

Package ‘rJST’

November 14, 2018

Type Package

Title Joint Sentiment Topic Modelling

Version 1.0

Date 2018-11-02

Maintainer Max Boiten <max_boiten@live.nl>

Description Estimates the Joint Sentiment Topic model and its reversed variety, as described by Lin and He, 2009 <DOI:10.1145/1645953.1646003> and Lin, He, Everson and Ruger (2012) <DOI:10.1109/TKDE.2011.48>.

License GPL-3

Imports magrittr, quanteda, reshape2, Rcpp (>= 0.12.12), RcppProgress, SnowballC

Depends methods

Suggests knitr, rmarkdown

LinkingTo Rcpp, RcppArmadillo, RcppProgress

Collate 'RcppExports.R' 'dictionary.R' 'topNwords.R' 'jst_reversed.R' 'jst.R' 'get_parameter.R' 'rJST.R'

Encoding UTF-8

RoxygenNote 6.0.1

VignetteBuilder knitr

NeedsCompilation yes

Author Max Boiten [cre, aut],
Chenghua Lin [cph],
Yulan He [cph]

Repository CRAN

Date/Publication 2018-11-14 09:10:03 UTC

R topics documented:

rJST-package	2
dictionary_wordstem	2
get_parameter	3
is.JST.result	4
is.JST_reversed.result	4
jst	5
JST.result-class	6
jst_reversed	6
JST_reversed.result-class	8
paradigm	8
top20words	9
topNwords	10

Index	11
--------------	-----------

rJST-package	<i>rJST: Joint Sentiment Topic modelling for R</i>
--------------	--

Description

rJST performs Joint Sentiment Topic modelling and includes a reversed model (for info on the latter, email me) and several methods to evaluate the results and make them more accessible. All models take **quanteda** dfm objects as inputs. Furthermore, all models have a defined tidy method. See **broom** for more information.

Author(s)

Maintainer: Max Boiten <max_boiten@live.nl>

Other contributors:

- Chenghua Lin [copyright holder]
- Yulan He [copyright holder]

dictionary_wordstem	<i>Wordstem a quanteda Dictionary2 object</i>
---------------------	---

Description

Applies Porter stemming to the words in a quanteda dictionary and then removes duplicates, if the stemming created these.

Usage

```
dictionary_wordstem(dict)
```

Arguments

dict A quanteda dictionary2 object

Value

A quanteda dictionary2 object with all elements stemmed

get_parameter *Get parameter from (reversed) JST results*

Description

Take the results object for the selected parameter and return a data.frame which is similar to tidy standards.

Usage

```
get_parameter(x, parameter = NULL)
```

Arguments

x A JST_reversed.result or JST.result object

parameter Character. The parameter to be tidied and returned. Note that no default is set.

Value

A data.frame.

Examples

```
data <- quanteda::dfm(quanteda::data_corpus_irishbudget2010)
model <- jst(data, paradigm(), numTopics = 5, numIters = 50)

phi <- get_parameter(model, 'phi')
```

is.JST.result

Check if an object is a JST.result object

Description

Check if an object is a JST.result object

Usage

is.JST.result(x)

Arguments

x object

Value

Boolean. True if x is a JST.result object.

is.JST_reversed.result

Check if an object is a JST_reversed.result object

Description

Check if an object is a JST_reversed.result object

Usage

is.JST_reversed.result(x)

Arguments

x object

Value

Boolean. True if x is a JST_reversed.result object.

jst

*Run a Joint Sentiment Topic model***Description**

Estimates a joint sentiment topic model using a Gibbs sampler, see Details for model description.

Usage

```
jst(dfm, sentiLexInput = list(), numSentiLabs = 3, numTopics = 10,
    numIters = 3, updateParaStep = -1, alpha = -1, beta = -1,
    gamma = -1, excludeNeutral = FALSE)
```

Arguments

dfm	A quanteda dfm object
sentiLexInput	Optional: A quanteda dictionary object for semi-supervised learning. If a dictionary is used, numSentiLabs will be overridden by the number of categories in the dictionary object. An extra category will by default be added for neutral words. This can be turned off by setting excludeNeutral = TRUE.
numSentiLabs	Integer, the number of sentiment labels (defaults to 3)
numTopics	Integer, the number of topics (defaults to 10)
numIters	Integer, the number of iterations (defaults to 3 for test runs, optimize by hand)
updateParaStep	Integer. The number of iterations between optimizations of hyperparameter alpha
alpha	Double, hyperparameter for (defaults to .05 * (average docsize/number of sentiment topics))
beta	Double, hyperparameter for (defaults to .01, with multiplier .9/.1 for sentiment dictionary presence)
gamma	Double, hyperparameter for (defaults to .05 * (average docsize/number of sentiment categories))
excludeNeutral	Boolean. If a dictionary is used, an extra category is added for neutral words. Words in the dictionary receive a low probability of being allocated there. If this is set to TRUE, the neutral sentiment category will be omitted. The variable is irrelevant if no dictionary is used. Defaults to FALSE.

