Universität Münster

Quantitative Textanalyse

Sitzung 8: Datenanalyse – Unsupervised Topic Models

Mirko Wegemann

Universität Münster
Institut für Politikwissenschaft

27. November 2024

Logistik

- **letzte Woche**: Vorbereitung eines Datensatzes in eine numerische Repräsentation (**d**ocument **f**eature **m**atrix)
- außerdem: wie können wir unseren Datensatz komprimieren und was sind mögliche Fallstricke dabei (`preText`-Package)?
- **diese Woche**: unsupervised topic models

Unsupervised methods

- Manchmal wollen wir Text **klassifizieren**, wissen im Vorfeld aber nicht genau in welche Kategorien
- in diesen Fällen können wir auf **unsupervised**-Ansätze zurückgreifen, welche "a class of methods that learn underlying features of text without explicitly imposing categories of interest" beschreiben (Grimmer and Stewart 2013, p. 281)
- wie Grimmer and Stewart (2013) aufzeigen, lassen sich diese noch in zwei Ansätze aufteilen
  1. **Clustering** (diese Woche)
  2. Scaling (nächste Woche)

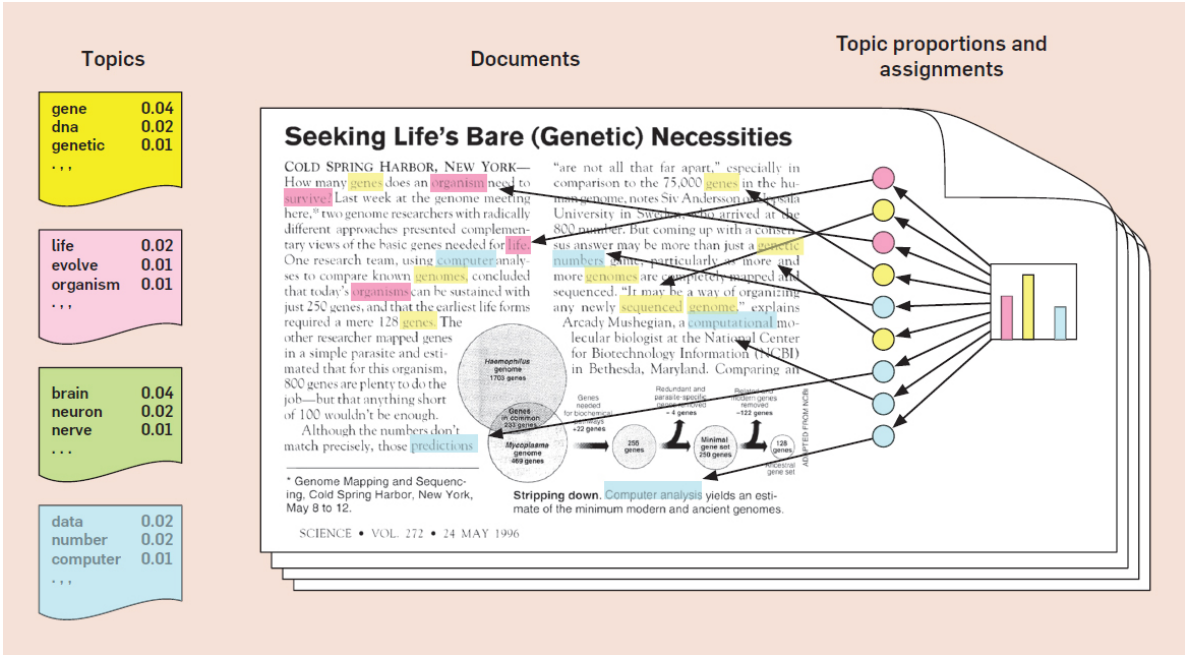
Topic models

- **Topic Models** sind eine Form des **Clusterings**
- Im Vergleich zu Modellen mit single membership (jeder Text wird **einer** Kategorie zugeordnet) sind Topic Models **mixed-membership models**
- Annahme: jeder Text verwendet Vokabular aus verschiedenen Themen
- Datenanforderung: große Textkorpora, u.U. pre-processed (s. letzte Woche)

Grundidee von Topic Models I

- Dokumente bestehen aus Wörtern, deren Kombination eine **(latente) semantische Bedeutung** repräsentiert
- Idee: Autor*innen eines Textes schreiben über ein Thema k und verwenden dabei Begriffe, die mit diesem assoziiert sind
- Abhängig von der Zusammensetzung der Wörter hat jedes Dokument eine **Wahrscheinlichkeit p, zum Thema k zu gehören**
- Zentrale Techniken sind die **L**atent **S**emantic **A**nalysis (**LSA**) und die **L**atent **D**irichlet-**A**llocation (**LDA**)

