Package ‘newsmap’
October 7, 2023

Type Package
Title Semi-Supervised Model for Geographical Document Classification
Version 0.8.3
Maintainer Kohei Watanabe <watanabe.kohei@gmail.com>
This package currently contains seed dictionaries in English, German, French, Spanish, Italian, Russian, Hebrew, Arabic Japanese and Chinese (Simplified and Traditional).
License MIT + file LICENSE
URL https://github.com/koheiw/newsmap
BugReports https://github.com/koheiw/newsmap/issues
LazyData TRUE
Encoding UTF-8
Depends R (>= 3.5), methods
Imports utils, Matrix, quanteda (>= 2.1), quanteda.textstats, stringi
Suggests testthat
Language en-GB
RoxygenNote 7.2.3
NeedsCompilation no
Author Kohei Watanabe [aut, cre, cph],
  Stefan Müller [aut],
  Dani Madrid-Morales [aut],
  Katerina Tertychnaya [aut],
  Ke Cheng [aut],
  Chung-hong Chan [aut],
  Claude Grasland [aut],
  Giuseppe Carteny [aut],
  Elad Segev [aut],
  Dai Yamao [aut],
  Barbara Ellynes Zucchi Nobre Silva [aut],
  Lanabi la Lova [aut],
**accuracy**

Evaluate classification accuracy in precision and recall

**Usage**

```r
accuracy(x, y)
```

**Arguments**

- `x`: vector of predicted classes
- `y`: vector of true classes

**Examples**

```r
class_pred <- c('US', 'GB', 'US', 'CN', 'JP', 'FR', 'CN') # prediction
class_true <- c('US', 'FR', 'US', 'CN', 'KP', 'EG', 'US') # true class
acc <- accuracy(class_pred, class_true)
print(acc)
summary(acc)
```
afe  

Compute average feature entropy (AFE)

Description
AFE computes randomness of occurrences features in labelled documents.

Usage
afe(x, y, smooth = 1)

Arguments
- x: a dfm for features
- y: a dfm for labels
- smooth: a numeric value for smoothing to include all the features

data_dictionary_newsmap_ar
Seed geographical dictionary in Arabic

Description
Seed geographical dictionary in Arabic

Author(s)
Dai Yamao <daiyamao@scs.kyushu-u.ac.jp>

data_dictionary_newsmap_de
Seed geographical dictionary in German

Description
Seed geographical dictionary in German

Author(s)
Stefan Müller <mullers@tcd.ie>
data_dictionary_newsmap_en

*Seed geographical dictionary in English*

**Description**

Seed geographical dictionary in English

**Author(s)**

Kohei Watanabe <watanabe.kohei@gmail.com>

data_dictionary_newsmap_es

*Seed geographical dictionary in Spanish*

**Description**

Seed geographical dictionary in Spanish

**Author(s)**

Dani Madrid-Morales <dani.madrid@my.cityu.edu.hk>

data_dictionary_newsmap_fr

*Seed geographical dictionary in French*

**Description**

Seed geographical dictionary in French

**Author(s)**

Claude Grasland <claude.grasland@parisgeo.cnrs.fr>
data_dictionary_newsmap_he

Seed geographical dictionary in Hebrew

Description
Seed geographical dictionary in Hebrew

Author(s)
Elad Segev <eladseg@gmail.com>

data_dictionary_newsmap_it

Seed geographical dictionary in Italian

Description
Seed geographical dictionary in Italian

Author(s)
Giuseppe Carteny <giuseppe.carteny@unimi.it>

data_dictionary_newsmap_ja

Seed geographical dictionary in Japanese

Description
Seed geographical dictionary in Japanese

Author(s)
Kohei Watanabe <watanabe.kohei@gmail.com>
**Description**

Seed geographical dictionary in Portuguese

**Author(s)**

Barbara Ellynes Zucchi Nobre Silva <barbara@zucchi.science>

---

**Description**

Seed geographical dictionary in Russian

**Author(s)**

Katerina Tertychnaya <katerina.tertychnaya@gmail.com>
Lanabi la Lova <S.Bilalova@lse.ac.uk>

---

**Description**

Seed geographical dictionary in Chinese (simplified)

**Author(s)**

Ke Cheng <kecheng.ac@gmail.com>
Description

Seed geographical dictionary in Chinese (traditional)