Details

Basic model description:

Lin, C. and He, Y., 2009, November. Joint sentiment/topic model for sentiment analysis. In Proceedings of the 18th ACM conference on Information and knowledge management (pp. 375-384). ACM.

Weak supervision adopted from:

Lin, C., He, Y., Everson, R. and Ruger, S., 2012. Weakly supervised joint sentiment-topic detection from text. IEEE Transactions on Knowledge and Data engineering, 24(6), pp.1134-1145.

Value

A JST.result object containing a data.frame for each estimated parameter

Examples

```
model <- jst(quanteda::dfm(quanteda::data_corpus_irishbudget2010),
             paradigm(),
             numTopics = 5,
             numIters = 150)
```

JST.result-class	<i>JST results object</i>
------------------	---------------------------

Description

Contains estimated Joint Sentiment Topic model

Slots

pi Document-level sentiment estimates
 theta Document-level sentitopic estimates
 phi Word-level sentitopic estimates
 phi.termScores Word-level term scores (suboptimal calculation, only useful for smaller models)
 numTopics Number of topics
 numSentiments Number of sentiment categories
 docvars Document-level metadata from the quanteda object used as input

jst_reversed	<i>Run a reversed Joint Sentiment Topic model</i>
--------------	---

Description

Estimates a reversed joint sentiment topic model using a Gibbs sampler, see Details for model description.

Usage

```
jst_reversed(dfm, sentiLexInput = list(), numSentiLabs = 3,
             numTopics = 10, numIters = 3, updateParaStep = -1, alpha = -1,
             beta = -1, gamma = -1, excludeNeutral = FALSE)
```

Arguments

dfm	A quanteda dfm object
sentiLexInput	Optional: A quanteda dictionary object for semi-supervised learning. If a dictionary is used, numSentiLabs will be overridden by the number of categories in the dictionary object. An extra category will by default be added for neutral words. This can be turned off by setting excludeNeutral = TRUE.
numSentiLabs	Integer, the number of sentiment labels (defaults to 3)
numTopics	Integer, the number of topics (defaults to 10)
numIters	Integer, the number of iterations (defaults to 3 for test runs, optimize by hand)
updateParaStep	Integer. The number of iterations between optimizations of hyperparameter alpha
alpha	Double, hyperparameter for (defaults to .05*(average docsize/number of topics))
beta	Double, hyperparameter for (defaults to .01, with multiplier .9/.1 for sentiment dictionary presence)
gamma	Double, hyperparameter for (defaults to .05 * (average docsize/number of senti-topics))
excludeNeutral	Boolean. If a dictionary is used, an extra category is added for neutral words. Words in the dictionary receive a low probability of being allocated there. If this is set to TRUE, the neutral sentiment category will be omitted. The variable is irrelevant if no dictionary is used. Defaults to FALSE.

Details

Lin, C., He, Y., Everson, R. and Ruger, S., 2012. Weakly supervised joint sentiment-topic detection from text. *IEEE Transactions on Knowledge and Data engineering*, 24(6), pp.1134-1145.

Value

A JST_reversed.result object containing a data.frame for each estimated parameter

Examples

```
model <- jst(quanteda::dfm(quanteda::data_corpus_irishbudget2010),
            paradigm(),
            numTopics = 5,
            numIters = 150)
```

JST_reversed.result-class

Reverse JST results object

Description

Contains estimated reversed Joint Sentiment Topic model, see Details for model description.

Slots

pi Document-level topic estimates

theta Document-level sentitopic estimates

phi Word-level sentitopic estimates

phi.termScores Word-level term scores (suboptimal calculation, only useful for smaller models)

numTopics Number of topics

numSentiments Number of sentiment categories

docvars Document-level metadata from the quanteda object used as input

paradigm

Paradigm word list (Lin and He, 2009, p.379)

Description

Returns the paradigm word list proposed by Lin and He. These words create groupings of positive and negative words which are considered to be domain-independent. However, as always, caution is warranted. The words are mostly suited for review analysis. Review the words before applying the dictionary.

