Universität Münster

Text as Data

Session 8: Data Analysis – Unsupervised Topic Models

Mirko Wegemann

Universität Münster
Institut für Politikwissenschaft

24 June 2026

Logistics

- **Last week**: Preparation of a dataset in a numerical representation (document `feature` matrix)
- In addition: How can we compress our dataset?
- Missing: Pitfalls in compressing our data (e.g., `preText`-package)?
- **This week**: unsupervised topic models

Unsupervised methods

- Sometimes we want to classify text **without** knowing in advance what categories exist.
- In these cases, we can rely on **unsupervised**-approaches, which are described as "a class of methods that learn underlying features of text without explicitly imposing categories of interest" (Grimmer and Stewart 2013, p. 281).
- As Grimmer and Stewart (2013) show, these can be further divided into two approaches:
  1. **Clustering** (this week)
  2. Scaling (not covered in this seminar)

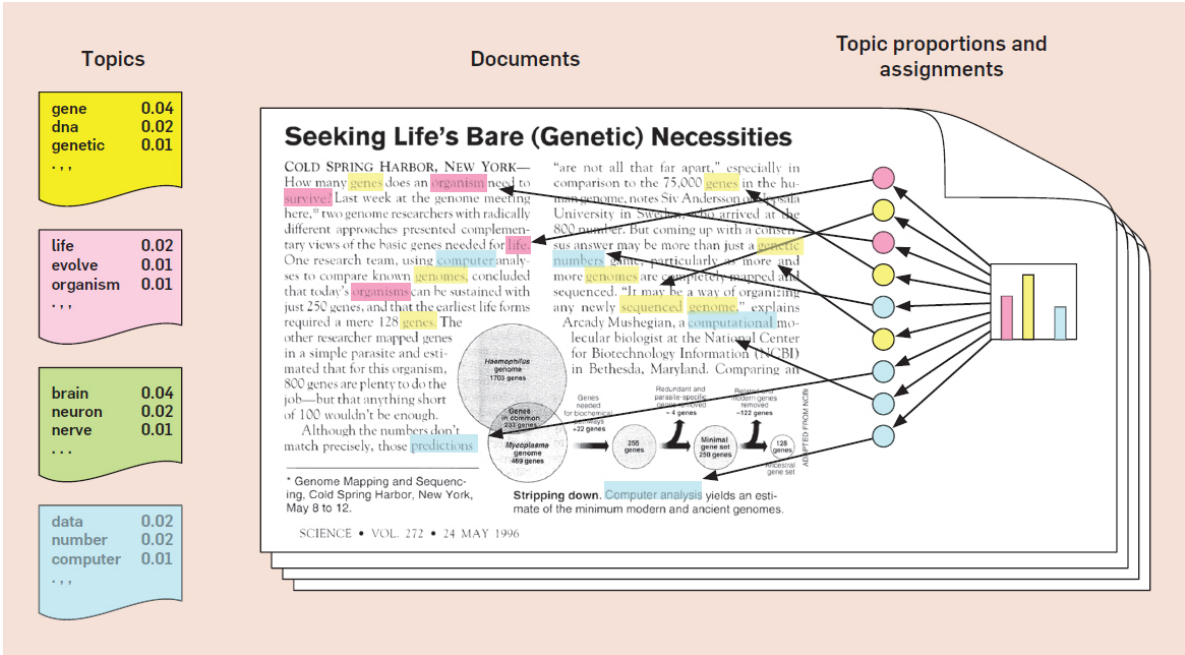
Topic models

- **Topic Models** are a form of **clustering**
- Compared to single-membership models (each text is assigned **one** category), Topic Models are **mixed-membership models**
- Assumption: Each text uses vocabulary from different topics
- Data requirement: large text corpora, u.U. pre-processed (s. last week)

Underlying idea of topic models I

- Documents consist of words, whose combination represents a **(latent) semantic meaning**
- Idea: Authors of a text write about a topic *k* and use terms associated with this
- Depending on the composition of the words, each document has a **probability p to belong to topic k**
- Central techniques are the **L**atent **S**emantic **A**nalysis (**LSA**) and the **L**atent **D**irichlet-**A**llocation (**LDA**)

