Package 'sentometrics'

November 13, 2017

Type Package

Title An Integrated Framework for Textual Sentiment Time Series Aggregation and Prediction

Version 0.2

Author David Ardia [aut],
Keven Bluteau [aut],
Samuel Borms [aut, cre],
Kris Boudt [aut]

Maintainer Samuel Borms <samuel.borms@unine.ch>

Description Time series analysis based on textual sentiment, accounting for the intrinsic challenge that sentiment can be computed and pooled across texts and time in many ways. As described in Ardia et al. (2017) <https://ssrn.com/abstract=3067734>, the package provides a means to model the impact of sentiment in texts on a target variable, by first computing a wide range of textual sentiment measures and then selecting those that are most informative.

Depends R (>= 3.4.2), data.table, ggplot2, foreach

License GPL (>= 2)

BugReports https://github.com/sborms/sentometrics/issues

URL https://github.com/sborms/sentometrics

Encoding UTF-8

LazyData true

Suggests testthat, e1071, randomForest

Imports utils, stats, quanteda, sentimentr, stringi, zoo, abind, glmnet, caret, compiler, Rcpp (>= 0.12.13), RcppRoll, ggthemes, ISOweek, MCS

LinkingTo Rcpp, RcppArmadillo

RoxygenNote 6.0.1

NeedsCompilation yes

Repository CRAN

Date/Publication 2017-11-13 10:34:52 UTC

R topics documented:

`sentometrics-package` *An Integrated Framework for Textual Sentiment Time Series Aggregation and Prediction*

Description

The **sentometrics** package is designed to do time series analysis based on textual sentiment. It accounts for the intrinsic challenge that, for a given text, sentiment can be computed in many ways, as well as the large number of possibilities to pool sentiment across text and time. This additional layer of manipulation does not exist in standard time series analysis and text mining packages. As a final outcome, this package provides an automated means to econometrically model the impact of sentiment in texts on a given variable, by first computing a wide range of textual sentiment time series and then selecting those that are most informative. The package created therefore integrates the *qualification* of sentiment from texts, the *aggregation* into different sentiment measures and the optimized *prediction* based on these measures.

Main functions

- Sentiment computation and aggregation into sentiment measures: `sento_corpus`, `ctr_agg`, `compute_sentiment`, `sento_measures`, `to_global`
- Sparse modelling: `ctr_model`, `sento_model`
- Prediction and post-modelling analysis: `predict.sentomodel`, `retrieve_attributions`, `perform_MCS`

Update

The development version of the package is available at https://github.com/sborms/sentometrics.

Note

The methodology behind the sentiment aggregation framework can be consulted in the working paper "Questioning the news about economic growth: Sparse forecasting using thousands of news-based sentiment values" (Ardia, Bluteau, and Boudt, 2017) at https://ssrn.com/abstract=2976084. The vignette "The R package sentometrics to compute, aggregate and predict with textual sentiment" (Ardia, Bluteau, Borms and Boudt, 2017) at https://ssrn.com/abstract=3067734 provides a hands-on introduction to the methodology and the package's functionalities.

Please cite the package in publications. Use `citation("sentometrics")`.

Author(s)

Maintainer: Samuel Borms <samuel.borms@unine.ch>

Authors:

- David Ardia <david.ardia@unine.ch>
- Keven Bluteau <keven.bluteau@unine.ch>
- Kris Boudt <kris.boudt@vub.be>

See Also

Useful links:

- https://github.com/sborms/sentometrics
- Report bugs at https://github.com/sborms/sentometrics/issues

add_features *Add feature columns to a sentocorpus*

Description

Adds new feature columns, either user-supplied or based on a simple keyword(s) search, to a provided sentocorpus object.

Usage

```
add_features(sentocorpus, featuresdf = NULL, keywords = NULL)
```

Arguments

sentocorpus	a sentocorpus object created with sento_corpus.
featuresdf	a named data.frame of type numeric where each columns is a new feature to be added to the inputted sentocorpus object. If the number of rows in featuresdf is not equal to the number of documents in sentocorpus, recycling will occur.
keywords	a named list. For every element, a new feature column is added with a value of 1 for the texts in which the keyword(s) appear(s), and 0 if not. If no texts match a keyword, no column is added. The list named elements are used as the names of the new features.

Details

If a provided feature name is already part of the corpus, it will be replaced. The featuresdf and keywords arguments can be provided at the same time, or only one of them, leaving the other at NULL.

Value

An updated sentocorpus object.

Author(s)

Samuel Borms

Examples

```
data("usnews")

# construct a corpus and add random features to it
corpus <- sento_corpus(corpusdf = usnews)
corpus1 <- add_features(corpus,
                        featuresdf = data.frame(random = runif(quanteda::ndoc(corpus))))
corpus2 <- add_features(corpus,
                        keywords = list(pres = "president", war = "war"))
```

`almons` *Compute Almon polynomials*

Description

Computes Almon polynomial weighting curves. Handy to self-select specific time aggregation weighting schemes for input in `ctr_agg`.

Usage

```
almons(n, orders = 1:3, do.inverse = TRUE, do.normalize = TRUE)
```

Arguments

`n`	a single `numeric` to indicate the length of the curve (the number of lags, cf., n).
`orders`	a `numeric` vector as the sequence of the Almon orders (cf., b). The maximum value corresponds to B.
`do.inverse`	`TRUE` if the inverse Almon polynomials should be calculated as well.
`do.normalize`	`TRUE` if polynomials should be normalized to unity.

Details

The Almon polynomial formula implemented is: $(1 - (i/n)^b)(i/n)^{B-b}$, where i is the lag index from 1 to n.

