Mini-Workshop (Structural) Topic Models

Marko Bachl

Sommersemester 2020 | IJK Hannover

Contents

Chapter 1

Überblick

1.1 Inhalt des virtuellen Mini-Workshops

- In diesem Mini-Workshop erläutere ich das praktische Vorgehen einer Datenanalyse mit *Structural Topic Models*. Wir behandeln die folgenden Schritte im Analyseprozess:
  - Schätzen eines ersten Modells
  - Modellvergleich zur Auswahl eines geeigneten Modells
  - Interpretation der Topics im finalen Modell
  - Darstellung der Ergebnisse
  - Weitere Analysen
    - Identifikation verwandter Themen
    - Zusammenhänge der Themenprävalenz mit Kovariaten.
- Wir verwenden das Paket `{stm}` (Roberts et al., 2019) zum Schätzen von Topic Models. Für die Variante der *Structural* Topic Models und die Implementation in diesem Paket sprechen *für mich* die folgenden Gründe
  - Gute Integration mit *R* und Paketen, die ich für die Arbeit mit Text-Daten verwende (insbesondere `{quanteda}` und `{tidytext}`)
  - Gute ergänzende Pakete zur Arbeit mit den Modellen (insbesondere `{stminsights}`)
  - Vergleichsweise schnelle Modellschätzung auch mit großen Datensätzen
  - Direktes Schätzen von Zusammenhängen von Topics mit Kovariaten
  - Initialisieren der Modellschätzung mit dem Spectral Algorithmus
  - Recht weit verbreitet in einem Feld, in dem ich viel lese (Politische Kommunkation nach einem weitem Verständnis)
- Die Darstellung basiert auf einer Analyse, die ich gemeinsam mit Elena Link durchgeführt habe. Wir untersuchten, wie das Thema Impfen in Online-Foren für Eltern diskutiert wurde. Wir verwenden aber nur einen *nicht repräsentativen* Ausschnitt aus dem Material, um die notwendige

Rechenleistung und -zeit zu verringern.
  - Einen Preprint zur Analyse könnt ihr hier lesen: Vaccine-related Discussions in Online Communities for Parents. A Quantitative Overview.
  - Die Dokumentation zur Studie ist hier verfügbar: https://bachl.github.io/vaccine_discussions/. Daten und Analyse-Skripts gibt es im OSF. In diesem Material werde auch die Datenerhebung mittels Web-Scraping und die Datenaufbereitung erläutert. Diese Inhalte sind *nicht* Teil dieses Workshops. Wenn ihr Fragen dazu habt, dürft ihr sie natürlich stellen.

1.2 Welche Inhalte wir *nicht* behandeln

- Auch wenn das im direkten Vergleich mit dem Parallel-Angebot zu Panel Data Analysis (meine Ausführlichkeit dort sind ein Grund für die spätere Lieferung dieser Materialien) enttäuschend sein mag: Die Inhalte in diesem Mini-Workshop entsprechen in ihrem Umfang wirklich nur dem, was ich zu Beginn des Digital-Semesters geplant und angekündigt hatte. Der Mini-Workshop ersetzt keine tiefer gehende Einarbeitung in die Methode, sondern ist als ein Einstieg zu verstehen.

- Wir behandeln hier keine theoretischen, statistischen oder auf die Software-Implementierung der Modellschätzung bezogenen Fragen. Die Grundlagen dazu können aus den Texten im LMS entnommen werden (Maier et al., 2018; Roberts et al., 2019).

- Es gibt neben `{stm}` viele andere Implementationen in *R* und ihn anderer Software. Gefühlt gibt es alle 6 Monate eine neue Variante von Topic Models, alle 3 Monate eine neue Implementierung und jeden Monat ein Paket mit zusätzlichen Tools für die Arbeit mit Topic Models. Meine Entscheidung für `{stm}` ist keine informierte Entscheidung gegen andere Varianten, Implementierungen und Tools. Dieser Workshop ist keine Aufforderung, ausschließlich `{stm}` zu nutzen. Informiert euch gegebenenfalls selbst über Software-Lösungen, die für eure Bedürfnisse geeignet sind.