Underlying idea of 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

Underlying idea of topic models III

- **LDA** is a probabilistic model, in which each term *t* has a certain probability of belonging to topic *k*, and each document *d* consists of several *k*s
  - Primarily used for Topic Modelling
  - **Generative Model**, where topics are initially randomly assigned to documents; each word has a probability of belonging to topic *k*
  - **Iterative** process, in which the log-likelihood is gradually reduced (the smaller, the better) [in the R-Output you can observe the iterative process in form of "expectation"-"maximization"]
  - The number of *k* must be defined in advance; there are no general guidelines here, but usually: the larger the corpus, the more topics it contains)

...more about LDA (Blei, Ng, and Jordan 2003)

Underlying idea of topic models IV

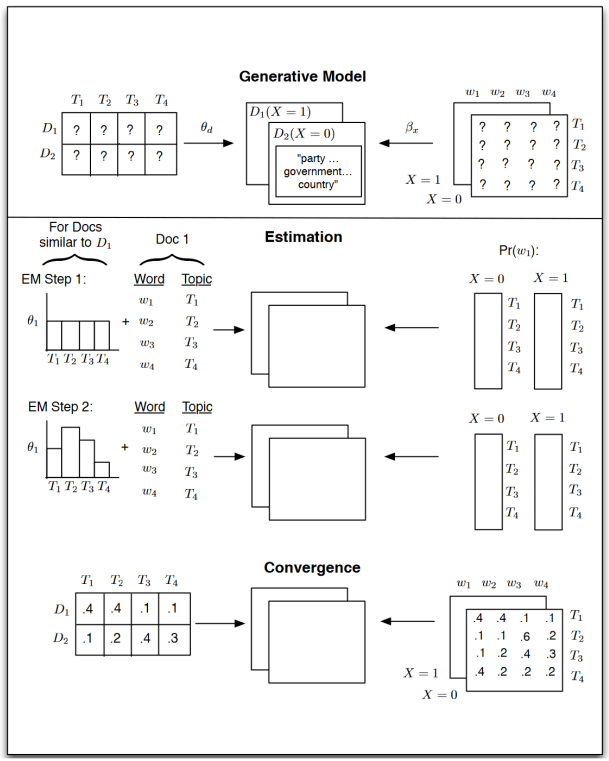


Figure: Process of Topic Models (Roberts et al. 2019, p. 4)

Topic Models in R I

There are different packages to use topic models in `R`. Two beginner-friendly packages are `lda` and `stm`.

Structural Topic Model (`stm`)

- Very similar to LDA
- Novelty: We can add **meta-information** to the topic model to obtain potentially more meaningful results
  1. **prevalence** of topics: Covariates that influence the frequency of a topic (e.g., in winter, the word "ice" may have a different usage than in summer)
  2. **content** of topics: Covariates that influence how a topic is discussed (parties have different understandings of economics and may use different vocabulary)

Topic Models in R II

- `stm` allows us technically to specify $k = 0$, where k is automatically determined; *Caution*: This does not mean that $\hat{k}$ is the true k.
- we can select further parameters, e.g.
  - `emtol = value`: specifies when to terminate the optimization
  - `seed = value`: for reproducible results
  - `init_type = value`: we can choose between different initializations; "spectral" is most reproducible
  - `LDAbeta = F`: SAGE-Updating of Topics

Topic Models in R III

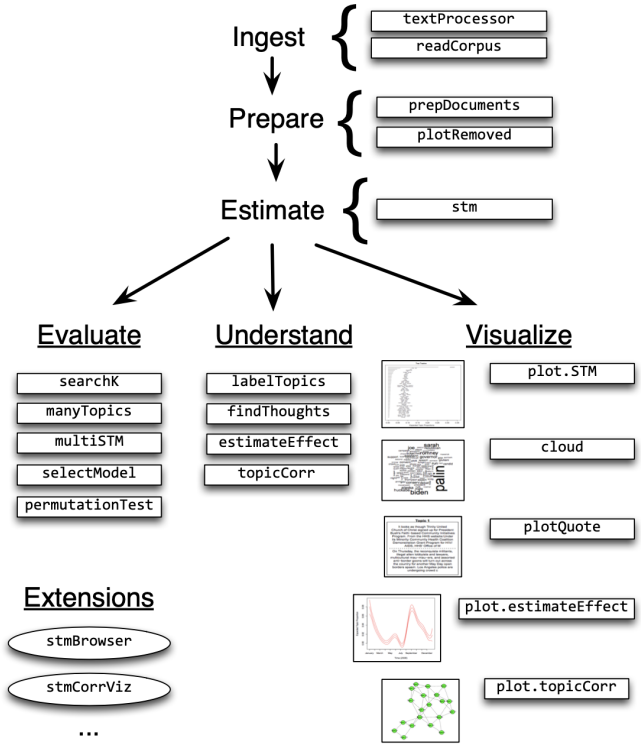


Figure: Functions in the `stm`-Package (Roberts et al. 2019, p. 5)