Value

A `data.frame` of all Almon polynomial weighting curves, of size `length(orders)` (times two if `do.inverse = TRUE`).

See Also

`ctr_agg`

`compute_sentiment` *Compute document-level sentiment across features and lexicons*

Description

Given a corpus of texts, computes sentiment per document using the bag-of-words approach based on the lexicons provided and a choice of aggregation across words per document. Relies partly on the **quanteda** package. The scores computed are net sentiment (sum of positive minus sum of negative scores).

Usage

```
compute_sentiment(sentocorpus, lexicons, how = get_hows()$words, dfm = NULL)
```

Arguments

sentocorpus	a sentocorpus object created with sento_corpus.
lexicons	output from a setup_lexicons call.
how	a single character vector defining how aggregation within documents should be performed. For currently available options on how aggregation can occur, see get_hows()$words.
dfm	optional; an output from a **quanteda** dfm call, such that users can specify their own tokenization scheme (via tokens) as well as other parameters related to the construction of a document-feature matrix (dfm). By default, a dfm is created based on a tokenization that removes punctuation, numbers, symbols and separators. We suggest to stick to unigrams, as the remainder of the sentiment computation and built-in lexicons assume the same.

Details

For a separate calculation of positive (resp. negative) sentiment, one has to provide distinct positive (resp. negative) lexicons. This can be done using the do.split option in the setup_lexicons function, which splits out the lexicons into a positive and a negative polarity counterpart. NAs are converted to 0, under the assumption that this is equivalent to no sentiment.

Value

A list containing:

corpus	the supplied sentocorpus object; the texts are altered if valence shifters are part of the lexicons.
sentiment	the sentiment scores data.table with a "date" and all lexicon–feature sentiment scores columns.
features	a character vector of the different features.
lexicons	a character vector of the different lexicons used.
howWithin	a character vector to remind how sentiment within documents was aggregated.

Author(s)

Samuel Borms

See Also

dfm, tokens

Examples

```
data("usnews")
data("lexicons")
data("valence")

# sentiment computation based on raw frequency counts
corpus <- sento_corpus(corpusdf = usnews)
corpusSample <- quanteda::corpus_sample(corpus, size = 1000)
l <- setup_lexicons(lexicons[c("LM_eng", "HENRY_eng")], valence[["valence_eng"]])
sent <- compute_sentiment(corpusSample, l, how = "counts")

## Not run:
# same sentiment computation based on a user-supplied dfm with default settings
dfm <- quanteda::dfm(quanteda::tokens(corpus), verbose = FALSE)
sent <- compute_sentiment(corpusSample, l, how = "counts", dfm = dfm)
## End(Not run)
```

---

ctr_agg *Set up control for aggregation into sentiment measures*

Description

Sets up control object for aggregation of document-level textual sentiment into textual sentiment measures (indices).

Usage

```
ctr_agg(howWithin = "proportional", howDocs = "equal_weight",
  howTime = "equal_weight", do.ignoreZeros = TRUE, by = "day", lag = 1L,
  fill = "zero", alphasExp = seq(0.1, 0.5, by = 0.1), ordersAlm = 1:3,
  do.inverseAlm = TRUE, do.normalizeAlm = TRUE, weights = NULL,
  dfm = NULL)
```

Arguments

howWithin	a single `character` vector defining how aggregation within documents will be performed. Should `length(howWithin) > 1`, the first element is used. For currently available options on how aggregation can occur, see `get_hows()$words`.
howDocs	a single `character` vector defining how aggregation across documents per date will be performed. Should `length(howDocs) > 1`, the first element is used. For currently available options on how aggregation can occur, see `get_hows()$docs`.
howTime	a `character` vector defining how aggregation across dates will be performed. More than one choice is possible. For currently available options on how aggregation can occur, see `get_hows()$time`.

do.ignoreZeros	a logical indicating whether zero sentiment values have to be ignored in the determination of the document weights while aggregating across documents. By default do.ignoreZeros = TRUE, such that documents with an exact score of zero are considered irrelevant.
by	a single character vector, either "day", "week", "month" or "year", to indicate at what level the dates should be aggregated. Dates are displayed as the first day of the period, if applicable (e.g. "2017-03-01" for March 2017).
lag	a single integer vector, being the time lag to be specified for aggregation across time. By default equal to 1L, meaning no aggregation across time.
fill	a single character vector, one of c("zero", "latest", "none"), to control how missing sentiment values across the continuum of dates considered are added. This impacts the aggregation across time, applying the fill_measures function before aggregating, except if fill = "none". By default equal to "zero", which sets the scores (and thus also the weights) of the added dates to zero in the time aggregation.
alphasExp	a numeric vector of all exponential smoothing factors to calculate weights for, used if "exponential" %in% howTime. Values should be between 0 and 1 (both excluded).
ordersAlm	a numeric vector of all Almon polynomial orders to calculate weights for, used if "almon" %in% howTime.
do.inverseAlm	a logical indicating if for every Almon polynomial its inverse has to be added, used if "almon" %in% howTime.
do.normalizeAlm	a logical indicating if every Almon polynomial weights column should sum to one, used if "almon" %in% howTime.
weights	an optional own weighting scheme, always used if provided as a data.frame with the number of rows equal to the desired lag. The automatic Almon polynomials are created sequentially; if the user wants only specific of such time weighting series it can use almons, select the columns it requires, combine it into a data.frame and supply it under this argument (see examples).
dfm	optional; see compute_sentiment.

Details

For currently available options on how aggregation can occur (via the howWithin, howDocs and howTime arguments), call get_hows.