- Dieser Mini-Workshop ist kein *R*-Tutorial. Wenn ihr Interesse habt, *R*-Kenntnisse zu erwerben und zu vertiefen, empfehle ich R4DS.

- Dieser Mini-Workshop ist keine allgemeine Einführung in die computergestützte Inhaltsanalyse. Wenn ihr allgemein mit *R* arbeiten möchtet, empfehle ich zu diesem Thema die Einführung von Cornelius Puschmann.

1.3 Aufbau des Workshops

- Inhaltlicher Aufbau: Siehe Kapitel-Gliederung

Material

- Dieses Dokument + R Skripte: (Hoffentlich) mehr oder weniger selbsterklärendes Material
  - Kuratierte Form ist dieses HTML-Dokument
  - Es gibt auch ein PDF, das ich aber nicht formatiert habe
- Daten: Ein Ausschnitt auf den Daten der oben genannten Beispielstudie. Eine genauere Beschreibung folgt im nächsten Abschnitt.
- Screencast: Zu einigen Analyseschritten stelle ich Screencasts zur Verfügung. Diese sind größtenteils ergänzend gedacht. Bis auf wenige Ausnahmen sollte das schriftliche Material selbsterklärend sein.
- Übungen: Zu einigen Analysen gibt es Übungsaufgaben.
  - XXX

Pakete

Wir verwenden die folgenden Pakete

```
if (!require("pacman")) install.packages("pacman")
pacman::p_load(tidyverse, stm, stminsights, tidytext, quanteda, lubridate)
theme_set(theme_bw())  # ggplot theme

tibble(package = c("R", sort(pacman::p_loaded()))) %>% mutate(version = map_chr(package,
    ~as.character(pacman::p_version(package = .x)))) %>% knitr::kable()
```

package	version
R	3.6.2
dplyr	0.8.4
forcats	0.4.0
ggplot2	3.3.1
lubridate	1.7.4
pacman	0.5.1
purrr	0.3.3
quanteda	2.0.0
readr	1.3.1
stm	1.3.5
stminsights	0.3.0
stringr	1.4.0
tibble	2.1.3
tidyr	1.0.2
tidytext	0.2.3
tidyverse	1.3.0