Author(s)

Chung-hong Chan <chainsawtiney@gmail.com>

predict.textmodel_newsmap

Prediction method for textmodel_newsmap

Description

Predict document class using trained a Newsmap model

Usage

## S3 method for class 'textmodel_newsmap'
predict(
  object,
  newdata = NULL,
  confidence = FALSE,
  rank = 1L,
  type = c("top", "all"),
  rescale = FALSE,
  min_conf = -Inf,
  min_n = 0L,
  ...
)

Arguments

object a fitted Newsmap textmodel.
newdata dfm on which prediction should be made.
confidence if TRUE, it returns likelihood ratio score.
rank rank of the class to be predicted. Only used when type = "top".
type if top, returns the most likely class specified by rank; otherwise return a matrix of likelihood ratio scores for all possible classes.
rescale if TRUE, likelihood ratio scores are normalized using `scale()`. This affects both types of results.

min_conf return NA when confidence is lower than this value.

min_n set the minimum number of polarity words in documents.

... not used.

summary.textmodel_newsmap_accuracy

*Calculate micro and macro average measures of accuracy*

**Description**

This function calculates micro-average precision (p) and recall (r) and macro-average precision (P) and recall (R) based on a confusion matrix from `accuracy()`.

**Usage**

```r
## S3 method for class 'textmodel_newsmap_accuracy'
summary(object, ...)
```

**Arguments**

- `object` output of `accuracy()`
- `...` not used.

**textmodel_newsmap**  
*Semi-supervised Bayesian multinomial model for geographical document classification*

**Description**

Train a Newsmap model to predict geographical focus of documents with labels given by a dictionary.

**Usage**

```r
textmodel_newsmap(
  x, 
  y, 
  label = c("all", "max"), 
  smooth = 1, 
  drop_label = TRUE, 
  verbose = quanteda_options("verbose"), 
  entropy = c("none", "global", "local", "average"), 
  ...
)
```
Arguments

- **x**: a dfm or fcm created by `quanteda::dfm()`.
- **y**: a dfm or a sparse matrix that record class membership of the documents. It can be created applying `quanteda::dfm_lookup()` to x.
- **label**: if "max", uses only labels for the maximum value in each row of y.
- **smooth**: a value added to the frequency of words to smooth likelihood ratios.
- **drop_label**: if TRUE, drops empty columns of y and ignore their labels.
- **verbose**: if TRUE, shows progress of training.
- **entropy**: [experimental] the scheme to compute the entropy to regularize likelihood ratios. The entropy of features are computed over labels if global or over documents with the same labels if local. Local entropy is averaged if average. See the details.
- **...**: additional arguments passed to internal functions.

Details

Newsmap learns association between words and classes as likelihood ratios based on the features in x and the labels in y. The large likelihood ratios tend to concentrate to a small number of features but the entropy of their frequencies over labels or documents helps to disperse the distribution.

References


Examples

```r
require(quanteda)
text_en <- c(text1 = "This is an article about Ireland.",
             text2 = "The South Korean prime minister was re-elected.")
toks_en <- tokens(text_en)
label_toks_en <- tokens_lookup(toks_en, data_dictionary_newsmap_en, levels = 3)
label_dfm_en <- dfm(label_toks_en)

feat_dfm_en <- dfm(toks_en, tolower = FALSE)

model_en <- textmodel_newsmap(feat_dfm_en, label_dfm_en)
predict(model_en)
```
Index

* data
  data_dictionary_newsmap_ar, 3
  data_dictionary_newsmap_de, 3
  data_dictionary_newsmap_en, 4
  data_dictionary_newsmap_es, 4
  data_dictionary_newsmap_fr, 4
  data_dictionary_newsmap_he, 5
  data_dictionary_newsmap_it, 5
  data_dictionary_newsmap_ja, 5
  data_dictionary_newsmap_pt, 6
  data_dictionary_newsmap_ru, 6
  data_dictionary_newsmap_zh_cn, 6
  data_dictionary_newsmap_zh_tw, 7

accuracy, 2
afe, 3

predict.textmodel_newsmap, 7
quanteda::dfm(), 9
quanteda::dfm_lookup(), 9

scale(), 8
summary.textmodel_newsmap_accuracy, 8

textmodel_newsmap, 8