Usage

paradigm()

Value

A quanteda dictionary2 object containing the paradigm word list

Note

Lin, C., & He, Y. (2009). Joint sentiment/topic model for sentiment analysis. In Proceedings of the 18th ACM conference on Information and knowledge management (pp. 375-384). ACM.

top20words	<i>Show top 20 words for topics/sentiments</i>
------------	--

Description

This method returns a vector containing the 20 words with the highest estimated parameter values for any of the models estimated in this package. If topic (and sentiment for the appropriate models) are not specified, the top 20 words of every topic will be returned. This method calls the generic [topNwords](#) method.

Usage

```
top20words(x, topic = NULL, sentiment = NULL)

## S4 method for signature 'JST_reversed.result,numeric,numeric'
top20words(x, topic = NULL,
  sentiment = NULL)

## S4 method for signature 'JST_reversed.result,ANY,ANY'
top20words(x)

## S4 method for signature 'JST.result,numeric,numeric'
top20words(x, topic = NULL,
  sentiment = NULL)

## S4 method for signature 'JST.result,ANY,ANY'
top20words(x)
```

Arguments

x	A results object from any of the models in the package
topic	(optional) Integer. The topic to return words from.
sentiment	(optional) Integer. The sentiment to return words from.

Value

If topic and sentiment are specified: A character vector containing the top 20 words for the requested topic-sentiment combination. Otherwise a data.frame containing the top 20 words for every topic-sentiment combination.

Examples

```
model <- jst(quanteda::data_dfm_lbgexample, paradigm())
top20words(model, topic = 1, sentiment = 1)
```

topNwords	<i>Show top N words for topics/sentiments</i>
-----------	---

Description

This method returns a vector containing the N words with the highest estimated parameter values for any of the models estimated in this package. If topic (and sentiment for the appropriate models) are not specified, the top N words of every topic will be returned.

Usage

```
topNwords(x, N, topic = NULL, sentiment = NULL)

## S4 method for signature 'JST_reversed.result,numeric,numeric,numeric'
topNwords(x, N,
  topic = NULL, sentiment = NULL)

## S4 method for signature 'JST_reversed.result,numeric,ANY,ANY'
topNwords(x, N)

## S4 method for signature 'JST.result,numeric,numeric,numeric'
topNwords(x, N, topic = NULL,
  sentiment = NULL)

## S4 method for signature 'JST.result,numeric,ANY,ANY'
topNwords(x, N)
```

Arguments

x	A results object from any of the models in the package
N	Integer. The number of words to return.
topic	(optional) Integer. The topic to return words from.
sentiment	(optional) Integer. The sentiment to return words from.

Value

If topic and sentiment are specified: A character vector containing the top N words for the requested topic-sentiment combination. Otherwise a data.frame containing the top N words for every topic-sentiment combination.

Examples

```
model <- jst(quanteda::data_dfm_lbgexample, paradigm())
topNwords(model, N = 30, topic = 2, sentiment = 1)
```

Index

dictionary_wordstem, 2

get_parameter, 3

is.JST.result, 4

is.JST_reversed.result, 4

jst, 5

JST.result (JST.result-class), 6

JST.result-class, 6

jst_reversed, 6

JST_reversed.result
(JST_reversed.result-class), 8

JST_reversed.result-class, 8

paradigm, 8

rJST (rJST-package), 2

rJST-package, 2

top20words, 9

top20words, JST.result, ANY, ANY-method
(top20words), 9

top20words, JST.result, numeric, numeric-method
(top20words), 9

top20words, JST.result-method
(top20words), 9

top20words, JST_reversed.result, ANY, ANY-method
(top20words), 9

top20words, JST_reversed.result, numeric, numeric-method
(top20words), 9

top20words, JST_reversed.result-method
(top20words), 9

topNwords, 9, 10

topNwords, JST.result, numeric, -method
(topNwords), 10

topNwords, JST.result, numeric, ANY, ANY-method
(topNwords), 10

topNwords, JST.result, numeric, numeric, numeric-method
(topNwords), 10

topNwords, JST_reversed.result, numeric, -method
(topNwords), 10

topNwords, JST_reversed.result, numeric, ANY, ANY-method
(topNwords), 10

topNwords, JST_reversed.result, numeric, numeric, numeric-method
(topNwords), 10