The World of Topic Modeling in R

Gregor Wiedemann*

This software review provides a systematic overview of R packages for topic modeling. These packages facilitate the application of computational text analysis to conduct research compatible with a wide variety of methodological frameworks employed in the social and communication sciences. For this overview, the analysis process is divided into four steps: preprocessing, inference, evaluation, and interpretation of topic models. The summary shows that the wide selection of R packages available today provides easy-to-use access to this increasingly popular computational social science method and contributes to its methodical standardization.

Keywords: R, topic model, visualization, text processing, text mining, automatic content analysis

Introduction

In recent years, topic modeling has been added to the analytical toolbox of an increasing number of communication scholars (Maier, Waldherr, Miltner, Wiedemann, Niekler, Keinert, Pfetsch, Heyer, Reber, Häussler, Schmid-Petri & Adam 2018). Topic modelling is a cost-efficient biclustering approach to the analysis of text collections capable of processing from a few hundred to millions of documents. As input, a topic model receives a topic number and a document-term matrix, that is, a count matrix of all different word types present in each document. An inference process then clusters the documents into thematically coherent groups and provides interpretable groups of semantically related words. These two latent output variables, the topic-document distribution, and the topic-word distribution are the basis for complex quantitative studies such as cross-sectional or longitudinal content analysis.

First developed in the discipline of computer science with an application for information retrieval in mind, topic modelling does not directly operationalise any theoretical or methodological concept from the social and communication sciences. Scholars from these disciplines who started to adopt topic modelling for their research, however, pointed out that this technique is compatible with a variety of methodological frameworks. Nelson (2020) reflects on computational grounded theory that involves topic modelling as playing a major role in extracting exploratory insights from large corpora. Jacobs & Tschötschel (2019) argue that by allowing the discovery of thematic patterns in very large text volumes, topic modelling fits the theoretical framework underlying (post-)structuralist discourse analysis. An increasing number of studies utilize topics as a proxy to perform frame analysis (e.g. Ylä-Anttila, Eranti & Kukkonen 2022). This methodological fit of topic models is, on the one hand, a consequence of their ability to discover thematic structures in a purely data-driven way, and on the other, of their innovative modelling of complex meaning structures from a linguistic point of view. Wiedemann & Fedtke (2021) point out that, for the first time, the latent semantic variables of topic models are based on a separation of the meaning

* Dr. Gregor Wiedemann, Leibniz-Institut für Medienforschung | Hans-Bredow-Institut (HBI), Rothenbaumchaussee 36, 20148 Hamburg, Deutschland, g.wiedemann@leibniz-hbi.de, ORCID: https://orcid.org/0000-0002-4239-295X.

 DOI: 10.5771/1615-634X-2022-3-286

representation of words from their linguistic surface form. This separation establishes a 'Saussurean turn' in automatic content analysis.

Today, researchers planning to perform topic modelling can choose from a wide range of software tools. This review provides a brief overview of the most important and widely used packages for this task for the scientific programming language R (R Core Team 2022). The process of topic modelling research can be broadly divided into four consecutive steps: *preprocessing*, *model inference*, *model evaluation*, and *(visual) interpretation*. The R community has developed several packages that support all four steps with a large variety of functions. Some packages provide efficient implementations for model inference algorithms such as variational inference for Latent Dirichlet Allocation (LDA; Blei, Ng & Jordan 2003). Other packages wrap around these basic functions and provide convenient solutions to common pre-processing methods, the evaluation and tuning of hyperparameters, or the visualization of results. Due to established standard data formats, for example, text corpora and sparse matrices, many of these packages are interoperable and can be combined to support individual and complex research designs. Maier et al. (2018) contain a summary of best practice suggestions for topic modelling. Table 1 summarises the most important R packages to conduct one or more steps of the research process along with these best practice suggestions.