Topic Models in R IV

Steps in R

1. **Preparation** (as before: from data.frame to corpus, to tokens-object, to data frequency matrix)
2. **Estimation** of the topic model
3. **Interpretation**
4. **Validation**

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

- **Research question**:
- **Argument**:
- **Data and analysis**:
- **Results**:
- **Implications**:

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

- **Research question**: What meaning(s) does the left-right scale have for citizens?
- **Argument**: People associate different associations with abstract concepts (ideologies, values, etc.)
- **Data and analysis**: Allbus 08 (General Population Survey of Social Sciences); unsupervised topic model with open answers
- **Results**: Participants have diverse perceptions of what *right* and *left* means (depending on their own rating on the scale and socio-economic background)
- **Implications**: Response behavior depends systematically on how people interpret theoretical concepts

Theory

Model of the 'survey response process', according to which individuals follow four steps to give an answer

1. **Understanding**
2. Retrieving information
3. Assessment
4. Response

→ Focus on understanding a question

Hypotheses

- H_1 People have a different understanding of the left-right scale.
- H_2 The understanding of the left-right scale affects the left-right self-categorization.
- H_3 The understanding of the left-right scale is systematically dependent on characteristics of the respondents.

Data and Model I

Main variables:

1. **Left-Right Self-Categorization** ("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. **Understanding of Left-Right** (...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'?)

Data and Model II

Two analytical steps:

1. LDA; for evaluation: Lift-Scores; to remind: "Lift weights words by dividing by their frequency in other topics" (Roberts et al. 2019, p. 13)
2. Multivariate regression models

For replication:

- Number of topics $k = 4$
- Stopwords removed
- Punctuation removed
- Only words that occur in at least five answers
- SAGE specification of the algorithm

Results I

- Depending on political orientation, people associate different terms with “left” and “right”.
- “Policies” and “Values” are associated by people with left-leaning political attitudes with “left”; “Parties” are associated with “right” by right-leaning respondents.
- Socioeconomic background is decisive for associations; e.g., higher educated people and East Germans highlight ”values” at “left”.

Questions 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"

- What do you think?

Questions II

The Allbus study first asks about the left-right self-categorization of respondents before they are asked about their associations with "left" and "right".

- What is the advantage of this approach for the study? What is the disadvantage?

And now in R

Validation

Validation can take place to different degrees (overview of the consequences in Bernhard et al. 2023):