Value

A list encapsulating the control parameters.

Author(s)

Samuel Borms, Keven Bluteau

See Also

fill_measures, almons, compute_sentiment

Examples

```
# simple control function
ctr1 <- ctr_agg(howTime = "linear", by = "year", lag = 3)

# more elaborate control function (particular attention to time weighting schemes)
ctr2 <- ctr_agg(howWithin = "tf-idf",
                howDocs = "proportional",
                howTime = c("equal_weight", "linear", "almon", "exponential", "own"),
                do.ignoreZeros = TRUE,
                by = "day",
                lag = 20,
                ordersAlm = 1:3,
                do.inverseAlm = TRUE,
                do.normalizeAlm = TRUE,
                alphasExp = c(0.20, 0.50, 0.70, 0.95),
                weights = data.frame(myWeights = runif(20)))

# set up control function with one linear and two chosen Almon weighting schemes
a <- almons(n = 70, orders = 1:3, do.inverse = TRUE, do.normalize = TRUE)
ctr3 <- ctr_agg(howTime = c("linear", "own"), by = "year", lag = 70,
                weights = data.frame(a1 = a[, 1], a2 = a[, 3]))
```

`ctr_merge` *Set up control for merging sentiment measures*

Description

Sets up control object for the optional merging (additional aggregation) of sentiment measures as done by `merge_measures`.

Usage

```
ctr_merge(sentomeasures, features = NA, lexicons = NA, time = NA,
  do.keep = FALSE)
```

Arguments

`sentomeasures`	a `sentomeasures` object created using `sento_measures`. This is necessary to check whether the other input arguments make sense.
`features`	a `list` with unique features to merge at given name, e.g., `list(feat12 = c("feat1", "feat2"))`. See `sentomeasures$features` for the exact names to use. Use `NA` to apply no merging across this dimension.
`lexicons`	a `list` with unique lexicons to merge at given name, e.g., `list(lex12 = c("lex1", "lex2"))`. See `sentomeasures$lexicons` for the exact names to use. Use `NA` to apply no merging across this dimension.

time — a list with unique time weighting schemes to merge at given name, e.g., list(tw12 = c("tw1", "tw2")). See sentomeasures$time for the exact names to use. Use NA to apply no merging across this dimension.

do.keep — a logical indicating if the original sentiment measures should be kept (i.e., the merged sentiment measures will be added to the current sentiment measures as additional indices if do.keep = TRUE).

Value

A list encapsulating the control parameters.

Author(s)

Samuel Borms

See Also

merge_measures

Examples

```
data("usnews")
data("lexicons")
data("valence")

# construct a sentomeasures object to start with
corpus <- sento_corpus(corpusdf = usnews)
corpusSample <- quanteda::corpus_sample(corpus, size = 750)
l <- setup_lexicons(lexicons[c("LM_eng", "HENRY_eng")], valence[["valence_eng"]])
ctr <- ctr_agg(howTime = c("equal_weight", "linear"), by = "year", lag = 3)
sentomeasures <- sento_measures(corpusSample, l, ctr)

# set up a correct control function
ctrMerge <- ctr_merge(sentomeasures,
                      time = list(W = c("equal_weight", "linear")),
                      lexicons = list(LEX = c("LM_eng", "HENRY_eng")),
                      features = list(journals = c("wsj", "wapo")),
                      do.keep = TRUE)

## Not run:
# produces an informative error message
ctrMerge <- ctr_merge(sentomeasures,
                      time = list(W = c("equal_weight", "almon1")),
                      lexicons = list(LEX = c("LM_eng", "HENRY_eng")),
                      features = list(journals = c("notInHere", "wapo")))
## End(Not run)
```

ctr_model *Set up control for sentiment measures-based regression modelling*

Description

Sets up control object for linear or nonlinear modelling of a response variable onto a large panel of textual sentiment measures (and potentially other variables). See sento_model for details on the estimation and calibration procedure.

Usage

```
ctr_model(model = c("gaussian", "binomial", "multinomial"), type = c("BIC",
  "AIC", "Cp", "cv"), intercept = TRUE, do.iter = FALSE, h = 0,
  alphas = seq(0, 1, by = 0.2), nSample = NULL, trainWindow = NULL,
  testWindow = NULL, oos = 0, start = 1, do.progress = TRUE,
  do.parallel = FALSE)
```

Arguments

model	a character vector with one of the following: "gaussian" (linear regression), "binomial" (binomial logistic regression), or "multinomial" (multinomial logistic regression).
type	a character vector indicating which model calibration approach to use. Supports "BIC", "AIC" and "Cp" (Mallows's Cp) as sparse regression adapted information criteria (cf., "On the 'degrees of freedom' of the LASSO"; Zou, Hastie, Tibshirani et al., 2007), and "cv" (cross-validation based on the train function from the **caret** package). The adapted information criteria are currently only available for a linear regression.
intercept	a logical, TRUE by default fits an intercept.
do.iter	a logical, TRUE induces an iterative estimation of models at the given nSample size and performs the associated one-step ahead out-of-sample prediction exercise through time.
h	an integer value that shifts the time series to have the desired prediction setup; h = 0 means no change to the input data (nowcasting assuming data is aligned properly), h > 0 shifts the dependent variable by h periods (i.e. rows) further in time (forecasting), h < 0 shifts the independent variables by h periods.
alphas	a numeric vector of the different alphas to test for during calibration, between 0 and 1. A value of 0 pertains to Ridge regression, a value of 1 to LASSO regression; values in between are pure elastic net. The lambda values tested for are chosen by the glmnet function or set to 10^seq(2, -2, length.out = 100) in case of cross-validation.
nSample	a positive integer as the size of the sample for model estimation at every iteration (ignored if iter = FALSE).
trainWindow	a positive integer as the size of the training sample in cross-validation (ignored if type != "cv").