Preprocessing

For a couple of years now, the package *quanteda* is the undisputed champion among all text processing packages in R. It provides very fast and broad basic functionality for text pre-processing, the creation of corpus objects, and conversion between different data formats. It allows for maximum freedom in the creation of a pre-processing pipeline at the cost of a user need to find out on their own what to do for an optimal result. In a more convenient approach, the packages *textmineR* and *tosca* provide functions that implement the most common techniques used in LDA text pre-processing (e. g. stopword removal, lower casing, and vocabulary pruning). To also include linguistic knowledge, *udpipe* provides functionality for the lemmatization of words in several languages while also making it possible to filter words in documents by their grammatical part of speech. This allows topic modelling based only on nouns and adjectives, for example. For Tweets, *Twitmo* provides functions for data collection and the cleaning and pooling of tweets, i. e. the concatenation of tweets by accounts to obtain longer documents that improve the topic inference process. Denny & Spirling (2018) highlight that pre-processing has significant effects on the outcome of topic models. Researchers are subsequently advised to carefully select and evaluate their pre-processing pipelines.

Inference

The core of topic modelling is the statistical inference of the two latent variables, the topic-document distribution, and the topic-term distribution, based on a document collection and some given hyperparameters such as the number of topics. R users today can choose from more than 10 variants of topic models available through external packages. So far, only the seminal LDA model and the structural topic model (STM) has gained wider attention in communication research. LDA is provided with the seminal packages *lda*, and *topicmodels*. Two more recent implementations are contained in *textmineR* and *text2vec*. They all contain high-speed native C++

Table 1: R packages for different steps of the topic modelling workflow

Package	Citation	Pre-processing	Model inference	Model evaluation	(Visual) interpretation
textmineR	Jones 2021	x	LDA (LSA, CTM)	x	
quanteda	Benoit et al. 2018	x			
Twitmo	Buchmüller et al. 2021	x	(LDA, STM)		x
tosca	Koppers et al. 2021	x	(LDA)	x	x
lda	Chang 2015		LDA, RTM, sLDA, NUBBI		
topicmodels	Grün & Hornik 2011		LDA, CTM		
stm	Roberts et al. 2019		STM	x	x
mallet	Mimno 2013		LDA		
text2vec	Selivanov et al. 2020		LDA, LSA	x	
topicmodels.etm	Wijffels 2021		ETM		
BTM	Wijffels 2020		BTM		
ldaPrototype	Rieger 2020		(LDA)	x	
ldatuning	Nikita 2020		(LDA)	x	
seededlda	Watanabe & Xuan-Hieu 2020		seededLDA		
keyATM	Eshima et al. 2020		keyATM [base, covariates, dynamic, label]		x
LDAvis	Sievert & Shirley 2014				x
rollinglda	Rieger 2021		(LDA)		
oolong	Chan & Sältzer 2020			x	x
udpipe	Wijffels 2022	x			

Model names in brackets in the model inference column indicate that this inference algorithm is imported by another package.

implementations of LDA inference algorithms. A wrapper around a fast, parallelized Java implementation is provided by the package *mallet*. Compared to the previously mentioned packages, *mallet* and *textmineR* have the advantage of providing automatic and asynchronous optimization of the alpha hyperparameter of the LDA model that steers the topic-document distribution. Asynchronous alpha parameters for each topic can lead to more stable and better interpretable topics. *stm* extends the basic model by incorporating covariates such as document authors or time into the modelling process. Through the default initialization of topics with spectral clustering and the variational

inference algorithm to determine the latent variables, STM also leads to deterministic modelling results, which is a major advantage regarding the reliability issues of LDA. An interesting option to capture the evolution of topics over time is provided by the *rollingLDA* package that maps the change of topic-word distributions across subsequent time slices. To employ topic modelling beyond explorative, inductive research, the approach of seeded topic modelling is provided by the packages *seededLDA*, and *keyATM*. These model variants allow for the inclusion of term lists as initial topic seeds to guide the inference process as it fits to pre-defined categories. *keyATM* in particular provides a very mature and fast implementation of different seeded model variants that are easy to use and evaluate. Models with so far underexplored potential in communication science are provided by the packages *topicmodels.etm*, which employs the modelling of topic and word semantics in embedding spaces, and *BTM* that implements the biterm topic model, which is based on co-occurrence windows of words to better model short text such as tweets. Unfortunately, both packages lack the maturity and convenience functions for quick results. As interesting modelling alternatives to LDA, however, they may be worth the effort of tweaking the R code to get them running. Models such as latent semantic analysis (LSA), supervised LDA (sLDA), relational topic model (RTM), and networks uncovered by Bayesian inference (NUBBI) as provided by the pioneering package *lda*, and the correlated topic model (CTM) implemented in *topicmodels*, have not yet gained significant attention in empirical research. I presume this is mainly due to the lowered semantic interpretability of their outcomes compared to standard LDA.