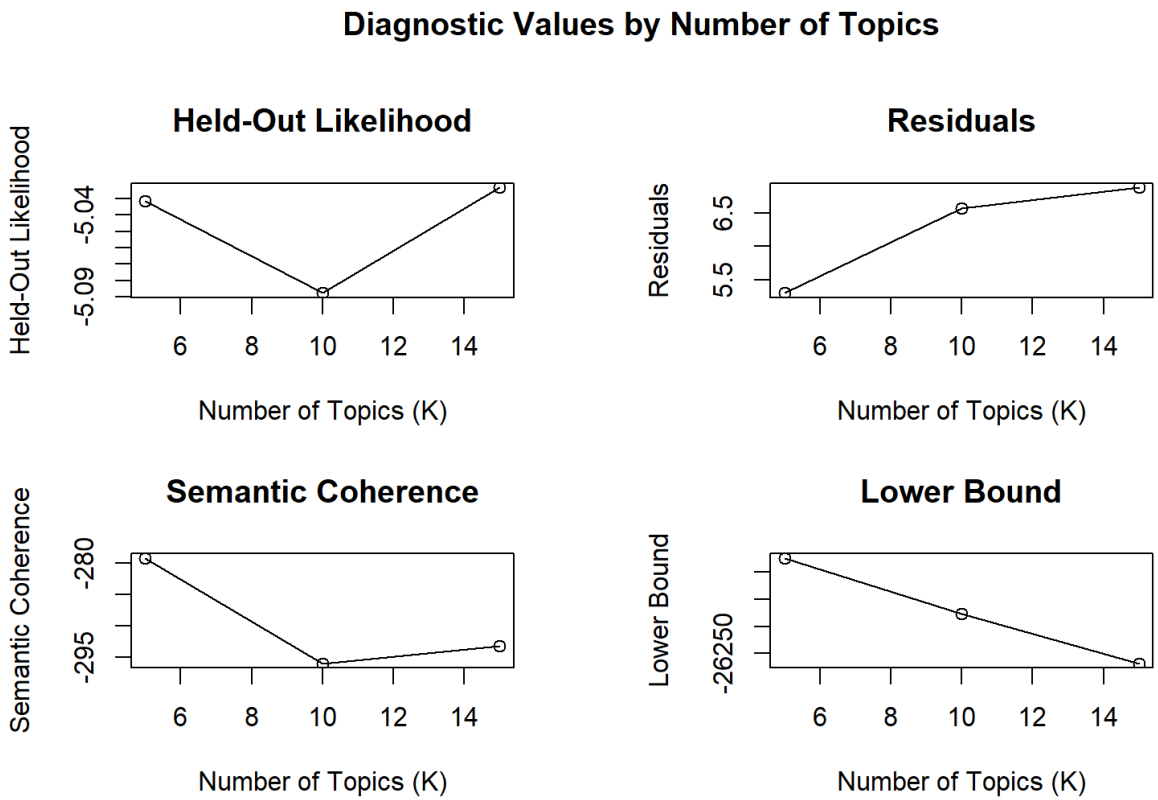
1. **semantic** validation (= interpretation of topics by frequency measures)
2. **statistical** validation (= smallest statistical error)
3. **external** validation according to Quinn et al. 2010 (= predictive power)
4. based on **manual coding**

Statistical Validation

We can use the `searchK()`-function from stm to determine the ideal number of topics k based on likelihood, lower bound, residuals and semantic coherence.

```
1
2       stm_left <- convert(dfm_left, "stm")
3
4       # Define number of 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)
```

Statistical Validation



Statistical Validation II

Alternatively, we can also, if the number of topics is clear, select the model that has the best parameters (`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)
```

Further Development of Topic Models

- Available in the R package `topicmodels`
- Correlated topic models: allow correlations between latent topics; executable in R with `CTM` (see for an application and Blei and Lafferty (2005) for methodological background)
- Topic models with user-input: semi-supervised approach; the topic model is given pre-specified keywords; executable in R with `textmodel_seededlda` (see for an application and Watanabe and Baturo (2024) for methodological background)
- `top2vec`: Topic models based on embeddings (more about this later in the course).
- `bertopic`: Topic models based on transformer models → more accurate representations, more computational intense; requires Python installation

Outlook

- Next week we will focus on supervised approaches
- Focus on creating predictors of text
- **Literature**:
  - James, G., Witten, D., Hastie, T., & Tibshirani. (2021). *An Introduction to Statistical Learning*. Retrieved September 19, 2024, from https://www.statlearning.com
  - Hvitfeldt, E., & Silge, J. (2022). *Supervised Machine Learning for Text Analysis in R*. CRC Press. Retrieved September 27, 2024, from https://smltar.com/
  - Stukal, D., Sanovich, S., Bonneau, R., & Tucker, J. A. (2017).Detecting Bots on Russian Political Twitter. *Big Data*, *5*(4), 310–324. https://doi.org/10.1089/big.2017.0038

Any questions?

Literature 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.

Literature 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

Hvitfeldt, E., & Silge, J. (2022). *Supervised Machine Learning for Text Analysis in R*. CRC Press. Retrieved September 27, 2024, from https://smltar.com/

James, G., Witten, D., Hastie, T., & Tibshirani. (2021). *An Introduction to Statistical Learning*. Retrieved September 19, 2024, from https://www.statlearning.com

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

Literature III

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

Stukal, D., Sanovich, S., Bonneau, R., & Tucker, J. A. (2017).Detecting Bots on Russian Political Twitter. *Big Data*, *5*(4), 310–324. https://doi.org/10.1089/big.2017.0038

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/10.1177/08944393231178605