Grundidee von Topic Models II



Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM, 55*(4), 77–84. https://doi.org/10.1145/2133806.2133826

Grundidee von Topic Models III

- **LDA** ist ein probabilistisches Modell, in dem jeder Begriff t eine bestimmte Wahrscheinlichkeit hat, zu Thema k zu gehören, und jedes Dokument d aus mehreren k besteht
  - Hauptsächlich für Topic Modelling verwendet
  - **Generatives Modell**, bei dem Themen zunächst zufällig Dokumenten zugewiesen werden; jedes Wort hat eine Wahrscheinlichkeit, zu Thema k zu gehören
  - **Iterativer** Prozess, bei dem schrittweise die Log-Likelihood reduziert wird (je kleiner, desto besser) [im R-Output könnt ihr den iterativen Prozess in Form von "expectation"-"maximization" beobachten]
  - Anzahl der k muss im Voraus definiert werden; es gibt hierbei keine allgemeingültigen Richtlinien, aber meist: Je größer der Korpus, desto mehr Themen enthält er)

...mehr zu LDA (Blei, Ng, and Jordan 2003)

Grundidee von Topic Models IV

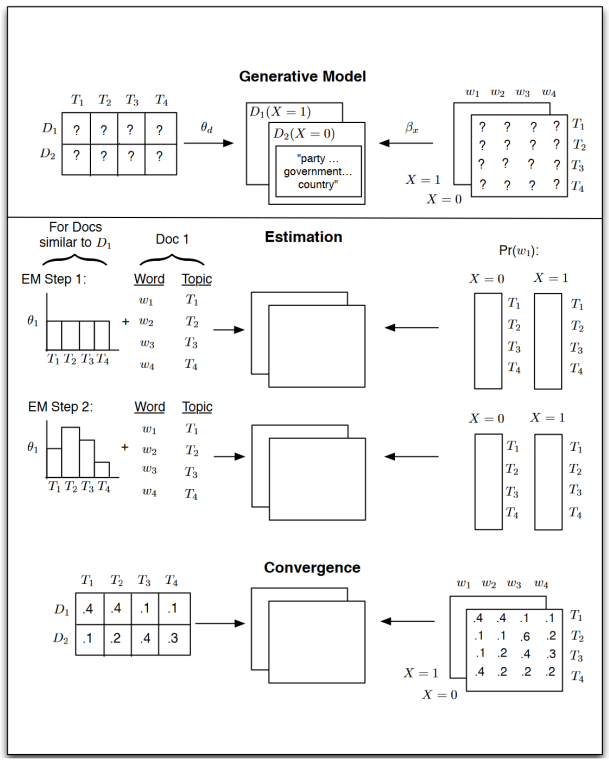


Figure: Prozess des Topic Models (Roberts et al. 2019, p. 4)

Topic Models in R I

Es gibt verschiedene Pakete, um Topic Models in `R` zu nutzen. Hierzu gehören u.a. `lda` und `stm`.

Structural Topic Model (`stm`)

- Sehr ähnlich zu LDA
- Neuheit: Wir können **Meta-Informationen** zum Topic Model hinzufügen, um potentiell aussagekräftigere Ergebnisse zu erhalten
  1. **prevalence** der Themen: Kovariaten, die die Häufigkeit eines Themas beeinflussen (im Winter hat das Wort "Eis" evtl. eine andere Nutzung als im Sommer)
  2. **content** der Themen: Kovariaten, die beeinflussen, wie über ein Thema gesprochen wird (Parteien haben unterschiedliches Verständnis von Wirtschaft und nutzen womöglich unterschiedliches Vokabular)

Topic Models in R II

- `stm` erlaubt es uns technisch auch, $k = 0$ anzugeben, wobei k automatisch bestimmt wird; *Achtung*: Das bedeutet nicht, dass $\hat{k}$ das wahre k ist
- wir können weitere Parameter auswählen, u.a.
  - `emtol=value`: gibt an, wann die Optimierung beendet werden soll
  - `seed=value`: für replizierbare Ergebnisse
  - `init_type=value`: wir können zwischen verschiedenen Initialisierung wählen; "spectral" ist am besten replizierbar
  - `LDAbeta=F`: SAGE-Updating der Topics

