Von Themen zu Topics

Über die Einsatzmöglichkeiten von Topic Modeling für quantitativ gestützte Inhaltsanalysen in den Geisteswissenschaften

Ulrike Henny, Universität Würzburg
ulrike.henny@uni-wuerzburg.de

Von Themen zu Topics

1. Was ist Topic Modeling?
2. Zur Anwendbarkeit von Topic Modeling
3. Beispiele für Textanalysen mit Topic Modeling
   - 3.1 Blevins: Text Analysis of Martha Ballard‘s Diary
   - 3.2 CLiGS: Topics im Textverlauf
4. Zur Interpretierbarkeit der Ergebnisse

Was ist Topic Modeling?

1. Was ist Topic Modeling?

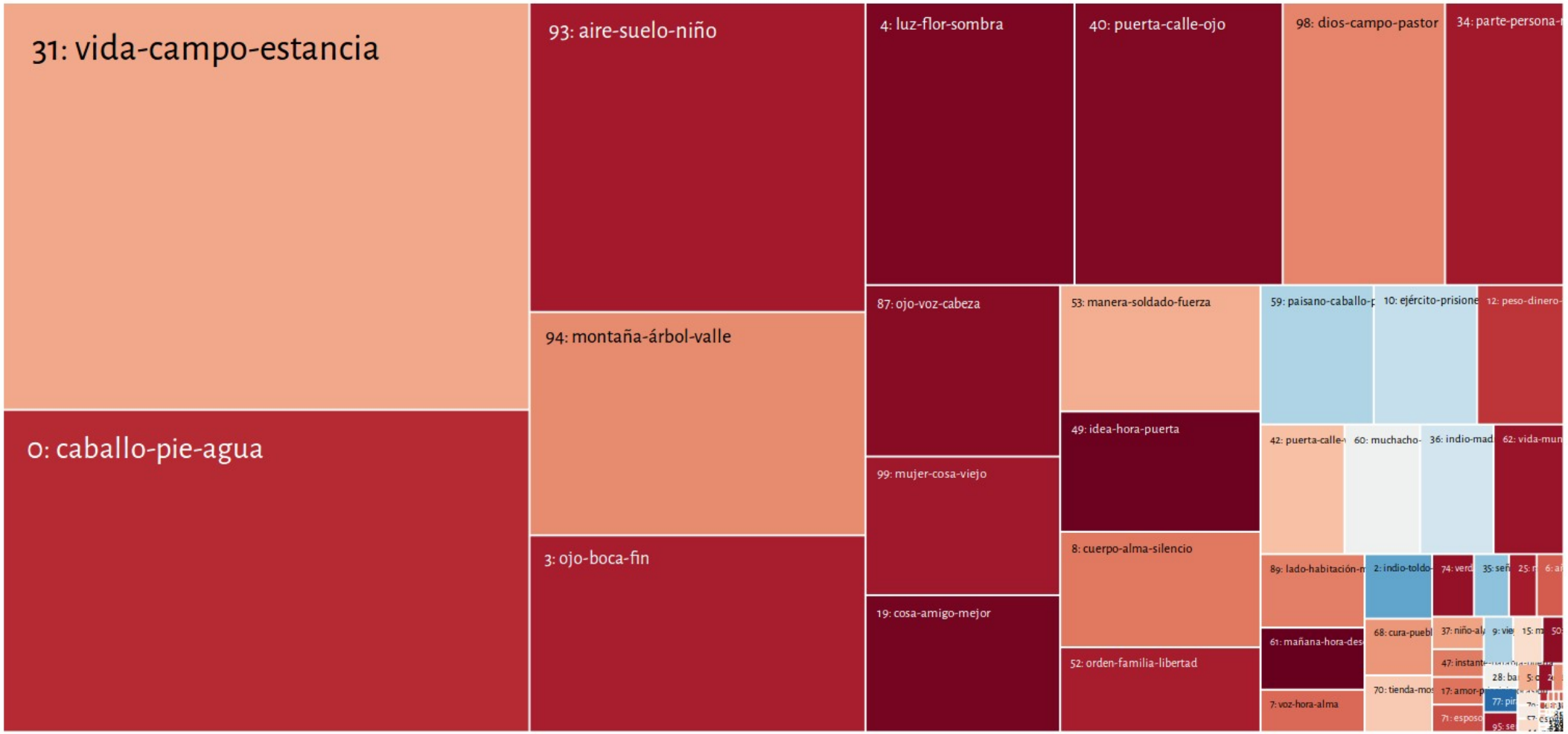
- quantitative Methode der Textanalyse
- in einem Korpus von Dokumenten (einer Textsammlung) werden Wortverteilungen statistisch ermittelt
- Ziel: Aufdecken „versteckter“ semantischer Strukturen
  - ohne explizites semantisches Wissen
  - Distributionelle Hypothese: Wörter, die in demselben Kontext vorkommen, tendieren dazu, eine ähnliche Bedeutung zu haben
  - wiederkehrende Themen, Motive, Diskurse werden automatisch identifiziert

1. Was ist Topic Modeling?

- Topic-Modell
  - Bestimmte Anzahl von Topics (für eine Dokumentsammlung)
  - Topic: Verteilung von Wahrscheinlichkeiten von Wörtern
  - Dokument: Verteilung von Wahrscheinlichkeiten von Topics
- Wie entsteht ein Topic-Modell?
  - gegeben: Dokumente, Wörter
  - versteckt: Topics
  - Anfangsmodell wird in iterativem Prozess optimiert

