

Topic Modeling

Aidar Zinnatullin

University of Kaiserslautern-Landau and University of Bologna

12. Juni 2024

Agenda

- What is behind topic modeling?

Agenda

- What is behind topic modeling?
- Showcase 1: Structural topic modeling (STM)

Agenda

- What is behind topic modeling?
- Showcase 1: Structural topic modeling (STM)
- Practicalities: selecting the optimal number of topics, interpreting the topics, and validating the output

Agenda

- What is behind topic modeling?
- Showcase 1: Structural topic modeling (STM)
- Practicalities: selecting the optimal number of topics, interpreting the topics, and validating the output
- Showcase 2: Keyword-assisted topic modeling

Agenda

- What is behind topic modeling?
- Showcase 1: Structural topic modeling (STM)
- Practicalities: selecting the optimal number of topics, interpreting the topics, and validating the output
- Showcase 2: Keyword-assisted topic modeling
- Other applications: Speaker Identity for Topic Segmentation (SITS), BERTopic

What is behind?

- Three basic concepts:
 - 1 Corpus (collection of text at our disposal)
 - 2 Document (item within the corpus)
 - 3 Terms or, simply, words
- Idea is to model a comprehensive **representation of the corpus** by inferring **latent content variables**, or *topics*
- Topics are between the corpus and documents (cluster of documents, in other words)
- Recurring patterns of word occurrence in documents
- Bag-of-words: words frequencies and distributions

Bag-of-Words and Document-Feature Matrix

Document	Make	America	Great	Again
Tweet 1	1	2	0	1
Tweet 2	0	1	3	1
Tweet 3	1	0	2	2

What is a topic?

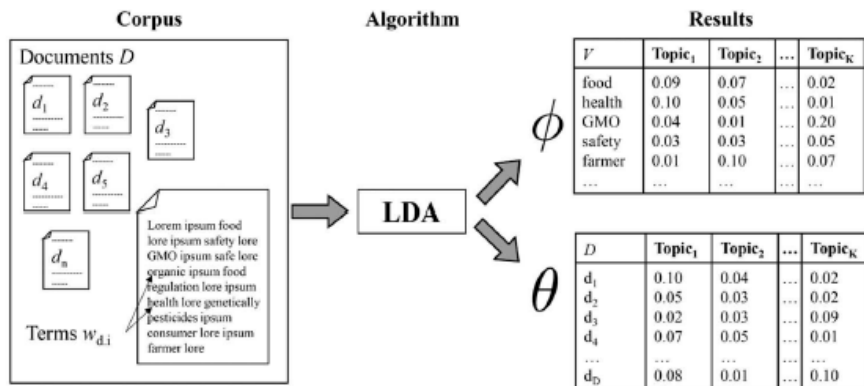
- **Substantively**: distinct subjects (war in Ukraine, energy crisis, Covid-19, etc.)
- **Statistically**: a topic is defined as a multinomial distribution over the words in the vocabulary of the corpus
- To discover topics, topic models use the patterns of words co-occurrence within and across documents

Data Generating Process

How to define the latent content structure of the corpus?

- We need to compute two matrices
 - 1 word-topic (ϕ)
 - 2 document-topic (θ)
- Bayesian approach to model the data generating process
- Dirichlet family of distributions for prior distributions of two matrices
- Prior distributions are governed by
 - 1 the number of topics, which is equal for both ϕ and θ
 - 2 abstract prior parameters α for ϕ , and β for θ
 - 3 α and β are chosen by a researcher
 - 4 But a topic modeling algorithm (for instance, LDA) is doing it for the researcher, i.e., it randomly assigns term probabilities to topics (ϕ) and topic probabilities to documents (θ)

Two Matrices for Latent Topical Structure



Source: [Maier et al., 2018](#)

Four Challenges

- Pre-processing

Four Challenges

- Pre-processing
- Selection of model parameters

Four Challenges

- Pre-processing
- Selection of model parameters
- Evaluation of model's reliability

Four Challenges

- Pre-processing
- Selection of model parameters
- Evaluation of model's reliability
- Interpreting the results and validation

Four Challenges

- Pre-processing
- Selection of model parameters
- Evaluation of model's reliability
- Interpreting the results and validation
- See more about it at [Maier et al. \(2018\)](#)

Showcase 1: Structural Topic Modeling (STM)

STM = LDA + Contextual Information

- More accurate estimation
- Better interpretability
- Topic prevalence and topic content across covariate levels

Preprocessing: important but...

