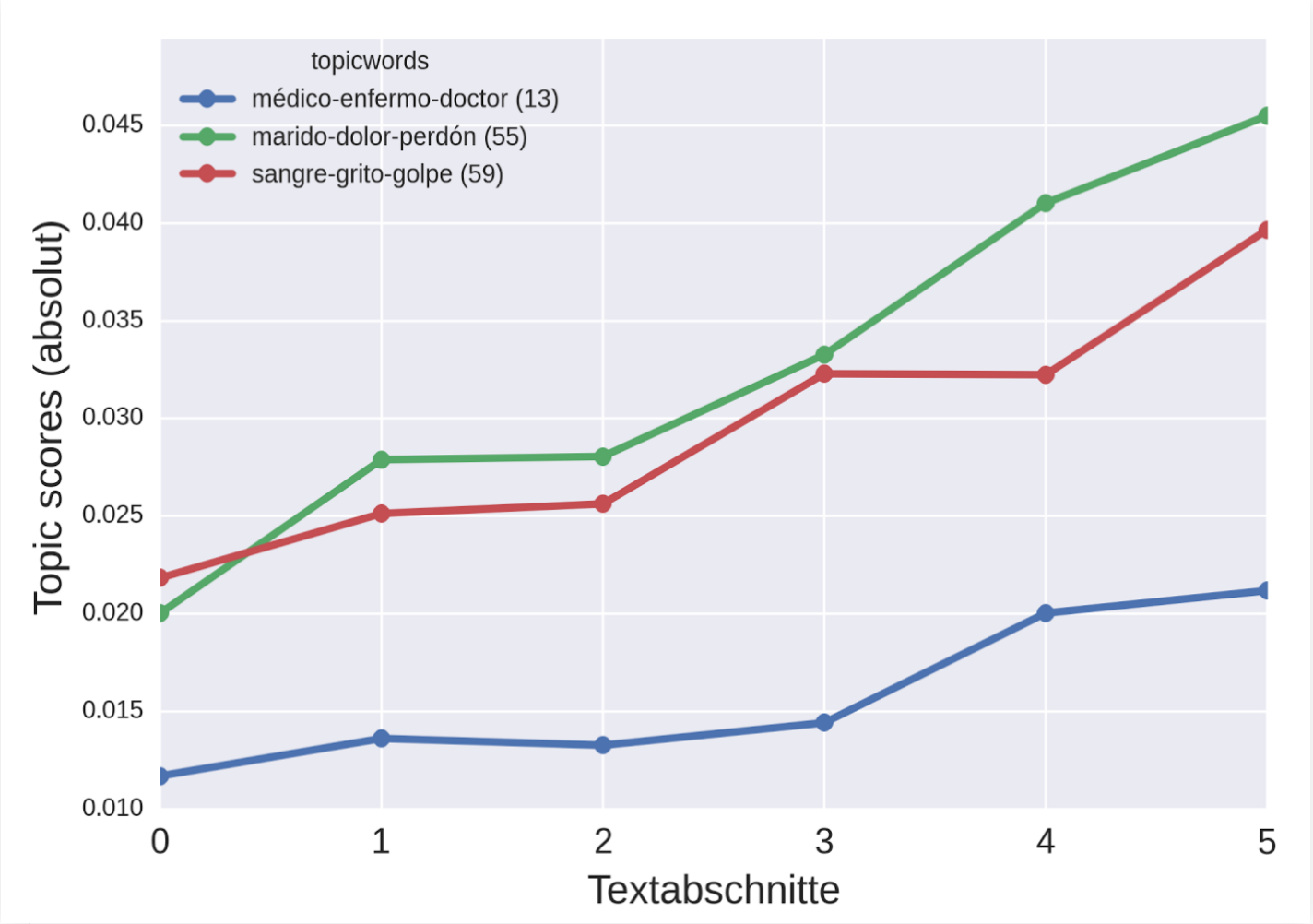
Textsammlung: 127 spanische Romane