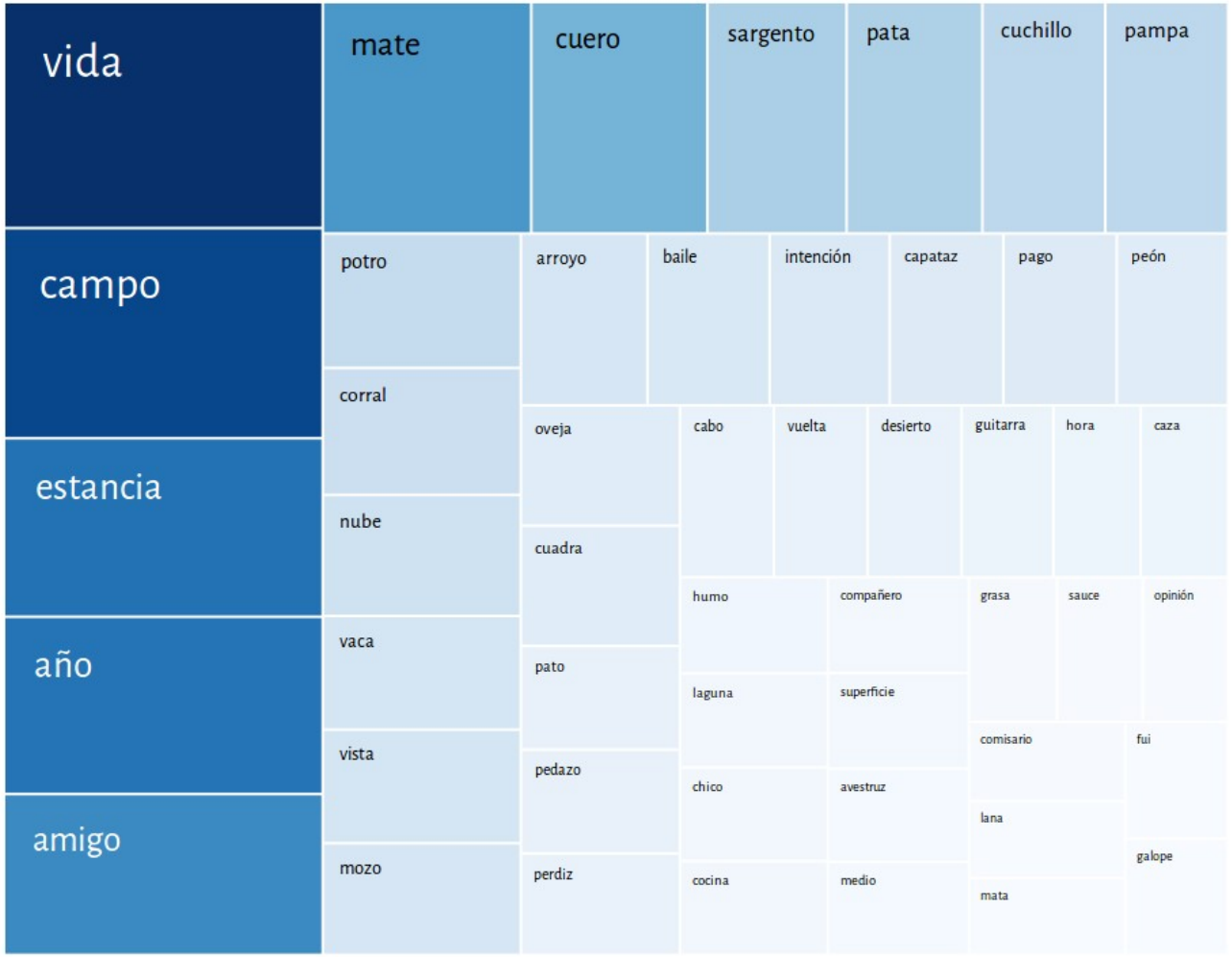
1. Was ist Topic Modeling?

Beispiel-Dokument: Ciro B. Ceballos, *Un adulterio* (1901), Kurzroman, Mexiko



1. Was ist Topic Modeling?

Beispiel-Topic: *vida-campo-estancia / Leben-Feld-Estanzia (Landgut)*



Zur Anwendbarkeit von Topic Modeling

2. Zur Anwendbarkeit von Topic Modeling

Voraussetzungen

- Textkorpus
  - (viele, ggf. normalisierte) Volltexte
  - Metadaten
- Kenntnis des Materials

Ablauf

- Vorarbeiten: Segmentierung, Tokenisierung, Lemmatisierung, POS-Tagging
- Modellierung
- Nacharbeiten: Zusammenführen mit Metadaten, Visualisierung, Auswertung, Evaluation

2. Zur Anwendbarkeit von Topic Modeling

Tools

- *Core*
  - Mallet (Java, http://mallet.cs.umass.edu/)
  - Gensim (Python, https://radimrehurek.com/gensim)
- Workflows
  - TMW (Christof Schöch, https://github.com/cligs/tmw)
  - „Cophi-Toolbox“ (im Aufbau, https://github.com/thvitt/cophi-toolbox)
- Visualisierung
  - Serendip (http://vep.cs.wisc.edu/serendip/, wird vermutlich nicht mehr entwickelt)
  - LDAVis (Demo: http://www.kennyshirley.com/LDAvis/, GitHub: https://github.com/cpsievert/LDAvis)

2. Zur Anwendbarkeit von Topic Modeling

- keine allzu großen sachlichen Voraussetzungen
- aber: ganz ohne Programmierkenntnisse ist es derzeit noch schwierig

Beispiele von Textanalysen mit Topic Modeling

3. Beispiele für Textanalysen mit Topic Modeling

1. Blevins: Text Analysis of Martha Ballard‘s Diary
2. CLiGS: Topics im Textverlauf

3.1 Blevins: Text Analysis of Martha Ballard‘s Diary

- Tagebuch einer Hebamme aus Maine, zwischen 1785 und 1812 geführt
- Von Cameron Blevins mit Text-Mining-Methoden analysiert
- Zuvor: Monographie „A Midwife‘s Tale“ von Laurel Ulrich
- Tagebuch:
  - Fast 10.000 Einträge
  - Fast tägliche Notizen

3.1 Blevins: Text Analysis of Martha Ballard‘s Diary

Ulrich: “*The problem is not that the diary is trivial but that it introduces more stories than can be easily recovered and absorbed.*”

