Package 'lda.svi'

July 12, 2019

Title Fit Latent Dirichlet Allocation Models using Stochastic Variational Inference

Version 0.1.0

Description Fits Latent Dirichlet Allocation topic models to text data using the stochastic variational inference algorithm described in Hoffman et. al. (2013) <arXiv:1206.7051v3>. This method is more efficient than the original batch variational inference algorithm for LDA, and allows users to fit LDA models with more topics and to larger text corpora than would be feasible using that older method.

Depends R (>= 3.5.0)

License MIT + file LICENSE

BugReports https://github.com/nerskin/lda.svi/issues

Encoding UTF-8

RoxygenNote 6.1.1

LinkingTo Rcpp, RcppArmadillo, BH

Imports Rcpp, reshape2, tm (>= 0.6), methods, Rdpack

Suggests topicmodels

SystemRequirements C++11

NeedsCompilation yes

Author Nicholas Erskine [aut, cre]

Maintainer Nicholas Erskine <nicholas.erskine95@gmail.com>

Repository CRAN

Date/Publication 2019-07-12 16:10:02 UTC

R topics documented:

lda_svi *Fit a Latent Dirichlet Allocation model to a text corpus*

Description

Fit a Latent Dirichlet Allocation model to a text corpus

Usage

```
lda_svi(dtm, passes = 10, batchsize = 256, maxiter = 100, K,
  eta = 1/K, alpha = 1/K, kappa = 0.7, tau_0 = 1024,
  tidy_output = TRUE)
```

Arguments

dtm	This must be a DocumentTermMatrix (with term frequency weighting) from the tm package.
passes	The number of passes over the whole corpus - how many times we update the local variables for each document.
batchsize	The size of the minibatches.
maxiter	The maximum iterations for the "E step" for each document (the updating of the per-document parameters within each minibatch). The default of 100 follows the reference implementation in python by the authors.
K	The number of topics
eta	Dirichlet prior hyperparameter for the document-specific topic proportions.
alpha	Dirichlet prior hyperparameter for the topic-specific term proportions.
kappa	learning rate parameter. Lower values give greater weight to later iterations. For guaranteed convergence to a local optimum, kappa must lie in the interval (0.5,1].
tau_0	learning rate parameter. Higher values reduce the influence of early iterations.
tidy_output	if true, the parameter estimates are returned as 'long' data frames; otherwise they are returned as matrices.

Details

The implementation here is based on the python implementation by Matthew D. Hoffman accompanying the paper

Value

A named list of length two. The element named 'beta' gives the proportions for the terms within the topics, while the element named 'theta' gives the proportions for the topics within the documents. If the tidy_output argument is true these are data frames in 'long' format; otherwise they are matrices.

References

Hoffman, M., Bach, FM., and Blei, DM. (2010) 'Online Learning for Latent Dirichlet Allocation', _Conference and Workshop on Neural Information Processing Systems_

Hoffman, M., Blei, DM., Wang, C, and Paisley, J. (2013) 'Stochastic Variational Inference', _Journal of Machine Learning Research_. Preprint: https://arxiv.org/abs/1206.7051_

Examples

```
library(topicmodels)
data(AssociatedPress)
ap_lda_fit <- lda_svi(AssociatedPress,passes=1,K=50)
#I use a single pass because CRAN requires examples to run quickly;
#generally one would use more. 20 often seems to be sufficient as a rule of thumb,
#but it might be worth experimenting with more or fewer
```

Index