testWindow	a positive integer as the size of the test sample in cross-validation (ignored if type != "cv").
oos	a non-negative integer to indicate the number of periods to skip from the end of the cross-validation training sample (out-of-sample) up to the test sample (ignored if type != "cv").
start	a positive integer to indicate at which point the iteration has to start (ignored if iter = FALSE). For example, given 100 possible iterations, start = 70 leads to model estimations only for the last 31 samples.
do.progress	a logical, if TRUE progress statements are displayed during model calibration.
do.parallel	a logical, if TRUE the %dopar% construct from the **foreach** package is applied for iterative model estimation. A proper parallel backend needs to be set up to make it work. No progress statements are displayed whatsoever when TRUE. For cross-validation models, parallelization can also be carried out for single-run models, whenever a parallel backend is set up. See the examples in sento_model.

Value

A list encapsulating the control parameters.

Author(s)

Samuel Borms, Keven Bluteau

See Also

sento_model

Examples

```
# information criterion based model control functions
ctrIC1 <- ctr_model(model = "gaussian", type = "BIC", do.iter = FALSE, h = 0,
                    alphas = seq(0, 1, by = 0.10))
ctrIC2 <- ctr_model(model = "gaussian", type = "AIC", do.iter = TRUE, h = 0, nSample = 100)

# cross-validation based model control functions
ctrCV1 <- ctr_model(model = "gaussian", type = "cv", do.iter = FALSE, h = 0, trainWindow = 250,
                    testWindow = 4, oos = 0, do.progress = TRUE)
ctrCV2 <- ctr_model(model = "binomial", type = "cv", h = 0, trainWindow = 250,
                    testWindow = 4, oos = 0, do.progress = TRUE)
ctrCV3 <- ctr_model(model = "multinomial", type = "cv", h = 0, trainWindow = 250,
                    testWindow = 4, oos = 0, do.progress = TRUE)
ctrCV4 <- ctr_model(model = "gaussian", type = "cv", do.iter = TRUE, h = 0, trainWindow = 45,
                    testWindow = 4, oos = 0, nSample = 70, do.progress = TRUE)
```

epu *Monthly Economic Policy Uncertainty Index*

Description

Monthly values of a news-based index of U.S. Economic Policy Uncertainty (EPU) between January 1980 and September 2014, including a binomial and a multinomial example series. For more information on its calculation, see this. Following columns are present:

- date. Date as `"yyyy-mm-01"`.
- index. A `numeric` monthly index value.
- above. A `factor` with value `"above"` if the index is greater than the mean of the entire series, else `"below"`.
- aboveMulti. A `factor` with values `"above+"`, `"above"`, `"below"` and `"below-"` if the index is greater than the 75% quantile and the 50% quantile, or smaller than the 50% quantile and the 25% quantile, respectively and in a mutually exclusive sense.

Usage

```
data("epu")
```

Format

A `data.frame` with 417 rows and 4 columns.

Source

Research on Economic Policy Uncertainty

exponentials *Compute exponential weighting curves*

Description

Computes exponential weighting curves. Handy to self-select specific time aggregation weighting schemes for input in `ctr_agg`.

Usage

```
exponentials(n, alphas = seq(0.1, 0.5, by = 0.1))
```

Arguments

`n`	a single `numeric` to indicate the length of the curve (the number of lags).
`alphas`	a `numeric` vector of decay factors.

Value

A data.frame of exponential weighting curves per value of alphas.

See Also

ctr_agg

---

fill_measures *Add and fill missing dates*

Description

Adds missing dates between earliest and latest date of a sentomeasures object, such that time series is continuous date-wise. Fills in these dates with either 0, the respective latest non-missing value or NA.

Usage

```
fill_measures(sentomeasures, fill = "zero")
```

Arguments

sentomeasures	a sentomeasures object created using sento_measures.
fill	an element of c("zero", "latest", NA); the first and last assume missing dates represent zero sentiment, the second assumes missing dates represent constant sentiment.

Value

A modified sentomeasures object.

Author(s)

Samuel Borms

Examples

```
data("usnews")
data("lexicons")
data("valence")

# construct a sentomeasures object to start with
corpus <- sento_corpus(corpusdf = usnews)
corpusSample <- quanteda::corpus_sample(corpus, sample = 500)
l <- setup_lexicons(lexicons[c("LM_eng", "HENRY_eng")], valence[["valence_eng"]])
ctr <- ctr_agg(howTime = c("equal_weight", "linear"), by = "year", lag = 3)
sentomeasures <- sento_measures(corpusSample, l, ctr)
```

```
# fill measures
f1 <- fill_measures(sentomeasures)
f2 <- fill_measures(sentomeasures, fill = "latest")
f3 <- fill_measures(sentomeasures, fill = NA)
```

`get_hows` *Options supported to perform aggregation into sentiment measures*

Description

Call for information purposes only. Used within `ctr_agg` to check if supplied aggregation hows are supported.

Usage

```
get_hows()
```

Value

A list with the supported aggregation hows for arguments `howWithin` (`"words"`), `howDows` (`"docs"`) and `howTime` (`"time"`), to be supplied to `ctr_agg`.