Evaluation

Since topic modelling is influenced by pre-processing steps, the choice of the model itself, the number of topics, and the other hyperparameters on the one hand, and variation due to random processes during inference, on the other hand, the modelling results tend to vary between different parameter settings and repeated runs. For this reason, model results always need to be evaluated carefully, and modelling parameters need to be optimised accordingly. The packages *ldatuning*, *textmineR, text2vec,* and *stm* support researchers by making suggestions for an optimal number of topics based on numeric evaluation criteria such as topic coherence. With little effort, the ldatuning package can be easily adapted to models other than the built-in topicmodels LDA inference. *ldaPrototype*, finally, deals with the indeterministic nature of the topic model inference. To avoid invalid interpretations from random results, it finds the model with the least distance to all other models from a series of repeated inferences. This model can then be interpreted as a representative model for the collection. After determining optimal parameters by some of the previous approaches, it is recommended to identify such a representative model as the final model in the last step of the model selection process.

(Visual) Interpretation

The clusters of words and documents emerging from a topic model need systematic interpretation and careful validation to generate a useful research result. This task is supported by a range of packages that provide standardised validation procedures and visual access to the results. *LDAvis* is a widely adopted package that visualises the similarity of topics through 2D projection together with the shares and exclusivity of top terms of each topic via an interactive browser interface. The packages *tosca* and

oolong provide a number of human-in-the-loop tests for the validity of topics such as topic intrusion and word (set) intrusion, which can be conducted to obtain qualitative evaluation metrics. *KeyATM* and *STM* provide a couple of plot functions to create visual representations of their specific models that incorporate covariates or temporal evolution. For LDA models that are applied to document collections with timestamps, *tosca* provides a variety of readily available plots that visualise the distribution of topics across time. For geotagged Twitter data, *Twitmo* provides functionality to visualize tweets of certain topics on a map.

In summary, R users are provided with a broad range of mature packages to perform each step of topic modelling research. These packages provide easy-to-use access to the inference of sophisticated topic models published in the scientific literature over the past few years with only a few lines of code. This facilitated access to the method carries the risk of trading systematic and profound research for the display of ad hoc interpretations based on topic model results that appear plausible at first glance. Instead, freed up resources through the convenient access via R packages should be invested in the standardisation and advancements of the methodology, i. e. to carefully carry out the process of text pre-processing, model selection, and evaluation followed by extended validity and reliability checks. Due to the highly dynamic R community, it is well worth searching out new packages from time to time.

References

Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Müller, S., & Matsuo, A. (2018). quanteda: An R package for the quantitative analysis of textual data. *Journal of Open Source Software*, *3*(30), 774. https://doi.org/10.21105/joss.00774.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022. https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf [13.07.2022].

Buchmueller, A., Kant, G., Weisser, C., & Saefken, B. (2021). *Twitmo: Twitter topic modeling and visualization for R* (Version 0.1.2) [Computer software]. https://CRAN.R-project.org/package=Twitmo [13.07.2022].

Chan, C., & Sältzer, M. (2020). oolong: An R package for validating automated content analysis tools. *Journal of Open Source Software*, *5*(55), 2461. https://doi.org/10.21105/joss.02461.

Chang, J. (2015). *lda: Collapsed Gibbs sampling methods for topic models* (Version 1.4.2) [computer software]. https://CRAN.R-project.org/package=lda [13.07.2022].

Denny, M. J., & Spirling, A. (2018). Text preprocessing for unsupervised learning: why it matters, when it misleads, and what to do about it. *Political Analysis*, *26*(2), 168–189. https://doi.org/10.1017/pan.2017.44.

Eshima, S., Imai, K., & Sasaki, T. (2020). *Keyword assisted topic models*. http://arxiv.org/pdf/2004.05964v2.

Grün, B., & Hornik, K. (2011). Topicmodels: an R package for fitting topic models. *Journal of Statistical Software*, *40*(13), 1–30. http://www.jstatsoft.org/v40/i13/ [13.07.2022].