Blevins: “*how does a reader (computer or human) recognize and conceptualize the recurrent themes that run through nearly 10,000 entries?*“

“*One answer lies in topic modeling*“

“*in the case of Martha Ballard’s diary, it worked. Beautifully*“

3.1 Blevins: Text Analysis of Martha Ballard‘s Diary

Mallet, 30 Topics, hier ein Sample (Top 20 Wörter, von Blevins mit Titeln versehen):

- **MIDWIFERY**: birth deld safe morn receivd calld left cleverly pm labour fine reward arivd infant expected recd shee born patient
- **CHURCH**: meeting attended afternoon reverend worship foren mr famely performd vers attend public supper st service lecture discoarst administred supt
- **DEATH**: day yesterday informd morn years death ye hear expired expird weak dead las past heard days drowned departed evinn
- **GARDENING**: gardin sett worked clear beens corn warm planted matters cucumbers gatherd potatoes plants ou sowd door squash wed seeds
- **SHOPPING**: lb made brot bot tea butter sugar carried oz chees pork candles wheat store pr beef spirit churnd flower
- **ILLNESS**: unwell mr sick gave dr rainy easier care head neighbor feet relief made throat poorly takeing medisin ts stomach

3.1 Blevins: Text Analysis of Martha Ballard‘s Diary

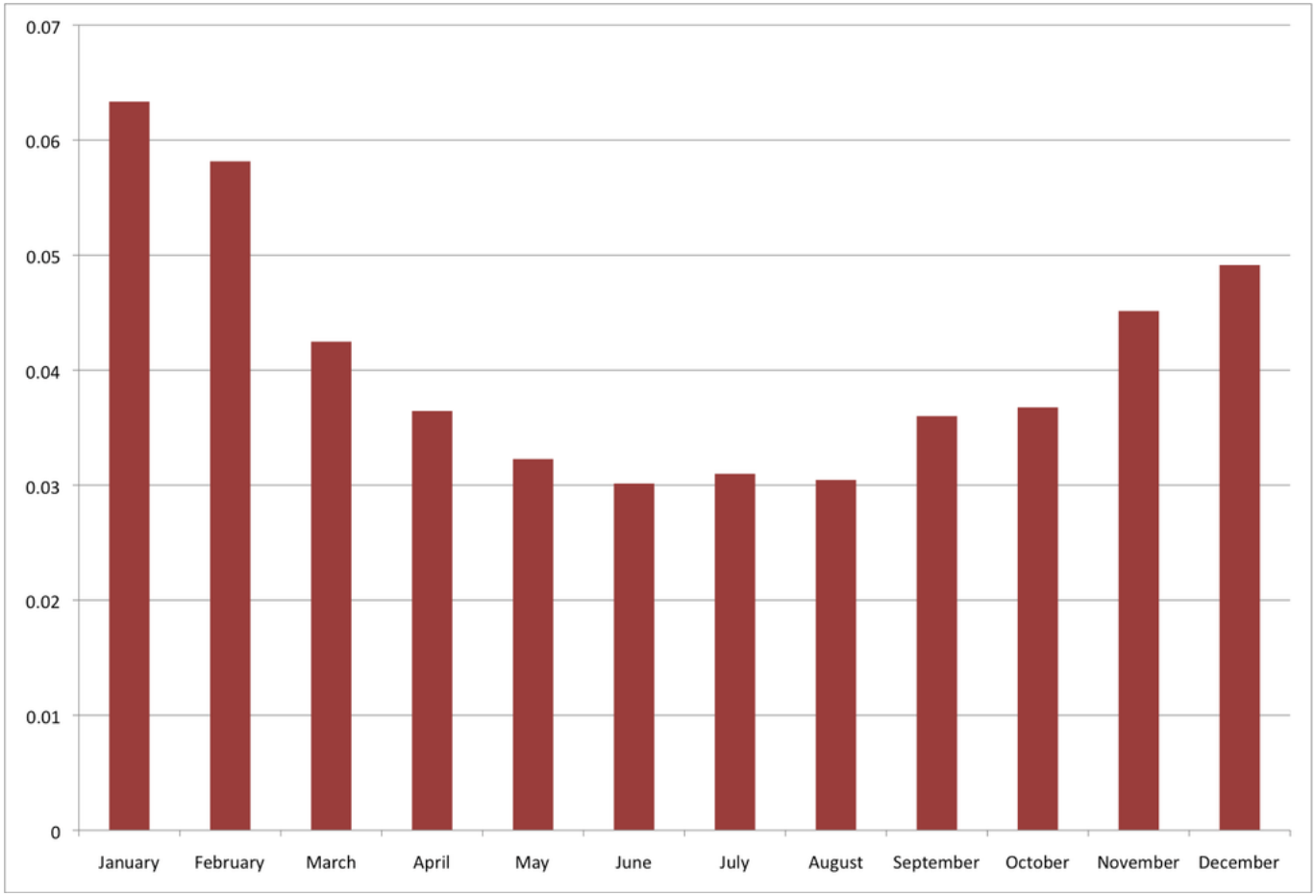
Blick in ein Dokument (Tagebucheintrag vom 28. November 1795):

“*Clear and pleasant. I am at mr Pages, had another fitt of ye Cramp, not So Severe as that ye night past. mrss Pages illness Came on at Evng and Shee was Deliverd at 11h of a Son which waid 12 lb. I tarried all night She was Some faint a little while after Delivery.*”

→dominantes Topic **MIDWIFERY** (passt)

3.1 Blevins: Text Analysis of Martha Ballard‘s Diary

Blevins: „*The power of topic modeling really emerges when we examine thematic trends across the entire diary.*“