See Also

`ctr_agg`

`lexicons` *Built-in lexicons*

Description

A list containing all built-in lexicons as a `data.table` with two columns: a `x` column with the words, and a `y` column with the polarities. The list element names incorporate consecutively the name and language, and `"_tr"` as suffix if the lexicon is translated. The lexicons are in the format required for further sentiment analysis. The built-in lexicons are the following:

- FEEL_eng_tr (FEEL: French Expanded Emotion Lexicon)
- FEEL_fr
- FEEL_nl_tr
- GI_eng (GI: General Inquirer, i.e. Harvard IV-4 combined with Laswell)
- GI_fr_tr
- GI_nl_tr
- HENRY_eng (HENRY: Henry)

- HENRY_fr_tr
- HENRY_nl_tr
- LM_eng (LM: Loughran and McDonald)
- LM_fr_tr
- LM_nl_tr

Usage

```
data("lexicons")
```

Format

A list with all built-in lexicons, appropriately named as `"NAME_language(_tr)"` .

Source

FEEL lexicon

GI lexicon

HENRY lexicon

LM lexicon

Examples

```
lexicons[c("FEEL_eng_tr", "LM_eng")]
```

---

merge_measures *Merge sentiment measures*

Description

Merge (further aggregate) measures by combining across provided lexicons, features, and time weighting schemes dimensions. The combination occurs by taking the mean of the relevant measures.

Usage

```
merge_measures(ctr)
```

Arguments

`ctr`	output from a `ctr_merge` call.

Value

A modified `sentomeasures` object, with only the sentiment measures required, including updated information and statistics, but the original sentiment scores `data.table` untouched.

Author(s)

Samuel Borms

See Also

ctr_merge

Examples

```
data("usnews")
data("lexicons")
data("valence")

# construct a sentomeasures object to start with
corpus <- sento_corpus(corpusdf = usnews)
corpusSample <- quanteda::corpus_sample(corpus, size = 500)
l <- setup_lexicons(lexicons[c("LM_eng", "HENRY_eng")], valence[["valence_eng"]])
ctr <- ctr_agg(howTime = c("equal_weight", "linear"), by = "year", lag = 3)
sentomeasures <- sento_measures(corpusSample, l, ctr)

# set up control function and perform the merging
ctrMerge <- ctr_merge(sentomeasures,
                      time = list(W = c("equal_weight", "linear")),
                      feat = list(journals = c("wsj", "wapo")),
                      do.keep = TRUE)
sentomeasuresMerged <- merge_measures(ctrMerge)
```

---

perform_agg *Aggregate textual sentiment across documents and time*

Description

Condense document-level textual sentiment scores into a panel of textual sentiment measures by aggregating across documents and time. This function is called within sento_measures, applied on the output of compute_sentiment.

Usage

```
perform_agg(toAgg, ctr)
```

Arguments

toAgg	output from a compute_sentiment call.
ctr	output from a ctr_agg call. The "howWithin" argument plays no more role.

Value

A sentomeasures object.

Author(s)

Samuel Borms, Keven Bluteau

See Also

compute_sentiment, ctr_agg, sento_measures

Examples

```
data("usnews")
data("lexicons")
data("valence")

# computation of sentiment and aggregation into sentiment measures
corpus <- sento_corpus(corpusdf = usnews)
corpusSample <- quanteda::corpus_sample(corpus, size = 1000)
l <- setup_lexicons(lexicons[c("LM_eng", "HENRY_eng")], valence[["valence_eng"]])
sent <- compute_sentiment(corpusSample, l, how = "counts")
ctr <- ctr_agg(howTime = c("linear"), by = "year", lag = 3)
sentomeasures <- perform_agg(sent, ctr)
```

---

perform_MCS *Apply model confidence set (MCS) procedure to a selection of models*

Description

Calculates the model confidence set (see "The Model Confidence Set"; Hansen, Lunde and Nason, 2011) as implemented in the **MCS** package, for a set of different `sentomodeliter` objects.

Usage

```
perform_MCS(models, loss = c("DA", "errorSq", "AD", "accuracy"), ...)
```

Arguments

`models`	a named `list` of `sentomodeliter` objects. All models should be of the same family, being either `"gaussian"`, `"binomial"` or `"multinomial"`, and have performance data of the same dimensions.
`loss`	a single `character` vector, either `"DA"` (directional *in*accuracy), `"errorSq"` (squared errors), `"AD"` (absolute errors) or `"accuracy"` (*in*accurate class predictions). This argument defines on what basis the model confidence set is calculated. The first three options are available for `"gaussian"` models, the last option applies only to `"binomial"` and `"multinomial"` models.
`...`	other parameters that can be supplied to the `MCSprocedure` function. If empty, its default values are used.

Value

An object as returned by the MCSprocedure function.

Author(s)