Topic Models in R III

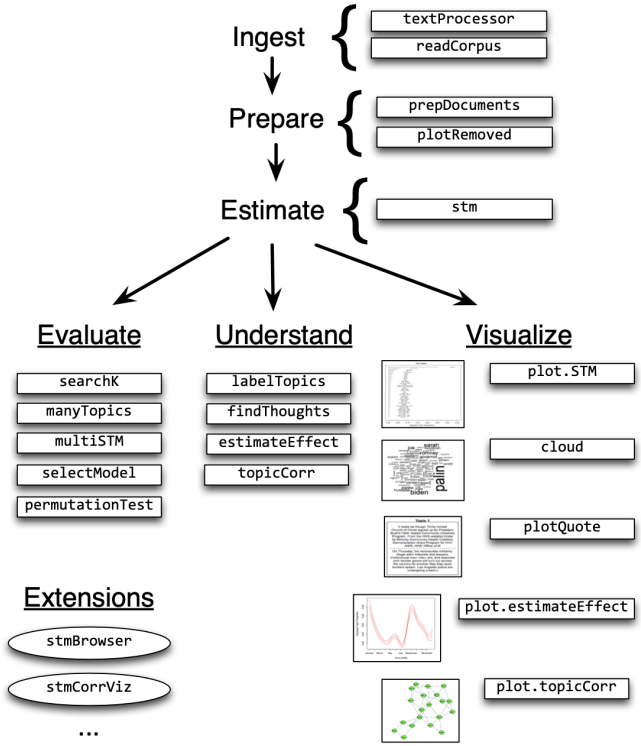


Figure: Funktionen im `stm`-Package (Roberts et al. 2019, p. 5)

Topic Models in R IV

Schritte in R

1. **Vorbereitung** (wie zuvor: vom data.frame zum corpus, zum tokens-Objekt, zur data frequency matrix)
2. **Schätzung** des Topic Models
3. **Interpretation**
4. **Validierung**

Is the Left-Right Scale a Valid Measure of Ideology? I

- **Forschungsfrage**:
- **Argument**:
- **Daten und Analyse**:
- **Ergebnisse**:
- **Implikationen**:

Is the Left-Right Scale a Valid Measure of Ideology? II

- **Forschungsfrage**: Welche Bedeutung(en) hat das Links-Rechts-Schema für Bürger*innen?
- **Argument**: Menschen verbinden unterschiedliche Assoziationen mit abstrakten Konzepten (Ideologien, Werte, etc.)
- **Daten und Analyse**: Allbus 08 (Allgemeine Bevölkerungsumfrage der Sozialwissenschaften); unsupervised topic model mit offenen Antworten
- **Ergebnisse**: Teilnehmende haben verschiedene Auffassungen davon, was *rechts* und *links* bedeutet (je nach eigener Einstufung auf der Skala und sozioökonomischen Hintergrund)
- **Implikationen**: Antwortverhalten hängt systematisch davon ab, wie Menschen theoretische Konzepte interpretieren

Theorie

Modell des ‘survey response process’, nach dem Individuen vier Schritten folgen, um eine Antwort abzugeben

1. **Verständnis**
2. Abrufen von Informationen
3. Beurteilung
4. Antwort

→ Fokus auf dem Verständnis einer Frage

Hypothesen

H_1 Menschen haben ein unterschiedliches Verständnis der Links-Rechts-Skala

H_2 Das Verständnis der Links-Rechts-Skala wirkt sich auf die Links-Rechts-Selbsteinstufung aus.

H_3 Das Verständnis der Links-Rechts-Skala ist systematisch von Merkmalen der Befragten abhängig.

Daten und Modell I

Hauptvariablen:

1. **Links-Rechts-Selbsteinstufung** ("Here we have a scale that runs from left to right. When you think of your own political views where would you position yourself on this scale?")
2. **Verständnis von Links-Rechts** ("...would you please tell me what you associate with the term 'left'? and would you please tell me what you associate with the term 'right'?")