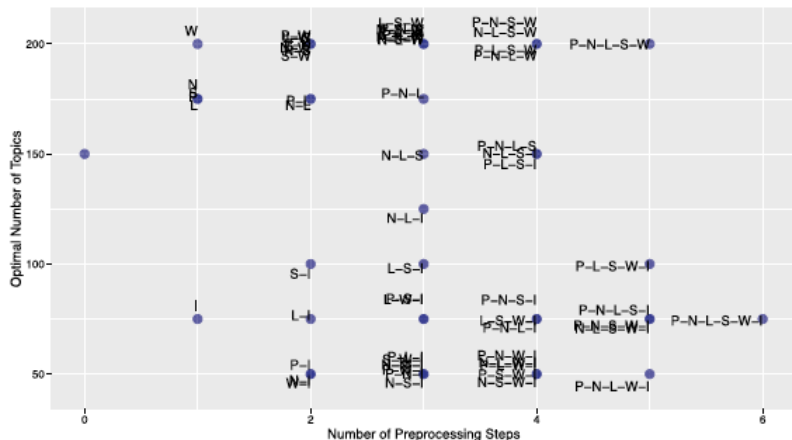
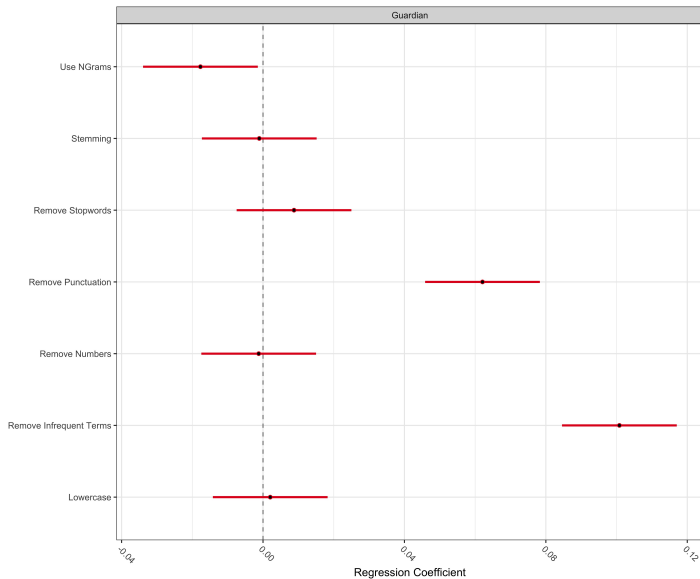


Figure 2. Plot depicting the optimal number of topics (as selected via perplexity) for each of 64 preprocessing specifications not including trigrams. On the x-axis is the number of preprocessing steps, and the y-axis is the number of topics. Each point is labeled according to its specification.

Source: [Denny & Spirling, 2017](#)

... do not follow blindly the quantitative metrics



Idea of preText

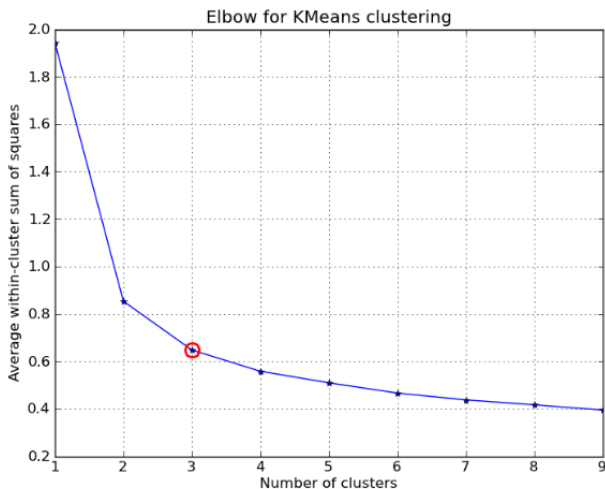
...if removing punctuation meant that the distance between the Labour 1983 manifesto and the Conservative 1983 manifesto was the largest pairwise distance observed in the DTM (something which makes substantive sense), but removing punctuation and numbers meant that this pairwise distance was the fifth largest observed (which makes little sense), red flags should be raised ([Denny & Spirling, 2017](#))

Number of Topics

- Trial and error
- Balance between semantic coherence and exclusivity
 - 1 Semantic coherence refers to the degree to which the top words in a topic make sense together and form a meaningful, interpretable concept
 - 2 Exclusivity denotes the uniqueness of words within a topic compared to other topics

Elbow Method

The elbow point on the plot is the point where the rate of decrease sharply slows down



Validation (1)

- reading words with the highest probability and FREX score

Validation (1)

- reading words with the highest probability and FREX score
- reading example documents

Validation (1)

- reading words with the highest probability and FREX score
- reading example documents
- but it resembles intuition and eyeballing