Samuel Borms

See Also

sento_model, MCSprocedure

Examples

```
## Not run:
data("usnews")
data("lexicons")
data("valence")
data("epu")

# construct two sentomeasures objects
corpusAll <- sento_corpus(corpusdf = usnews)
corpus <- quanteda::corpus_subset(corpusAll, date >= "1997-01-01" & date < "2014-10-01")
l <- setup_lexicons(lexicons[c("LM_eng", "HENRY_eng")], valence[["valence_eng"]])

ctr1 <- ctr_agg(howWithin = "tf-idf", howDocs = "proportional",
                howTime = c("equal_weight", "linear"), by = "month", lag = 3)
sentMeas1 <- sento_measures(corpus, l, ctr1)

ctr2 <- ctr_agg(howWithin = "counts", howDocs = "equal_weight",
                howTime = c("equal_weight", "linear"), by = "month", lag = 3)
sentMeas2 <- sento_measures(corpus, l, ctr2)

# prepare y and other x variables
y <- epu[epu$date >= sentMeas1$measures$date[1], ]$index
length(y) == nrow(sentMeas1$measures) # TRUE
x <- data.frame(runif(length(y)), rnorm(length(y))) # two other (random) x variables
colnames(x) <- c("x1", "x2")

# estimate different type of regressions
ctr1 <- ctr_model(model = "gaussian", type = "AIC", do.iter = TRUE,
                  h = 0, nSample = 120, start = 50)
out1 <- sento_model(sentMeas1, y, x = x, ctr = ctr1)

ctr2 <- ctr_model(model = "gaussian", type = "AIC", do.iter = TRUE,
                  h = 0, nSample = 120, start = 50)
out2 <- sento_model(sentMeas1, y, x = NULL, ctr = ctr2)

ctr3 <- ctr_model(model = "gaussian", type = "AIC", do.iter = TRUE,
                  h = 0, nSample = 120, start = 50)
out3 <- sento_model(sentMeas2, y, x = x, ctr = ctr3)

ctr4 <- ctr_model(model = "gaussian", type = "AIC", do.iter = TRUE,
```

```
                  h = 0, nSample = 120, start = 50)
out4 <- sento_model(sentMeas2, y, x = NULL, ctr = ctr4)

mcs <- perform_MCS(models = list(m1 = out1, m2 = out2, m3 = out3, m4 = out4),
                   loss = "errorSq")
## End(Not run)
```

`plot.sentomeasures` *Plot sentiment measures*

Description

Straightforward plotting method that shows all sentiment measures from the provided `sentomeasures` object in one plot, or the average along one of the lexicons, features and time weighting dimensions. We suggest to make use of the `select_measures` function when you desire to plot only a subset of the sentiment measures.

Usage

```
## S3 method for class 'sentomeasures'
plot(x, group = "all", ...)
```

Arguments

`x`	a `sentomeasures` object created using `sento_measures`.
`group`	a value from `c("lexicons", "features", "time", "all")`. The first three choices display the average of all measures from the same group, in a different color. The choice `"all"` displays every single sentiment measure in a separate color, but this may look visually overwhelming very fast, and can be quite slow.
`...`	not used.

Value

Returns a simple `ggplot` object, which can be added onto (or to alter its default elements) by using the + operator (see examples). By default, a legend is positioned at the top if there are at maximum twelve line graphs plotted.

Author(s)

Samuel Borms

Examples

```
data("usnews")
data("lexicons")
data("valence")

# construct a sentomeasures object to start with
corpus <- sento_corpus(corpusdf = usnews)
corpusSample <- quanteda::corpus_sample(corpus, size = 500)
l <- setup_lexicons(lexicons[c("LM_eng")], valence[["valence_eng"]])
ctr <- ctr_agg(howTime = c("equal_weight", "linear"), by = "year", lag = 3)
sentomeasures <- sento_measures(corpusSample, l, ctr)

# plot sentiment measures
plot(sentomeasures)
plot(sentomeasures, group = "features")

# adjust appearance of plot
p <- plot(sentomeasures)
p <- p +
  ggthemes::theme_base() +
  scale_x_date(name = "month-year") +
  scale_y_continuous(name = "newName")
p
```

`plot.sentomodeliter` *Plot iterative predictions versus realized values*

Description

Displays a plot of all predictions made through the iterative model computation as incorporated in the input `sentomodeliter` object, as well as the corresponding true values.

Usage

```
## S3 method for class 'sentomodeliter'
plot(x, ...)
```

Arguments

`x`	a `sentomodeliter` object created using `sento_model`.
`...`	not used.

Value

Returns a simple `ggplot` object, which can be added onto (or to alter its default elements) by using the + operator (see examples).

Author(s)

Samuel Borms

Examples

```
data("usnews")
data("lexicons")
data("valence")
data("epu")

# construct a sentomeasures object to start with
corpusAll <- sento_corpus(corpusdf = usnews)
corpus <- quanteda::corpus_subset(corpusAll, date >= "2007-01-01" & date < "2014-10-01")
l <- setup_lexicons(lexicons[c("LM_eng", "HENRY_eng")], valence[["valence_eng"]])
ctr <- ctr_agg(howWithin = "tf-idf", howDocs = "proportional",
               howTime = c("equal_weight", "linear"),
               by = "month", lag = 3)
sentomeasures <- sento_measures(corpus, l, ctr)

# prepare y variable
y <- epu[epu$date >= sentomeasures$measures$date[1], ]$index
length(y) == nrow(sentomeasures$measures) # TRUE

# estimate regression iteratively based on a sample of 60, skipping first 25 iterations
ctr <- ctr_model(model = "gaussian", type = "AIC", do.iter = TRUE,
                 h = 0, nSample = 60, start = 26)
out <- sento_model(sentomeasures, y, ctr = ctr)
summary(out)

# plotting
p <- plot(out)
p <- p +
  ggthemes::theme_few()
p
```

---

plot_attributions *Plot prediction attributions at specified level*

Description

Shows a plot of the attributions along the dimension provided, stacked per date.

Usage

```
plot_attributions(attributions, group = "features")
```

Arguments

attributions	an output from a retrieve_attributions call.
group	a value from c("lexicons", "features", "time").

Details

See sento_model for an elaborate modelling example including the calculation and plotting of attributions. This function does not handle the plotting of the attribution of individual documents, since there are often a lot of documents involved and they appear only once at one date (even though a document may contribute to predictions at several dates, depending on the number of lags in the time aggregation).