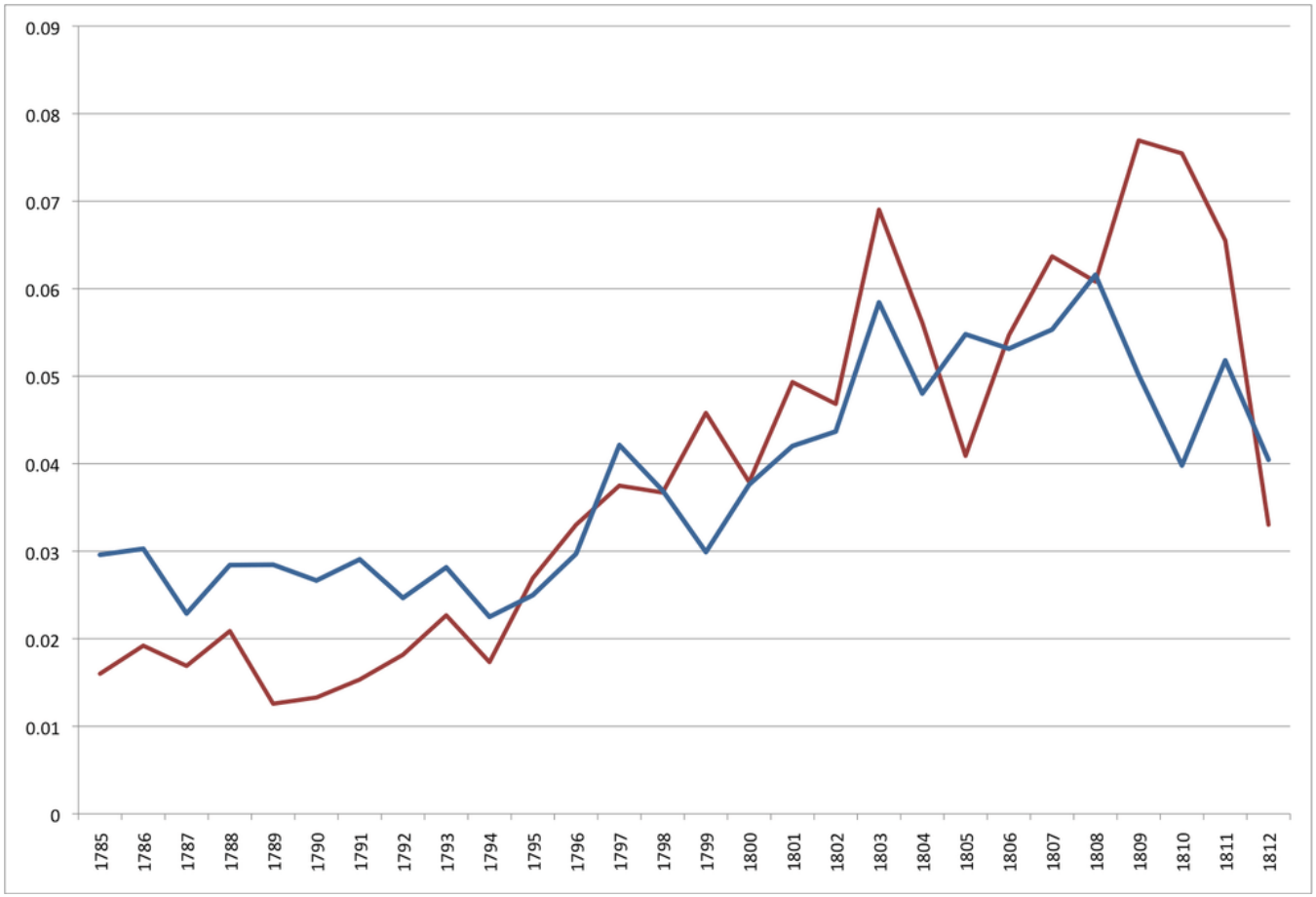


COLD WEATHER-Topic

cold windy chilly snowy air...

3.1 Blevins: Text Analysis of Martha Ballard‘s Diary

zwei **HOUSEHOLD**-Topics über die Zeit



Warum Anstieg am Ende?

3.1 Blevins: Text Analysis of Martha Ballard‘s Diary

Blevins Fazit:

„*I am absolutely intrigued by the potential for topic modeling in historic source material. In many ways, it seems that Martha Ballard’s diary is ideally suited for this kind of analysis. Short, content-driven entries that usually touch upon a limited number of topics appear to produce remarkably cohesive and accurate topics.*“

