The Journal of Open Source Software

ldaPrototype: A method in R to get a Prototype of multiple Latent Dirichlet Allocations

Jonas Rieger[1]

1 TU Dortmund University

DOI: 10.21105/joss.02181

Software

- Review
- Repository
- Archive

Editor: Karthik Ram

Reviewers:

- @tommyjones
- @bstewart

Submitted: 10 March 2020
Published: 16 July 2020

Summary

Topic Modeling (Blei, 2012) is one of the biggest subjects in the field of text data analysis. Here, the Latent Dirichlet Allocation (Blei, Ng, & Jordan, 2003) takes a special position. A large part of scientific text data analyses are based on this model (LDA). The LDA method has a far-reaching disadvantage. Random initialization and conditional reassignments within the iterative process of the Gibbs sampler (Griffiths & Steyvers, 2004) can result in fundamentally different models when executed several times on the same data and with identical parameter sets. This fact greatly limits the scientific reproducibility.

Up to now, the so-called eye-balling method has been used in practice to select suitable results. From a set of models, subjective decisions are made to select the model that seems to fit the data best or, in the worst case, the result that best supports one's hypothesis is chosen. The latter contradicts basically good scientific practice. A different method of objective and automated selection has also become established. A model from a set of LDAs can be determined optimizing the log-likelihood using the perplexity on held-out data. The `R` (R Core Team, 2020) package `topicmodels` (Grün & Hornik, 2011) provides a workflow for this procedure. As an extension, Nguyen, Boyd-Graber, & Resnik (2014) proposed to average different iterations of the Gibbs sampling procedure to achieve an increase of perplexity. The averaging technique has the weakness, that the user does not get token specific assignments to topics, but only averaged topic counts or proportions per text. In addition, Chang, Boyd-Graber, Gerrish, Wang, & Blei (2009) were able to show that selection mechanisms aiming for optimizing likelihood-based measures do not correspond to the human perception of a well-adapted model of text data. Instead, the authors propose a so-called intruder procedure based on human codings. The corresponding methodology is implemented in the package `tosca` (Koppers, Rieger, Boczek, & von Nordheim, 2019).

The `R` package `ldaPrototype` on the other hand determines a prototypical LDA by automated selection from a set of LDAs. The method improves reliability of findings drawn from LDA results (Rieger, Koppers, Jentsch, & Rahnenführer, 2020), which is achieved following a typical statistical approach. For a given combination of parameters, a number of models is calculated (usually about 100), from which that LDA is determined that is most similar to all other LDAs from a set of models. For this purpose pairwise model similarities are calculated using the S-CLOP measure (**S**imilarity of Multiple Sets by **C**lustering with **Lo**cal **P**runing), which can be determined by a clustering procedure of the individual topic units based on topic similarities of the two LDA results considered. The package offers visualization possibilities for comparisons of LDA models based on the clustering of the associated topics. Furthermore, the package supports the repetition of the modeling procedure of the LDA by a simple calculation of the repeated LDA runs.

In addition to the possibility of local parallel computation by connecting to the package `parallelMap` (Bischl & Lang, 2019), there is the possibility to calculate using batch systems on high performance computing (HPC) clusters by integrating helpful functions from the

package batchtools (Lang, Bischl, & Surmann, 2017). This is especially helpful if the text corpora contains several hundred of thousands articles and the sequential calculation of 100 or more LDA runs would extend over several days. The modeling of single LDA runs is done with the help of the computation time optimized R package lda (Chang, 2015), which implements the calculation in C++ code. In general, the package ldaPrototype is based on S3 objects and thus extends the packages lda and tosca by user-friendly display and processing options. Other R packages for estimating LDA are topicmodels and mallet (Mimno, 2013), whereas stm (Roberts, Stewart, & Tingley, 2019) offers a powerful framework for Structural Topic Models and quanteda (Benoit et al., 2018) is a popular framework for preprocessing and quantitative analysis of text data.

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. doi:10.21105/joss.00774

Bischl, B., & Lang, M. (2019). *parallelMap: Unified Interface to Parallelization Back-Ends*. Retrieved from https://CRAN.R-project.org/package=parallelMap

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

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*, 993–1022. doi:10.1162/jmlr.2003.3.4-5.993

Chang, J. (2015). *lda: Collapsed Gibbs Sampling Methods for Topic Models*. Retrieved from https://CRAN.R-project.org/package=lda

Chang, J., Boyd-Graber, J., Gerrish, S., Wang, C., & Blei, D. M. (2009). Reading Tea Leaves: How Humans Interpret Topic Models. In *Proceedings of the 22nd international conference on neural information processing systems*, NIPS (pp. 288–296). Red Hook, NY, USA: Curran Associates Inc. ISBN: 9781615679119

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences*, *101*(suppl 1), 5228–5235. doi:10.1073/pnas.0307752101

Grün, B., & Hornik, K. (2011). topicmodels: An R Package for Fitting Topic Models. *Journal of Statistical Software*, *40*(13), 1–30. doi:10.18637/jss.v040.i13

Koppers, L., Rieger, J., Boczek, K., & von Nordheim, G. (2019). *tosca: Tools for Statistical Content Analysis*. doi:10.5281/zenodo.3591068

Lang, M., Bischl, B., & Surmann, D. (2017). batchtools: Tools for R to work on batch systems. *The Journal of Open Source Software*, (10). doi:10.21105/joss.00135

Mimno, D. (2013). *mallet: A wrapper around the Java machine learning tool MALLET*. Retrieved from https://CRAN.R-project.org/package=mallet

Nguyen, V.-A., Boyd-Graber, J., & Resnik, P. (2014). Sometimes Average is Best: The Importance of Averaging for Prediction using MCMC Inference in Topic Modeling. In *Proceedings of the 2014 EMNLP-Conference* (pp. 1752–1757). ACL. doi:10.3115/v1/D14-1182

R Core Team. (2020). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Rieger, J., Koppers, L., Jentsch, C., & Rahnenführer, J. (2020). Improving Reliability of Latent Dirichlet Allocation by Assessing Its Stability Using Clustering Techniques on Replicated Runs. Retrieved from https://arxiv.org/abs/2003.04980

Roberts, M. E., Stewart, B. M., & Tingley, D. (2019). stm: An R Package for Structural Topic Models. *Journal of Statistical Software*, *91*(2), 1–40. doi:10.18637/jss.v091.i02