Value

Returns a simple ggplot object, which can be added onto (or to alter its default elements) by using the + operator (see examples). By default, a legend is positioned at the top if the number of components of the dimension (thus, individual line graphs) is at maximum twelve.

Author(s)

Samuel Borms, Keven Bluteau

---

predict.sentomodel *Make predictions from a sentomodel object*

Description

Prediction method for sentomodel class, with usage along the lines of predict.glmnet, but simplified in terms of allowed parameters.

Usage

```
## S3 method for class 'sentomodel'
predict(object, newx, type, offset = NULL, ...)
```

Arguments

object	a sentomodel object created with sento_model.
newx	a matrix of numeric values with all explanatory variables to be used for the prediction(s), structured row-by-row; see documentation for predict.glmnet. The number of variables should be equal to sentomodel$nVar, being the sum of the number of original sentiment measures and the number of additional explanatory variables. Variables discarded in the regression process are discarded again here, based on sentomodel$discarded.
type	type of prediction required, a value from c("link", "response", "class"), see documentation for predict.glmnet.
offset	not used. Any values here will be ignored.
...	not used.

Value

A prediction output depending on the type argument.

Author(s)

Samuel Borms

See Also

predict.glmnet, sento_model

---

retrieve_attributions *Retrieve top-down sentiment attributions given model object*

Description

Computes the attributions to predictions for a (given) number of dates at all possible sentiment dimensions, based on the coefficients associated to each sentiment measure, as estimated in the provided model object.

Usage

```
retrieve_attributions(model, sentomeasures, do.normalize = FALSE,
  refDates = NULL, factor = NULL)
```

Arguments

model	a sentomodel or sentomodeliter object created with sento_model.
sentomeasures	the sentomeasures object, as created with sento_measures, used to estimate the model from the first argument.
do.normalize	a logical, TRUE divides each element of every attribution vector at a given date by its L2-norm at that date, normalizing the values between -1 and 1. The document attributions are not normalized.
refDates	the dates at which attribution is to be performed. These should be between the latest date available in the input sentomeasures object and the first estimation sample date, i.e. model$dates[1] if model is a sentomodel object. All dates should also be present in sentomeasures$measures$date. If NULL (default), attribution is calculated for all in-sample dates. Ignored if model is a sentomodeliter object, for which attribution is calculated for all out-of-sample prediction dates.
factor	the factor level as a single character vector for which attribution has to be calculated in case of (a) multinomial model(s). Ignored for linear and binomial models.

Details

See sento_model for an elaborate modelling example including the calculation and plotting of attributions. The attribution for logistic models is represented in terms of log odds. For binomial models, it is calculated with respect to the last factor level or factor column.

Value

A list with all dimensions for which aggregation is computed, being "documents", "lexicons", "features" and "time". The last three dimensions are data.tables having a "date" column and the other columns the different components of the dimension, with the attributions as values. Document-level attribution is further decomposed into a data.table per date, with "id", "date" and "attrib" columns.

Author(s)

Samuel Borms, Keven Bluteau

See Also

sento_model

---

scale.sentomeasures *Scaling and centering of sentiment measures*

Description

Scales and centers the sentiment measures from a sentomeasures object, column-per-column. By default, the measures are normalized. NAs are removed first.

Usage

```
## S3 method for class 'sentomeasures'
scale(x, center = TRUE, scale = TRUE)
```

Arguments

x	a sentomeasures object created using sento_measures.
center	a logical, see documentation for the generic scale.
scale	a logical, see documentation for the generic scale.

Value

A modified sentomeasures object, with the measures replaced by the scaled measures as well as updated statistics.

Author(s)

Samuel Borms

Examples

```
data("usnews")
data("lexicons")
data("valence")

# construct a sentomeasures object to start with
corpus <- sento_corpus(corpusdf = usnews)
corpusSample <- quanteda::corpus_sample(corpus, size = 500)
l <- setup_lexicons(lexicons[c("LM_eng", "HENRY_eng")], valence[["valence_eng"]])
ctr <- ctr_agg(howTime = c("equal_weight", "linear"), by = "year", lag = 3)
sentomeasures <- sento_measures(corpusSample, l, ctr)

# scale sentiment measures
scaled <- scale(sentomeasures)
```

`select_measures` *Select a subset of sentiment measures*

Description

Selects the subset of sentiment measures which include either all of the given selection components combined, or those who's name consist of at least one of the selection components. One can also extract measures within a subset of dates.

Usage

```
select_measures(sentomeasures, toSelect = "all", do.combine = TRUE,
  dates = NA)
```

Arguments

`sentomeasures`	a `sentomeasures` object created using `sento_measures`.
`toSelect`	a `"character"` vector of the lexicon, feature and time weighting scheme names, to indicate which measures need to be selected. By default equal to `"all"`, which means no selection of the sentiment measures is made; this may be used if one only wants to extract a subset of dates via the `dates` argument.
`do.combine`	a `logical` indicating if only measures for which all (`do.combine = TRUE`) or at least one (`do.combine = FALSE`) of the selection components should occur in each sentiment measure's name in the subset. If `do.combine = TRUE`, the `toSelect` argument can only consist of one lexicon, one feature, and one time weighting scheme at maximum.
`dates`	any expression, in the form of a `character` vector, that would correctly evaluate to a `logical` vector, features the variable `date` and has dates specified as `"yyyy-mm-dd"`, e.g. `dates = "date >= '2000-01-15'"`. This argument may also be a vector of class `Date` which extracts all dates that show up in that vector. See the examples. By default equal to `NA`, meaning no subsetting based on dates is done.

Value

A modified `sentomeasures` object, with only the sentiment measures required, including updated information and statistics, but the original sentiment scores `data.table` untouched.