3.2 CLiGS: Topics im Textverlauf

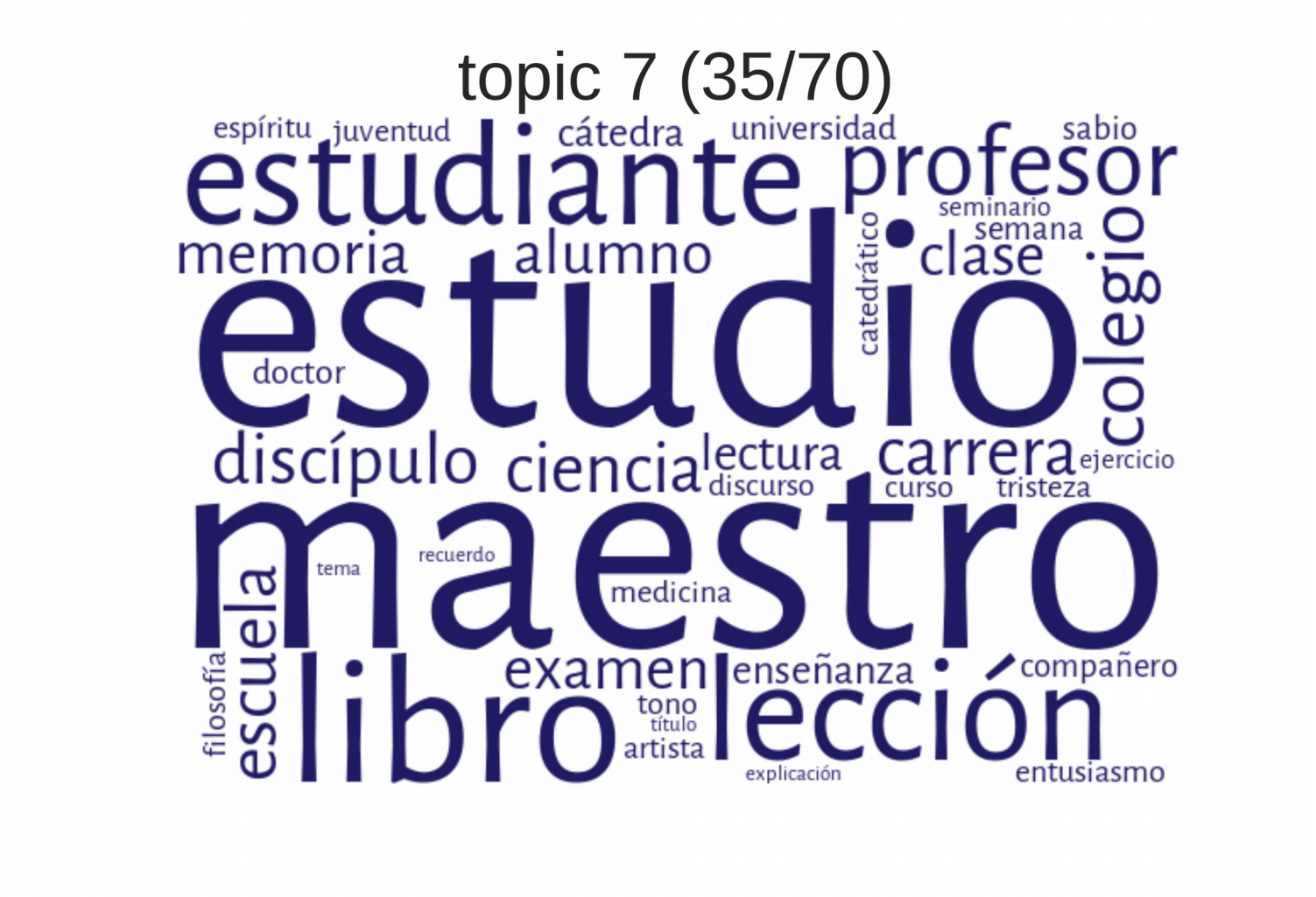
- Untersuchung von Topics im Textverlauf von Untergattungen des spanischen und hispanoamerikanischen Romans (1850-1930)
  - Welche Beziehungen gibt es zwischen Topics und Untergattungen?
  - Welche Beziehungen gibt es zwischen Topics und dem Textverlauf?
  - Gibt es untergattungsspezifische Topics im Textverlauf?

3.2 CLiGS: Topics im Textverlauf

- Datengrundlage: 130 Romane aus Spanien, Argentinien, Kuba und Mexiko; rund 7,3 Mio. Tokens
- Untergattungen: sentimental, historisch, politisch-sozial, „subjektiv“
- Topic Modeling mit Python (Mallet + TMW)
- 70 Topics, 6 „Bins“ pro Text