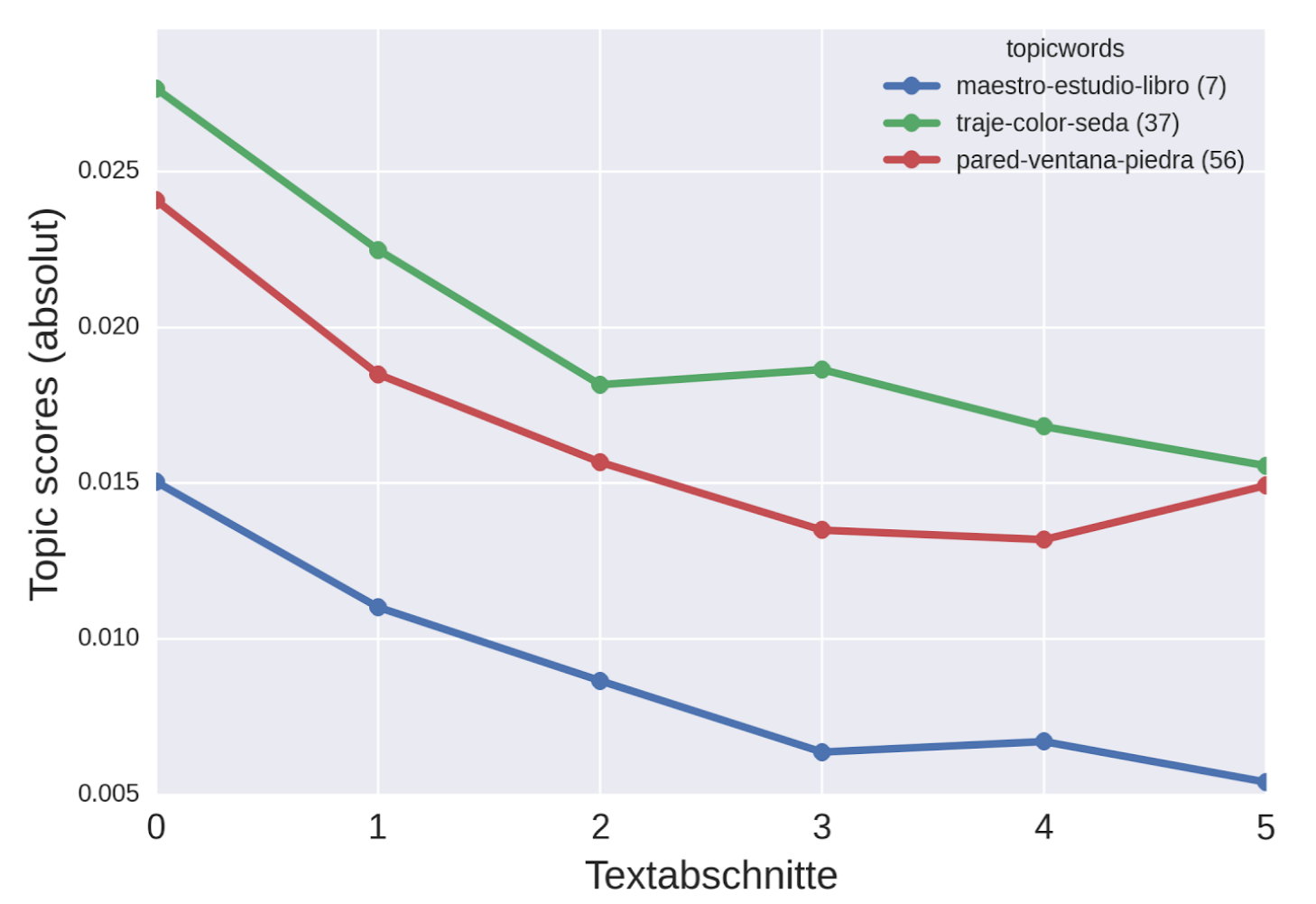
Topic/Dokument-Matrix



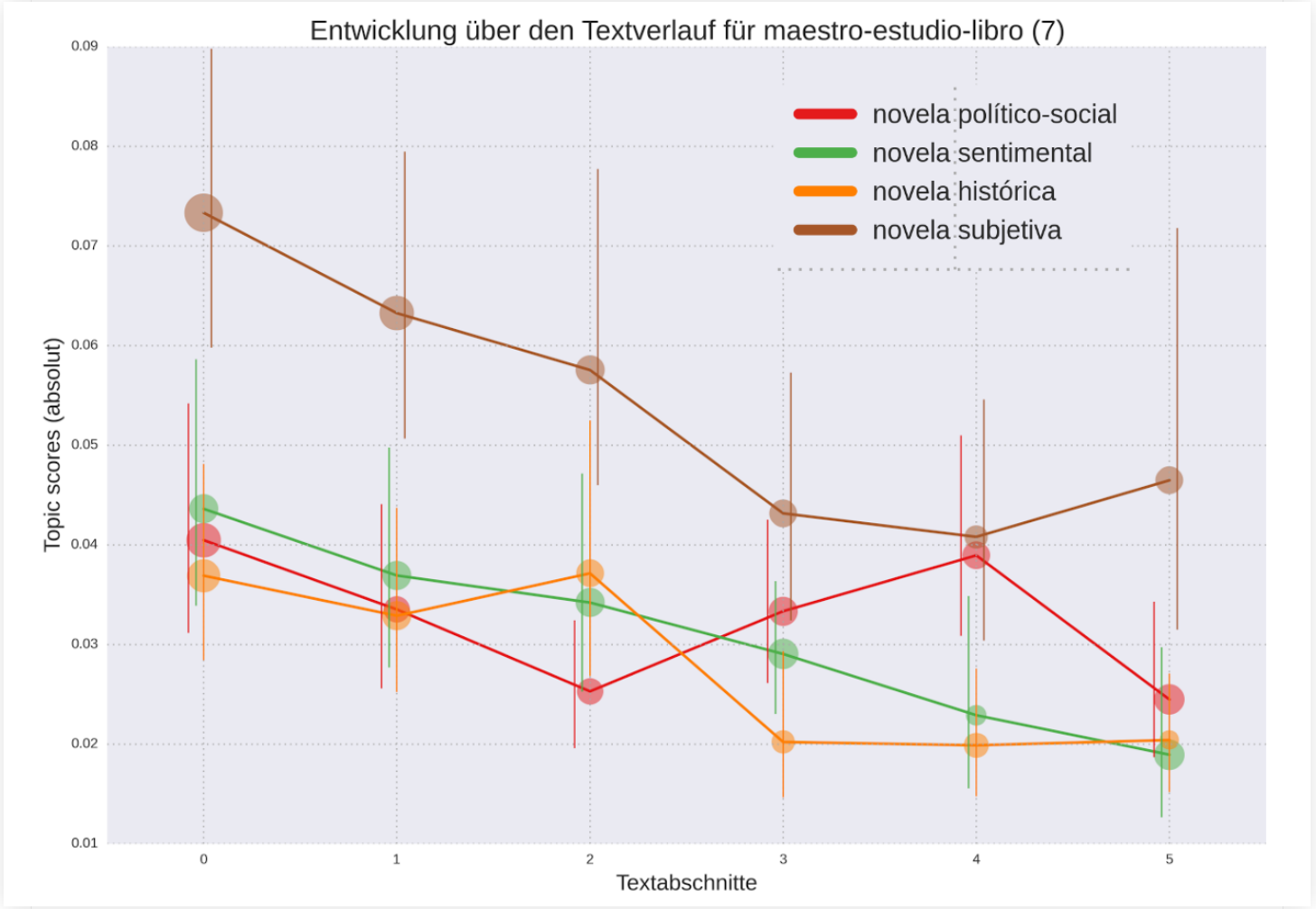
Topic/Dokument-Matrix



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