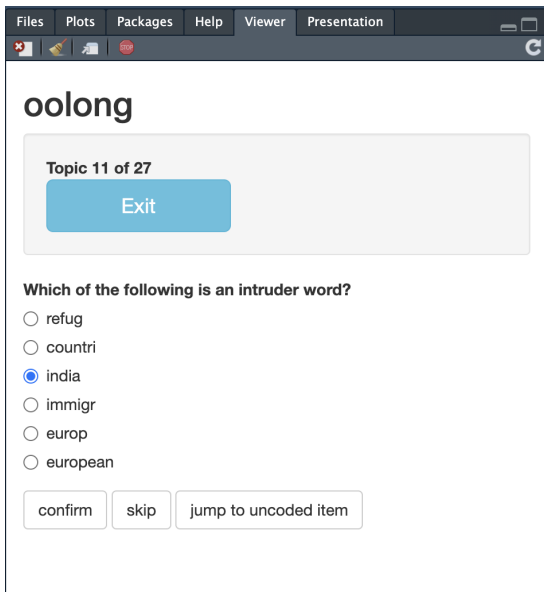
Validation (2)

- **Word intrusion:** This checks how well the words in a topic fit together. It tests if the topics make sense to humans
- **Topic intrusion:** This checks how well the topics in a document match what humans think the document is about
- [Chang et al. 2009](#)

Word and Topic Intrusion Test Implementation in R (1)

- **Word intrusion:** This checks how well the words in a topic fit together. It tests if the topics make sense to humans
- **Topic intrusion:** This checks how well the topics in a document match what humans think the document is about
- [Chang et al. 2009](#)
- Implementation in R: [Chan & Sältzer, 2020](#)

Word and Topic Intrusion Test Implementation in R (2)



The screenshot shows an R application window with a dark-themed menu bar containing 'Files', 'Plots', 'Packages', 'Help', 'Viewer', and 'Presentation'. Below the menu bar is a toolbar with icons for file operations and a 'STOP' button. The main content area is titled 'oolong' and displays 'Topic 11 of 27'. A blue 'Exit' button is visible. Below this, a question is posed: 'Which of the following is an intruder word?'. Five radio button options are listed: 'refug', 'countri', 'india' (which is selected), 'immigr', and 'europ'. At the bottom, there are three buttons: 'confirm', 'skip', and 'jump to uncoded item'.

oolong

Topic 11 of 27

Exit

Which of the following is an intruder word?

- ☐ refug
- ☐ countri
- ☒ india
- ☐ immigr
- ☐ europ
- ☐ european

confirm skip jump to uncoded item

Showcase 2: Keyword-assisted Topic Modeling

- specifies keywords for each topic, guiding the model to generate more relevant and interpretable topics

Showcase 2: Keyword-assisted Topic Modeling

- specifies keywords for each topic, guiding the model to generate more relevant and interpretable topics
- offers different variants (e.g., keyATM-basic, keyATM-covariates, keyATM-dynamic) to handle various types of research needs

Showcase 2: Keyword-assisted Topic Modeling

- specifies keywords for each topic, guiding the model to generate more relevant and interpretable topics
- offers different variants (e.g., keyATM-basic, keyATM-covariates, keyATM-dynamic) to handle various types of research needs
- familiar setting for the users of `quanteda`

Showcase 2: Keyword-assisted Topic Modeling

- specifies keywords for each topic, guiding the model to generate more relevant and interpretable topics
- offers different variants (e.g., keyATM-basic, keyATM-covariates, keyATM-dynamic) to handle various types of research needs
- familiar setting for the users of `quanteda`
- <https://keyatm.github.io/keyATM/index.html>

Speaker Identity for Topic Segmentation (SITS)

- Rossiter, E.L. (2022), Measuring Agenda Setting in Interactive Political Communication. *American Journal of Political Science*, 66: 337-351, <https://onlinelibrary.wiley.com/doi/abs/10.1111/ajps.12653>

Speaker Identity for Topic Segmentation (SITS)

- Rossiter, E.L. (2022), Measuring Agenda Setting in Interactive Political Communication. American Journal of Political Science, 66: 337-351, <https://onlinelibrary.wiley.com/doi/abs/10.1111/ajps.12653>
- three sets of latent quantities of interest:
 - ① what topics are on the agenda
 - ② where shifts in the agenda occur
 - ③ each actor's agenda-setting power

BERTopic

- combines the power of BERT's contextual embeddings with traditional topic modeling techniques, allowing for the extraction of coherent topics from large text corpora
- covariates can be included
- BERTopic utilizes hierarchical clustering to group similar documents together based on their BERT embeddings
- Grootendorst M., BERTopic: Neural topic modeling with a class-based TF-IDF procedure

Word Embeddings

Distributional Representation: Illustration

If we label the dimensions in a hypothetical word vector (there are no such pre-assigned labels in the algorithm of course), it might look a bit like this:



Such a vector comes to represent in some abstract way the 'meaning' of a word

Source: [StackOverflow](#), Understanding embedding vectors dimension