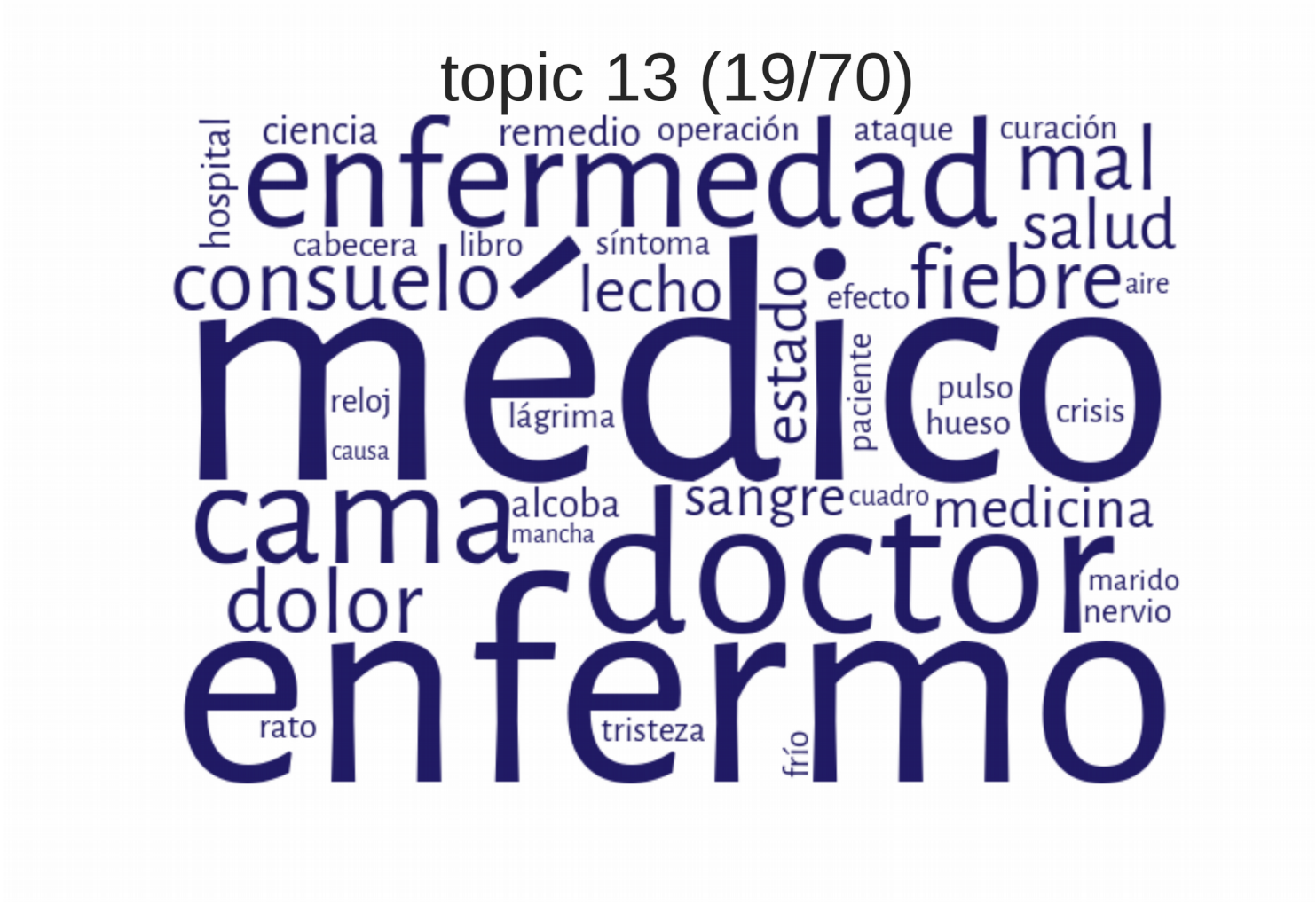
3.2 CLiGS: Topics im Textverlauf

Topic: *Lehrer-Studium-Schüler (Schule)*

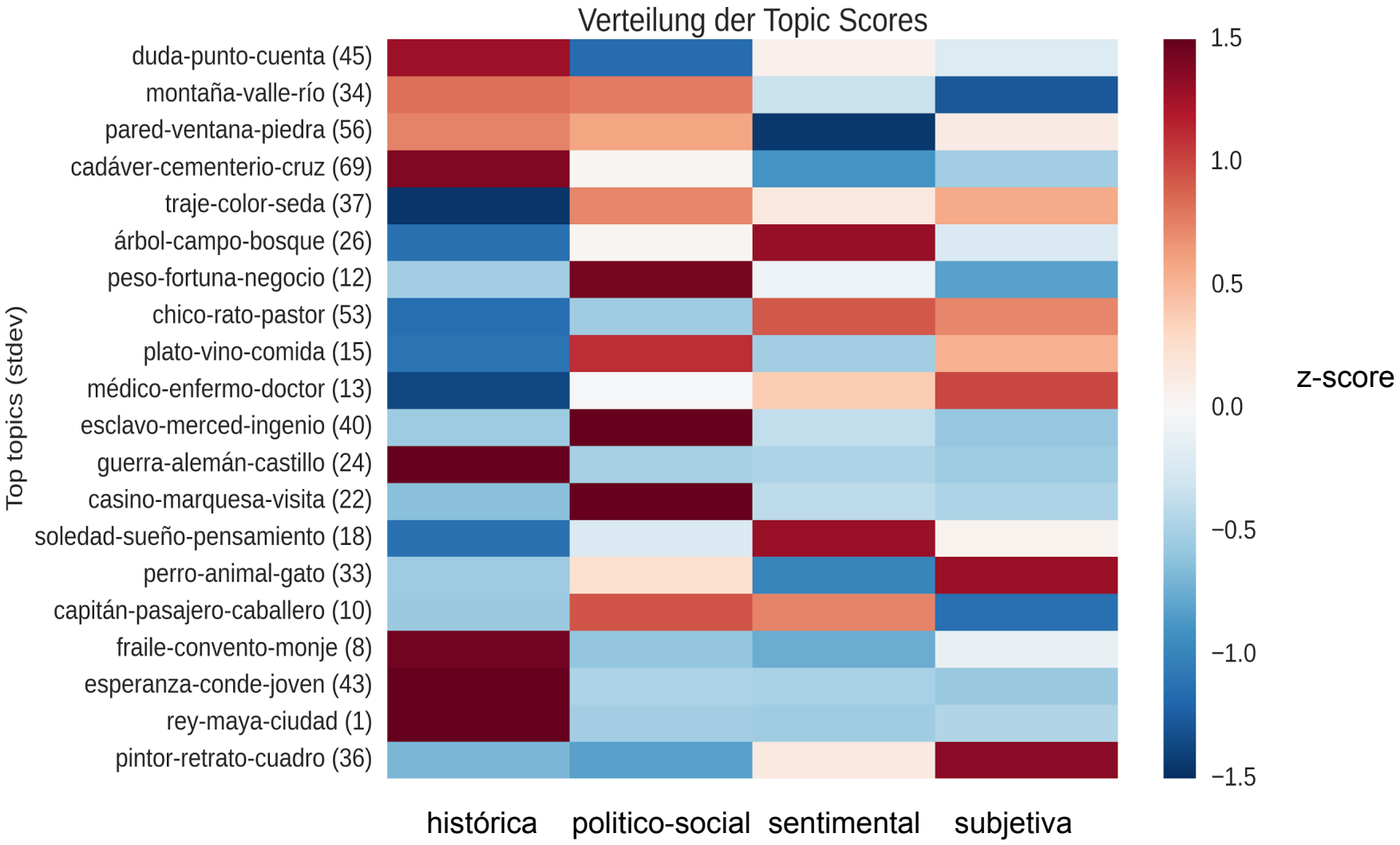


3.2 CLiGS: Topics im Textverlauf

Topic: *Arzt-Kranker-Doktor (Krankheit)*

