Package 'seededlda'

September 10, 2020

Type Package

Title Seeded-LDA for Topic Modeling

Version 0.5

Description Implements the seeded-
LDA model (Lu, Ott, Cardie & Tsou 2010) <doi:10.1109/ICDMW.2011.125> us-
ing the quanteda package and the GibbsLDA++ library for semisupervised topic modeling.
Seeded-LDA allows users to pre-define topics with keywords to perform theory-driven analy-
sis of textual data in social sciences and humanities (Watan-
abe & Zhou 2020) <doi:10.1177/0894439320907027>.

License GPL-3

URL https://github.com/koheiw/seededlda

BugReports https://github.com/koheiw/seededlda/issues

Encoding UTF-8

LazyData true

Depends R (>= 3.5.0), quanteda (> 2.0), methods

Imports Matrix

LinkingTo Rcpp, RcppParallel, RcppArmadillo (>= 0.7.600.1.0), quanteda

Suggests testthat, quanteda.textmodels

RoxygenNote 7.1.1

NeedsCompilation yes

Author Kohei Watanabe [aut, cre, cph],
Phan Xuan-Hieu [aut, cph] (GibbsLDA++)

Maintainer Kohei Watanabe <watanabe.kohei@gmail.com>

Repository CRAN

Date/Publication 2020-09-10 09:00:03 UTC

R topics documented:

`terms` *Extract most likely terms*

Description

Extract most likely terms

Usage

```
terms(x, n = 10)
```

Arguments

`x`	a fitted LDA model
`n`	number of terms to be extracted

`textmodel_lda` *Semisupervised Latent Dirichlet allocation*

Description

`textmodel_seededlda()` implements semisupervised Latent Dirichlet allocation (seeded-LDA). The estimator's code adopted from the GibbsLDA++ library (Xuan-Hieu Phan, 2007). `textmodel_seededlda()` allows identification of pre-defined topics by semisupervised learning with a seed word dictionary.

Usage

```
textmodel_lda(
  x,
  k = 10,
  max_iter = 2000,
  alpha = NULL,
  beta = NULL,
  verbose = quanteda_options("verbose")
)

textmodel_seededlda(
  x,
  dictionary,
  valuetype = c("glob", "regex", "fixed"),
  case_insensitive = TRUE,
  residual = FALSE,
  weight = 0.01,
  max_iter = 2000,
```

```
  alpha = NULL,
  beta = NULL,
  verbose = quanteda_options("verbose")
)
```

Arguments

x	the dfm on which the model will be fit
k	the number of topics
max_iter	the maximum number of iteration in Gibbs sampling.
alpha	the hyper parameter for topic-document distribution
beta	the hyper parameter for topic-word distribution
verbose	logical; if TRUE print diagnostic information during fitting.
dictionary	a quanteda::dictionary() with seed words as examples of topics.
valuetype	see quanteda::valuetype
case_insensitive	see quanteda::valuetype
residual	if TRUE a residual topic (or "garbage topic") will be added to user-defined topics.
weight	pseudo count given to seed words as a proportion of total number of words in x.

References

Lu, Bin et al. (2011). Multi-aspect Sentiment Analysis with Topic Models. *Proceedings of the 2011 IEEE 11th International Conference on Data Mining Workshops*.

Watanabe, Kohei & Zhou, Yuan (2020). Theory-Driven Analysis of Large Corpora: Semisupervised Topic Classification of the UN Speeches. *Social Science Computer Review*.

See Also

topicmodels

Examples

```
require(quanteda)

data("data_corpus_moviereviews", package = "quanteda.textmodels")
corp <- head(data_corpus_moviereviews, 500)
dfmt <- dfm(corp, remove_number = TRUE) %>%
    dfm_remove(stopwords('en'), min_nchar = 2) %>%
    dfm_trim(min_termfreq = 0.90, termfreq_type = "quantile",
             max_docfreq = 0.1, docfreq_type = "prop")

# unsupervised LDA
lda <- textmodel_lda(dfmt, 6)
terms(lda)
```

```
# semisupervised LDA
dict <- dictionary(list(people = c("family", "couple", "kids"),
                        space = c("areans", "planet", "space"),
                        moster = c("monster*", "ghost*", "zombie*"),
                        war = c("war", "soldier*", "tanks"),
                        crime = c("crime*", "murder", "killer")))
slda <- textmodel_seededlda(dfmt, dict, residual = TRUE)
terms(slda)
```

`topics` *Extract most likely topics*

Description

Extract most likely topics

Usage

```
topics(x)
```

Arguments

x	a fitted LDA model

Index