Chapter 2

Beispiel-Daten und Aufbereitung

2.1 Laden der Daten und Übersicht

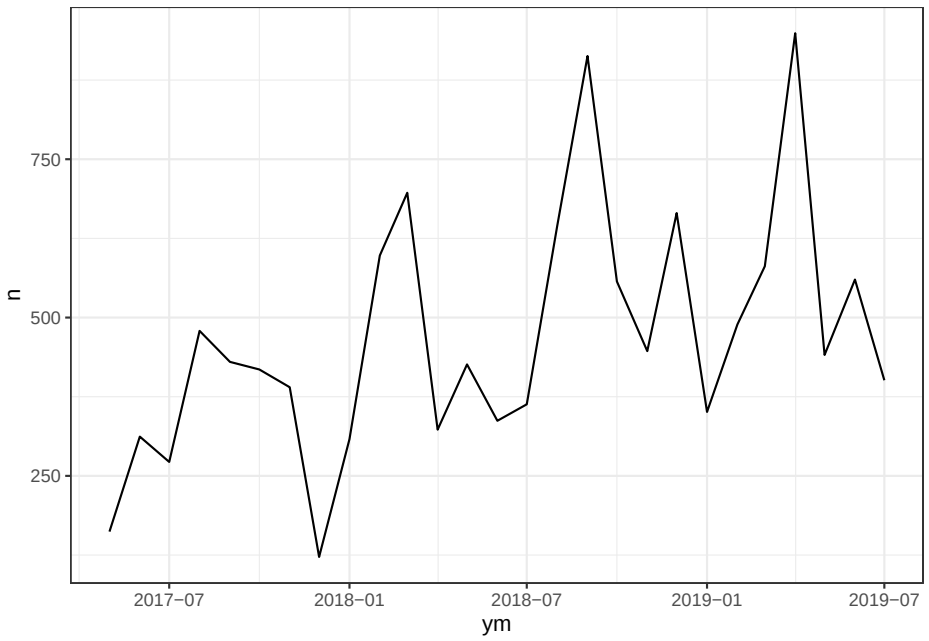
- Wir verwenden einen Ausschnitt der Daten aus der Beispielstudie. Konkret handelt es sich um Posts mit dem Suchwort *impf*, die zwischen dem 1. Mai 2016 und dem 8. Juli 2019 im Elternforum Urbia veröffentlicht wurden. Ausgeschlossen wurden unter anderem
  - sehr kurze Posts (weniger als 19 Wörter)
  - Posts mit dem Wort *schimpf*
  - Posts zur Impfung von Haustieren (nach einem kurzen Diktionär)
- Die Dokumentation zur Studie gibt weitere Informationen zur Erhebung und Bereinigung der Rohdaten.
- Diese Daten können aus Copyright- und Privacy-Gründen nicht auf GitHub veröffentlicht werden. Ich habe Sie daher im LMS hochgeladen. Bitte ladet die ZIP-Datei herunter.
  - Wenn ihr sie mit dem Code aus dem Repository integrieren wollt, müsst ihr sie in den Ordner "data" unter "R" entpacken.

```
# Laden der Daten
d = read_rds("R/data/exampe_data.rds")
d %>%
  print(n = 5)
```

```
## # A tibble: 12,635 x 5
##   post                          author   postdate      wc thread_title
##   <chr>                         <chr>    <date>     <int> <chr>
## 1 Wenn Impfungen zu Todesfäll~ zwerg-b~ 2018-04-06    26 HPV-Impfung
## 2 Hallo Moni Danke für deine ~ Inaktiv  2017-06-03    21 Warum so oft Scheidenp~
```

```
## 3 Hallo ja sind glaube ich dr~ danerl   2017-06-05    42 Warum so oft Scheidenp~
## 4 Guten Morgen, gibt es hier ~ butterf~ 2017-05-14   133 Impfung Deutschland/Ös~
## 5 In Österreich wird im 3., 5~ butterf~ 2017-05-15    68 Impfung Deutschland/Ös~
## # ... with 1.263e+04 more rows
```

```
d %>%
  mutate(ym = round_date(postdate, "month")) %>%
  count(ym) %>%
  ggplot(aes(ym, n)) + geom_line()
```




```
d %>%
  pull("wc") %>%
  summary()
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      20      37      60      87     102    2493
```

- Der Datensatz besteht aus 12,635 Posts.
  - Die Variable `post` enthält den vollen Text des Posts.
  - Die Variable `author` enthält den Accountnamen, von dem der Post abgegeben wurde.
  - Die Variable `date` enthält den Tag der Veröffentlichung.
  - Die Variable `wc` enthält die Zahl der Wörter des Posts.
  - Die Variable `thread_title` enthält den Titel des Diskussions-Threads.
- Pro Monat sind zwischen ca. 120 und 1.000 Posts in unserer Stichprobe.
- Typische Posts haben einen Umfang von zwischen 40 und 100 Wörtern

(Zur Erinnerung: Sehr kurze Post wurden bereits ausgeschlossen).

Bibliography

Maier, D., Waldherr, A., Miltner, P., Wiedemann, G., Niekler, A., Keinert, A., Pfetsch, B., Heyer, G., Reber, U., Häussler, T., Schmid-Petri, H., and Adam, S. (2018). Applying LDA topic modeling in communication research: Toward a valid and reliable methodology. *Communication Methods and Measures*, 12(2-3):93–118.

Roberts, M. E., Stewart, B. M., and Tingley, D. (2019). Stm: An R Package for Structural Topic Models. *Journal of Statistical Software*, 91(1):1–40.