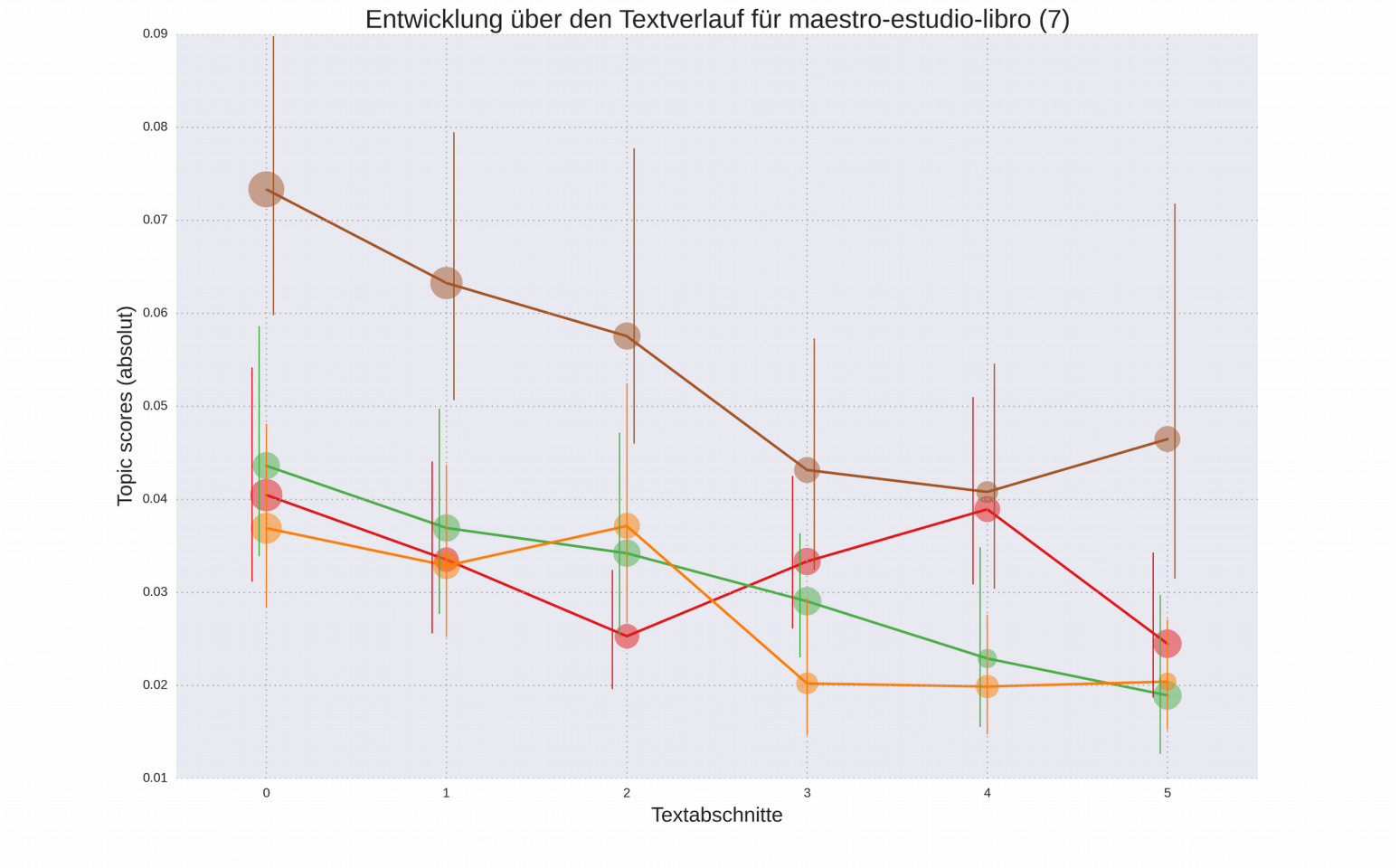


3.2 CLiGS: Topics im Textverlauf – Distinktive Topics für Untergattungen



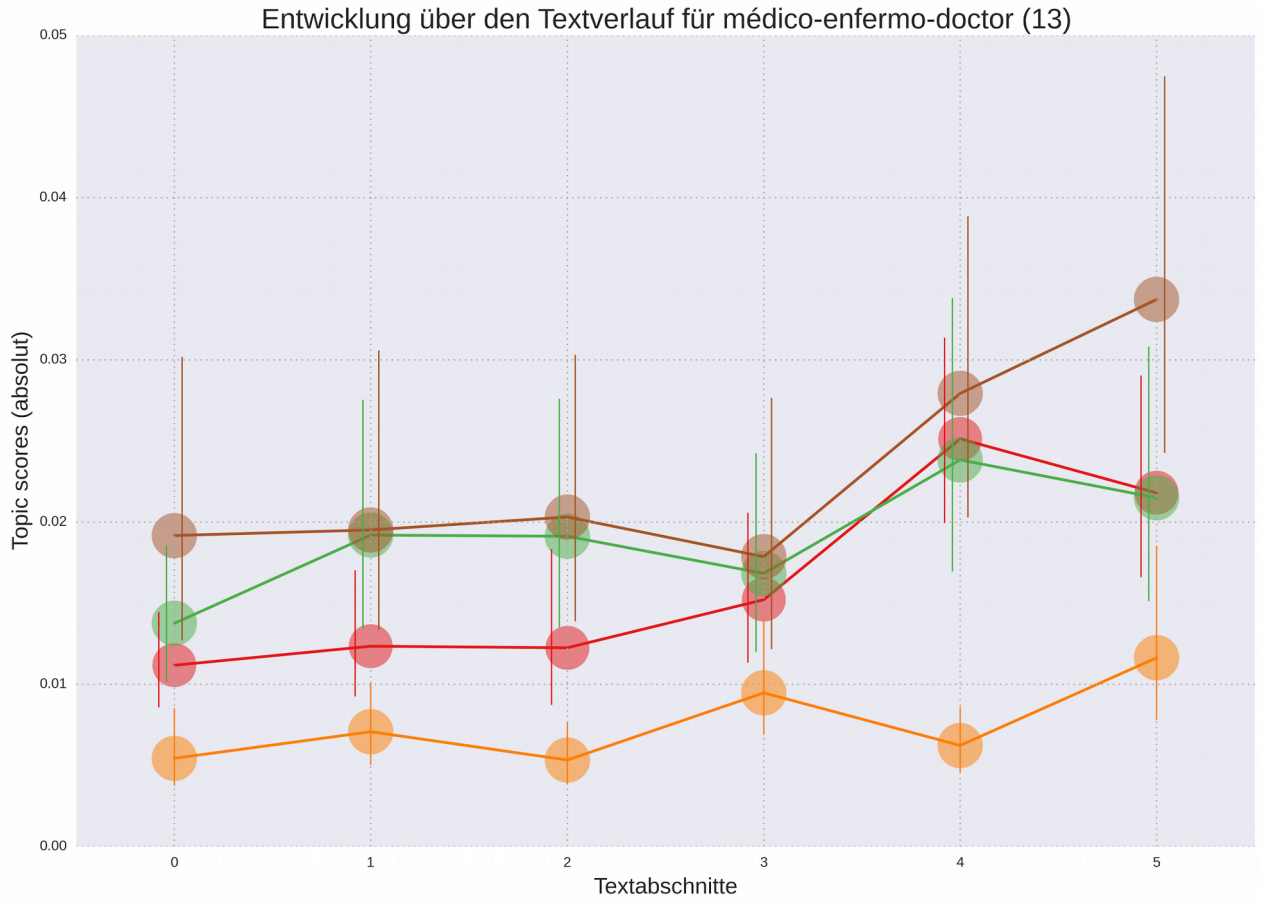
3.2 CLiGS: Topics im Textverlauf – Distinktive Topics für Untergattungen

