Package ‘stylest2’

March 23, 2024

Title Estimating Speakers of Texts

Version 0.1

Description Estimates the authors or speakers of texts. Methods developed in Huang, Perry, and Spirling (2020) <doi:10.1017/pan.2019.49>. The model is built on a Bayesian framework in which the distinctiveness of each speaker is defined by how different, on average, the speaker's terms are to everyone else in the corpus of texts. An optional cross-validation method is implemented to select the subset of terms that generate the most accurate speaker predictions. Once a set of terms is selected, the model can be estimated. Speaker distinctiveness and term influence can be recovered from parameters in the model using package functions. Once fitted, the model can be used to predict authorship of new texts.

Depends R (>= 4.2),

License GPL-3

Imports Matrix, quanteda

Suggests devtools, knitr, rmarkdown, testthat

Collate 'stylest2_select_vocab.R' 'stylest2_fit.R'
'stylest2_predict.R' 'data.R' 'stylest2.R'

Encoding UTF-8

VignetteBuilder knitr, rmarkdown

RoxygenNote 7.3.1

NeedsCompilation no

Author Christian Baehr [aut, cre, cph],
Arthur Spirling [aut, cph],
Leslie Huang [aut]

Maintainer Christian Baehr <cbaehr@princeton.edu>

Repository CRAN

Date/Publication 2024-03-23 10:50:06 UTC

R topics documented:

`novels` *Excerpts from English novels*

Description

A dataset of text from English novels by Jane Austen, George Eliot, and Elizabeth Gaskell.

Usage

```
data(novels)
```

Format

A dataframe with 21 rows and 3 variables.

Source

Novel excerpts obtained from Project Gutenberg full texts in the public domain in the USA. http://gutenberg.org

`novels_dfm` *Novel excerpts in quanteda dfm object*

Description

A dataset of text from English novels by Jane Austen, George Eliot, and Elizabeth Gaskell. It has been tokenized and processed as a document-feature matrix in quanteda.

Usage

```
data(novels_dfm)
```

Format

A quanteda `dfm` with a document variable titled "author".

Source

Novel excerpts obtained from Project Gutenberg full texts in the public domain in the USA. http://gutenberg.org

`stylest` *stylest2: A package for estimating authorship of texts.*

Description

stylest2 provides a set of functions for fitting a model of speaker distinctiveness, including tools for selecting the optimal vocabulary for the model and predicting the most likely speaker (author) of a new text.

`stylest2_fit` *Fit speaker model to document-feature matrix*

Description

This function generates a model of speaker/author attribution, given a document-feature matrix.

Usage

```
stylest2_fit(
  dfm,
  smoothing = 0.5,
  terms = NULL,
  term_weights = NULL,
  fill_weight = NULL
)
```

Arguments

`dfm`	a quanteda `dfm` object
`smoothing`	the smoothing parameter value for smoothing the dfm. Should be a numeric scalar, default to 0.5.
`terms`	If not `NULL`, terms to be used in the model. If `NULL`, use all terms.
`term_weights`	Named vector of distances (or any weights) per term in the vocab. Names should correspond to the term.
`fill_weight`	Numeric value to fill in as weight for any term which does not have a weight specified in `term_weights`.

Value

An S3 object, a model with with each term that occurs in the text, the frequency of use for each author, and the frequency of that terms' occurrence through the texts.

Examples

```
data(novels_dfm)
stylest2_fit(dfm = novels_dfm)
```

stylest2_predict *Predict authorship of texts.*

Description

This function generates predicted probabilities of authorship for a set of texts. It takes as an input a document-feature matrix of texts for which authorship is to be predicted, as well as a stylest2 model containing potential authors.

Usage

```
stylest2_predict(
  dfm,
  model,
  speaker_odds = FALSE,
  term_influence = FALSE,
  prior = NULL
)
```

Arguments

dfm	a quanteda dfm object. Each row should represent a text whose authorship is to be predicted.
model	A stylest2 model.
speaker_odds	Should the model return log odds of authorship for each text, in addition to posterior probabilities?
term_influence	Should the model return the influence of each term in determining authorship over the prediction set, in addition to returning posterior probabilities?
prior	Prior probability, defaults to NULL.

Value

A list object:

Examples

```
data(novels_dfm)
mod <- stylest2_fit(novels_dfm)
stylest2_predict(dfm=novels_dfm, model=mod)
```

stylest2_select_vocab *Cross-validation based term selection*

Description

K-fold cross validation to determine the optimal cutoff on the term frequency distribution under which to drop terms.

Usage

```
stylest2_select_vocab(
  dfm,
  smoothing = 0.5,
  cutoffs = c(50, 60, 70, 80, 90, 99),
  nfold = 5,
  terms = NULL,
  term_weights = NULL,
  fill = FALSE,
  fill_weight = NULL,
  suppress_warning = TRUE
)
```

Arguments

`dfm`	a quanteda `dfm` object.
`smoothing`	the smoothing parameter value for smoothing the dfm. Should be a numeric scalar, default to 0.5.
`cutoffs`	a numeric vector of cutoff candidates.
`nfold`	number of folds for the cross-validation
`terms`	If not `NULL`, terms to be used in the model. If `NULL`, use all terms.
`term_weights`	Named vector of distances (or any weights) per term in the vocab. Names should correspond to the term.
`fill`	Should missing values in term weights be filled? Defaults to FALSE.
`fill_weight`	Numeric value to fill in as weight for any term which does not have a weight specified in `term_weights`.
`suppress_warning`	TRUE/FALSE, indicate whether to suppress warnings from `stylest2_fit()`.

Value

List of: best cutoff percent with the best speaker classification rate; cutoff percentages that were tested; matrix of the mean percentage of incorrectly identified speakers for each cutoff percent and fold; and the number of folds for cross-validation.

Examples

```
data(novels_dfm)
stylest2_select_vocab(dfm=novels_dfm)
```

stylest2_terms *Select terms above frequency cutoff*

Description

A function to select terms for inclusion in a stylest2 model, based on a document-feature matrix of texts to predict and a specified cutoff.

Usage

```
stylest2_terms(dfm, cutoff)
```

Arguments

`dfm`	a quanteda `dfm` object.
`cutoff`	a single numeric value - the quantile of term frequency under which to drop terms.

Value

A character vector of terms falling above the term frequency cutoff.

Examples

```
data(novels_dfm)
best_cut <- stylest2_select_vocab(dfm=novels_dfm)
stylest2_terms(dfm = novels_dfm, cutoff=best_cut$cutoff_pct_best)
```

Index