Jacobs, T., & Tschötschel, R. (2019). Topic models meet discourse analysis: a quantitative tool for a qualitative approach. *International Journal of Social Research Methodology*, *22*(5), 469–485. https://doi.org/10.1080/13645579.2019.1576317.

Jones, T. (2021). *textmineR: functions for text mining and topic modeling* (Version 3.0.5) [computer software]. https://CRAN.R-project.org/package=textmineR [13.07.2022].

Koppers, L., Rieger, J., Boczek, K., & Nordheim, G. von. (2021). *tosca: tools for statistical content analysis* (Version 0.3-2) [Computer software]. https://github.com/Docma-TU/tosca [13.07.2022].

Maier, D., Waldherr, A., Miltner, P., Wiedemann, G., Niekler, A., Keinert, A., Pfetsch, B., Heyer, G., Reber, U., Häussler, T., Schmid-Petri, H., & Adam, S. (2018). Applying LDA topic modeling in communication research: toward a valid and reliable methodology. *Com-*

munication Methods and Measures, *12*(2-3), 93–118. https://doi.org/10.1080/19312458.2018.1430754.
Mimno, D. (2013). *mallet: A wrapper around the Java machine learning tool MALLET* (Version 1.0) [computer software]. https://CRAN.R-project.org/package=mallet [13.07.2022].
Nelson, L. K. (2020). Computational grounded theory: a methodological framework. *Sociological Methods & Research*, *49*(1), 3–42. https://doi.org/10.1177/0049124117729703.
Nikita, M. (2020). *ldatuning: tuning of the latent Dirichlet allocation models parameters* (Version 1.0.2) [computer software]. https://github.com/nikita-moor/ldatuning [13.07.2022].
R Core Team. (2022). *R: a language and environment for statistical computing* [computer software]. Vienna, Austria. https://www.R-project.org/ [13.07.2022].
Rieger, J. (2020). ldaPrototype: A method in R to get a Prototype of multiple Latent Dirichlet Allocations. *Journal of Open Source Software*, *5*(51), 2181. https://doi.org/10.21105/joss.02181.
Rieger, J. (2021). *rollinglda: construct consistent time series from textual data* (Version 0.1.1) [computer software]. https://github.com/JonasRieger/rollinglda [13.07.2022].
Roberts, M. E., Stewart, B. M., & Tingley, D. (2019). stm: an R package for structural topic models. *Journal of Statistical Software*, *91*(2). https://doi.org/10.18637/jss.v091.i02.
Selivanov, D., Bickel, M., & Wang, Q. (2020). *text2vec: modern text mining framework for R* (Version 0.6) [computer software]. https://CRAN.R-project.org/package=text2vec [13.07.2022].
Sievert, C., & Shirley, K. E. (2014). LDAvis: A method for visualizing and interpreting topics. In *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces* (pp. 63–70). ACL.
Watanabe, K., & Xuan-Hieu, P. (2020). *seededlda: Seeded-LDA for topic modeling* (Version 0.5) [computer software]. https://github.com/koheiw/seededlda [13.07.2022].
Wiedemann, G., & Fedtke, C. (2021). From frequency counts to contextualized word embeddings: The Saussurean turn in automatic content analysis. In U. Engel, A. Quan-Haase, S. X. Liu, & L. Lyberg (Eds.), *Handbook of Computational Social Science, Volume 2* (pp. 366–385). Routledge. https://doi.org/10.4324/9781003025245-25.
Wijffels, J. (2020). *BTM: Biterm topic models for short text* (Version 0.3.4) [computer software]. https://CRAN.R-project.org/package=BTM [13.07.2022].
Wijffels, J. (2021). *topicmodels.etm: topic modelling in embedding spaces* (Version 0.1.0) [computer software]. https://CRAN.R-project.org/package=topicmodels.etm [13.07.2022].
Wijffels, J. (2022). *udpipe: tokenization, parts of speech tagging, lemmatization and dependency parsing with the 'UDPipe' 'NLP' Toolkit* (Version 0.8.9) [computer software]. https://CRAN.R-project.org/package=udpipe [13.07.2022].
Ylä-Anttila, T., Eranti, V., & Kukkonen, A. (2022). Topic modeling for frame analysis: a study of media debates on climate change in India and USA. *Global Media and Communication*, *18*(1), 91–112. https://doi.org/10.1177/17427665211023984.