Textverlauf nach Untergattung: Schule



3.2 CLiGS: Topics im Textverlauf – Distinktive Topics für Untergattungen

Textverlauf nach Untergattung: Krankheit



3.2 CLiGS: Topics im Textverlauf – Distinktive Topics für Untergattungen

- Ergebnisse:
  - Topics gefunden, die für den Textanfang oder das Textende einer bestimmten Roman-Untergattung typisch sind
  - aber:
    - solche Befunde eher die Ausnahme
    - Untergattungen ggf. differenzierter betrachten
    - es sollten alle Ergebnisse des Topic Models einbezogen werden, nicht nur die positiven
    - Rolle von Korpusgröße und -zusammenstellung
    - Einfluss von Parametern des Topic Models (z.B. Topiczahl, Segmentlänge)

Zur Interpretierbarkeit der Ergebnisse

4. Zur Interpretierbarkeit der Ergebnisse

- Topics vs. Themen
  - Thema: das, worum es in einem Text, Diskurs, Gespräch geht
  - Topic: Wahrscheinlichkeitsverteilung über ein Wort-Vokabular

4. Zur Interpretierbarkeit der Ergebnisse

- Begriffe und Konzepte im Topic Model:
  - Korpus: Sammlung von „Dokumenten“
  - Dokument: Sammlung von „Wörtern“ (*bag-of-words* – Modell!)
  - Wort: Token
  - Topic: Wahrscheinlichkeitsverteilung über ein Wort-Vokabular

    → in der Praxis steht nicht fest, was ein Wort ist und was ein Dokument!

4. Zur Interpretierbarkeit der Ergebnisse

Ein Topic Model *kann* Topics hervorbringen, die nach Themen aussehen.

Es können aber auch andere Arten semantischer Relationen sichtbar werden: Motive, Redeweisen, …

Oder es ist kein semantischer Zusammenhang erkennbar.

Bei einer Interpretation sollten möglichst alle Ergebnisse des Topic Models berücksichtigt werden.

4. Zur Interpretierbarkeit der Ergebnisse

Ein Topic ist unter Umständen nicht mehr unmittelbar auf einzelne Texte zu beziehen. Ein Topic Model bezieht sich vor allem auf die Text*sammlung*.

Ciro B. Ceballos, *Un adulterio* (1901), Kurzroman, Mexiko

Wort	Gewichtung	Vorkommen im Text
vida (Leben)	193	20
campo (Feld)	179	5
estancia (Landgut)	152	1
año (Jahr)	151	5
amigo (Freund)	137	11
mate (Mate-Tee)	129	0
cuero (Leder)	109	0
sargento (Feldwebel)	86	0
pata (Pfote)	84	0
cuchillo (Messer)	76	0

4. Zur Interpretierbarkeit der Ergebnisse

- Weitere Aspekte:
  - Zufälligkeit der Ergebnisse?
  - Evaluation von Topic Models
    - Was wird erwartet?
      - z.B. semantische Kohärenz von Topics
      - dass Topic Models die Dokumente „gut“ beschreiben
      - (dass das Modell sich gut für andere Aufgaben einsetzen lässt)
    - Wie kann das überhaupt gemessen werden?

Fazit

- Topic Modeling ist relativ einfach einzusetzen, es fehlt derzeit aber vor allem noch an Werkzeugen, welche die Modellierung selbst um Vor- und Nachbereitung ergänzen.
- Eine Topic Modeling-Analyse ist vor allem *distant reading.* Sie kann der Erschließung großer Textsammlungen dienen, sie kann einen neuen Blick auf Texte ermöglichen.
- Ein Topic Model ist vor dem Hintergrund der Methode zu sehen. Wie die Ergebnisse an traditionelle Fragen angebunden werden können, ist noch weitgehend offen.

Fazit

„*As Stephen Ramsay argues in* Reading Machines*, using algorithms need not propel us towards applying an ersatz scientific and scientistic evidentiary standard to literary interpretation, but rather should reveal and perhaps help amplify our already part-algorithmic literary-critical reading practices, the regular sets of protocols and procedures of analog literary criticism with which we are very—perhaps sometimes too—familiar*“

(Rachel Sagner Buurma: The Fictionality Of Topic Modeling: Machine Reading Anthony Trollope's Barsetshire Series)

Referenzen

- Blei, David M. (2006): „Probabilistic Topic Models“. In: Communications of the ACM, 55 (4), S. 77-84. http://www.cs.princeton.edu/~blei/papers/Blei2012.pdf
- Blevins, Cameron (2010): „Topic Modeling Martha Ballard‘s Diary“. [Blog Posts] http://www.cameronblevins.org/posts/topic-modeling-martha-ballards-diary/
- Buurma, Rachel Sagner (2015): „The Fictionality Of Topic Modeling: Machine Reading Anthony Trollope‘s Barsetshire Series“. Big Data and Society. Bd. 2, Nr. 2. http://works.swarthmore.edu/fac-english-lit/286
- Schöch, Christof et al. (2016): „Topic, Genre, Text. Topics im Textverlauf von Untergattungen des Spanischen und Hispanoamerikanischen Romans (1880-1930)“. Leipzig: nisaba verlag, S. 235-238. http://dhd2016.de/boa.pdf