Daten und Modell II

Zwei Analyseschritte:

1. LDA; zur Evaluierung: Lift-Scores; zur Erinnerung: "Lift weights words by dividing by their frequency in other topics" (Roberts et al. 2019, p. 13)
2. multivariate Regressionsmodelle

Daten und Modell III

Für die Replikation interessant:

- Anzahl der Topics $k = 4$
- Stopwörter entfernt
- Satzzeichen entfernt
- nur Wörter, die in mindestens fünf Antworten vorkommen
- SAGE-Spezifizierung des Algorithmus

Ergebnisse I

- Je nach politischer Einstellung assoziieren Menschen unterschiedliche Begriffe mit "Links" und "Rechts"
- "Policies" und "Values" verbinden jeweils Leute mit linker politischen Einstellung mit "links"; "Parteien" assoziieren rechtseingestellte Befragten mit "rechts"
- sozioökonomischer Hintergrund entscheidend für Assoziationen; z.B. heben höher gebildete Menschen und Ostdeutsche "Werte" bei "links" hervor

Fragen I

"We cannot be sure whether this difference is due to variation in their associations with the concept 'left' or due to a real difference in their political ideology"

- Was denkt ihr?

Fragen II

Die Allbus-Studie fragt zuerst nach der Links-Rechts-Selbsteinstufung der Befragten, bevor diese nach ihren Assoziationen zu "Links" und "Rechts" befragt werden.

- Was ist der Vorteil von diesem Vorgehen für die Studie? Was ist der Nachteil?

Und jetzt in R

Validierung

Validierung kann in unterschiedlichem Maße stattfinden (Überblick über die Konsequenzen in Bernhard et al. (2023)):

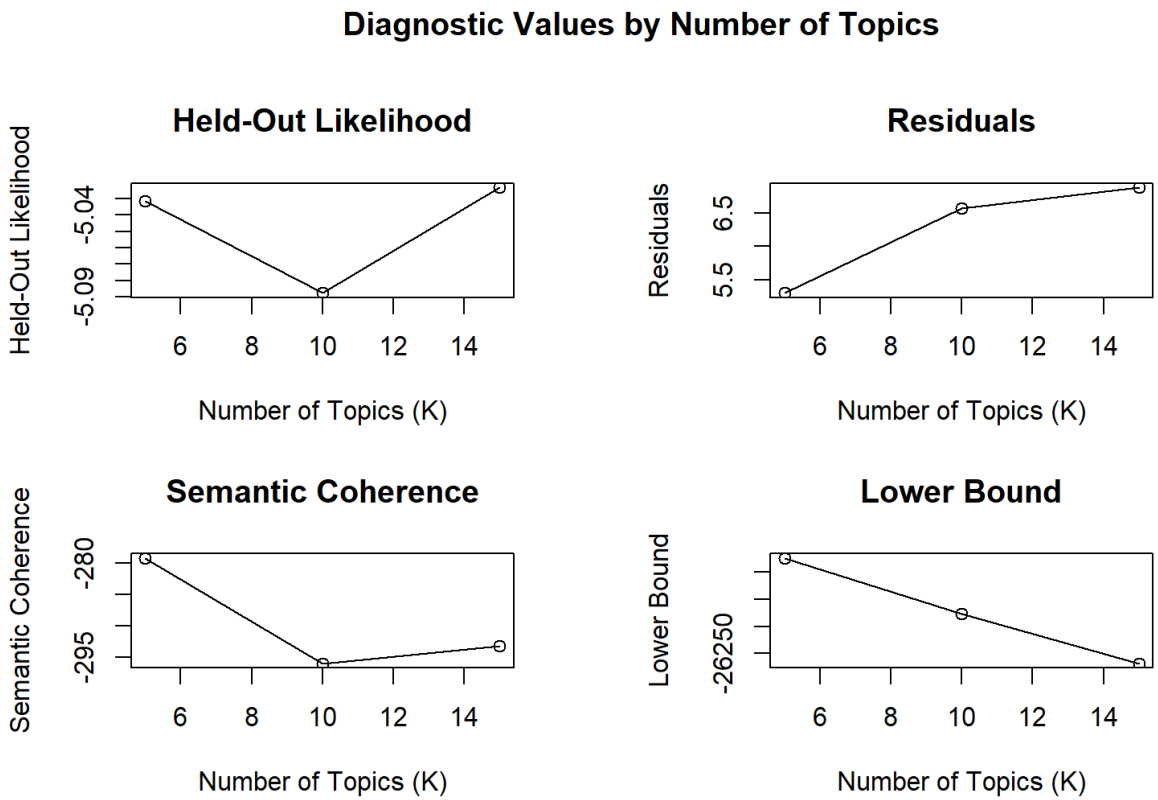
1. **semantische** Validierung (= Interpretation der Topics durch Häufigkeitsmaße)
2. **statistische** Validierung (= geringster statistischer Fehler)
3. **externe** Validierung nach Quinn et al. (2010) (= Vorhersagekraft)
4. anhand **manueller** Kodierung

Statistische Validierung

Wir können die `searchK()`-Funktion von stm verwenden, um die ideale Anzahl an Themen k basierend auf Likelihood, Untergrenze, Residuen und semantischer Kohärenz zu bestimmen.

```
1  # stm-Format fuer diese Funktion notwendig
2  stm_left <- convert(dfm_left, "stm")
3
4  # Vektor mit Zahl der Topics
5  K <- c(5,10,15)
6
7  best_k <- searchK(stm_left$documents,
       stm_left$vocab, K, seed=421, emtol=0.001,
       LDAbeta = F)
```

Statistische Validierung



Statistische Validierung II

Alternativ können wir auch, wenn die Anzahl an Topics klar ist, das Modell auswählen, welches die besten Parameter hat (`selectModel()`)

```
1    best_m <- selectModel(stm_left$documents,
         stm_left$vocab, 4, runs=10, seed=421,
         emtol=0.001, init.type = "Spectral", LDAbeta
         = F)
2    plotModels(best_m)
```

Weiterentwicklung von Topic Models

- verfügbar im R-Paket `topicmodels`
- correlated topic models: erlauben Korrelationen zwischen latenten Themen; in R ausführbar mit `CTM` (siehe für eine Anwendung und Blei and Lafferty (2005) für methodischen Hintergrund)
- topic models mit user-input: semi-supervised Aproach; dem Topic Model werden vorab spezifizierte Schlüsselwörter vorgegeben; in R ausführbar mit `textmodel_seededlda` (siehe für eine Anwendung und Watanabe and Baturo (2024) für methodischen Hintergrund)
- `top2vec`: Topic Models, welche auf Embeddings basieren (mehr dazu im Verlauf des Seminars)

Ausblick

- nächste Woche widmen wir uns semi-supervised Approaches
- Fokus auf Scaling politischer Texte
- **Literatur**:
Watanabe, K. (2021). Latent Semantic Scaling: A Semisupervised Text Analysis Technique for New Domains and Languages. *Communication Methods and Measures*, *15*(2), 81–102. https://doi.org/10.1080/19312458.2020.1832976
Zollinger, D. (2024). Cleavage Identities in Voters' Own Words: Harnessing Open-Ended Survey Responses. *American Journal of Political Science*, *68*(1), 139–159. https://doi.org/10.1111/ajps.12743

Noch Fragen?

Literatur I

Bernhard, J., Teuffenbach, M., & Boomgaarden, H. G. (2023). Topic Model Validation Methods and their Impact on Model Selection and Evaluation. *Computational Communication Research*, *5*(1), 1. https://doi.org/10.5117/CCR2023.1.13.BERN

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77–84. https://doi.org/10.1145/2133806.2133826

Blei, D. M., & Lafferty, J. D. (2005). Correlated topic models. *Proceedings of the 18th International Conference on Neural Information Processing Systems*, 147–154.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *J. Mach. Learn. Res.*, *3*(null), 993–1022.

Literatur II

Grimmer, J., & Stewart, B. M. (2013). Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis*, *21*(3), 267–297. https://doi.org/10.1093/pan/mps028

Quinn, K. M., Monroe, B. L., Colaresi, M., Crespin, M. H., & Radev, D. R. (2010). How to Analyze Political Attention with Minimal Assumptions and Costs. *American Journal of Political Science*, *54*(1), 209–228. https://doi.org/10.1111/j.1540-5907.2009.00427.x

Roberts, M. E., Stewart, B. M., & Tingley, D. (2019). Stm: An R Package for Structural Topic Models. *Journal of Statistical Software*, *91*, 1–40. https://doi.org/10.18637/jss.v091.i02

Watanabe, K. (2021). Latent Semantic Scaling: A Semisupervised Text Analysis Technique for New Domains and Languages. *Communication Methods and Measures*, *15*(2), 81–102. https://doi.org/10.1080/19312458.2020.1832976

Literatur III

Watanabe, K., & Baturo, A. (2024). Seeded Sequential LDA: A Semi-Supervised Algorithm for Topic-Specific Analysis of Sentences. *Social Science Computer Review*, *42*(1), 224–248. https://doi.org/10.1177/0894439231178605

Zollinger, D. (2024). Cleavage Identities in Voters' Own Words: Harnessing Open-Ended Survey Responses. *American Journal of Political Science*, *68*(1), 139–159. https://doi.org/10.1111/ajps.12743