Author(s)

Samuel Borms

Examples

```
data("usnews")
data("lexicons")
data("valence")

# construct a sentomeasures object to start with
corpus <- sento_corpus(corpusdf = usnews)
corpusSample <- quanteda::corpus_sample(corpus, size = 1000)
l <- setup_lexicons(lexicons[c("LM_eng", "HENRY_eng")], valence[["valence_eng"]])
ctr <- ctr_agg(howTime = c("equal_weight", "linear"), by = "year", lag = 3)
sentomeasures <- sento_measures(corpusSample, l, ctr)

# different selections
sel1 <- select_measures(sentomeasures, c("equal_weight"))
sel2 <- select_measures(sentomeasures, c("equal_weight", "linear"), do.combine = FALSE)
sel3 <- select_measures(sentomeasures, c("linear", "LM_eng"))
sel4 <- select_measures(sentomeasures, c("linear", "LM_eng", "wsj", "economy"),
                        do.combine = FALSE)
sel5 <- select_measures(sentomeasures, c("linear", "LM_eng"),
                        dates = "date >= '1996-12-31' & date <= '2000-12-31'")
d <- seq(as.Date("2000-01-01"), as.Date("2013-12-01"), by = "month")
sel6 <- select_measures(sentomeasures, c("linear", "LM_eng"), dates = d)
```

---

sento_corpus *Create a sentocorpus object*

Description

Formalizes a collection of texts into a well-defined corpus object, by calling, amongst others, the `corpus` function from the **quanteda** package. This package is a (very) fast text mining package; for more info, see quanteda. Their formal corpus structure is required for better memory management, corpus manipulation, and sentiment calculation. This function mainly performs a set of checks on the input data and prepares the corpus for further sentiment analysis.

Usage

```
sento_corpus(corpusdf, do.clean = FALSE)
```

Arguments

corpusdf	a data.frame with as named columns and *in this order*: a document "id" column, a "date" column, a "text" column (i.e. the columns where all texts to analyze reside), and a series of feature columns of type numeric, with values pointing to the applicability of a particular feature to a particular text. The latter columns are often binary (1 means the feature is applicable to the document in the same row) or as a percentage to specify the degree of connectedness of a feature to a document. Features could be topics (e.g., legal, political, or economic), but also article sources (e.g., online or printed press), amongst many more options. If you have no knowledge about features or no particular features are of interest to your analysis, provide no feature columns. In that case, the corpus constructor automatically adds an additional feature column named "dummy". Provide the date column as "yyyy-mm-dd". The id column should be in character mode. All spaces in the names of the features are replaced by underscores.
do.clean	a logical, if TRUE all texts undergo a cleaning routine to eliminate common textual garbage. This includes a brute force replacement of HTML tags and non-alphanumeric characters by an empty string.

Details

A sentocorpus object can be regarded as a specialized instance of a **quanteda** corpus. In theory, all **quanteda** functions applicable to its corpus object can also be applied to a sentocorpus object. However, changing a given sentocorpus object too drastically using some of **quanteda**'s functions might alter the very structure the corpus is meant to have (as defined in the corpusdf argument) to be able to be used as an input in other functions of the **sentometrics** package. There are functions, including corpus_sample or corpus_subset, that do not change the actual corpus structure and may come in handy. To add additional features, we recommend to use add_features.

Value

A sentocorpus object, derived from a **quanteda** corpus classed list keeping the elements "documents", "metadata", and "settings". The first element incorporates the corpus represented as a data.frame.

Author(s)

Samuel Borms

See Also

corpus

Examples

```
data("usnews")

# corpus construction
corpus <- sento_corpus(corpusdf = usnews)
```

```
# take a random subset making use of quanteda
corpusSmall <- quanteda::corpus_sample(corpus, size = 500)

# deleting a feature
quanteda::docvars(corpus, field = "wapo") <- NULL

# corpus creation when no features are present
corpusDummy <- sento_corpus(corpusdf = usnews[, 1:3])
```

`sento_measures` *One-way road towards a sentomeasures object*

Description

Wrapper function which assembles calls to `compute_sentiment` and `perform_agg`, and includes the input `sentocorpus` and computed sentiment scores in its output. Serves as the most direct way towards a panel of textual sentiment measures as a `sentomeasures` object.

Usage

```
sento_measures(sentocorpus, lexicons, ctr)
```

Arguments

`sentocorpus`	a `sentocorpus` object created with `sento_corpus`.
`lexicons`	output from a `setup_lexicons` call.
`ctr`	output from a `ctr_agg` call.

Value

A `sentomeasures` object, which is a `list` containing:

`measures`	a `data.table` with a `"date"` column and all textual sentiment measures as remaining columns.
`features`	a `character` vector of the different features.
`lexicons`	a `character` vector of the different lexicons used.
`time`	a `character` vector of the different time weighting schemes used.
`by`	a single `character` vector specifying the time interval of aggregation used.
`stats`	a `data.frame` with a series of elementary statistics (mean, standard deviation, maximum, minimum, and average correlation with all other measures) for each individual sentiment measure.
`sentiment`	the sentiment scores `data.table` with `"date"` and lexicon–feature sentiment scores columns. If `ctr$do.ignoreZeros = TRUE`, all zeros are replaces by `NA`.
`howWithin`	a single `character` vector to remind how sentiment within documents was aggregated.

howDocs	a single character vector to remind how sentiment across documents was aggregated.
fill	a single character vector that specifies if and how missing dates have been added before aggregation across time was carried out.
do.ignoreZeros	a single character vector to remind if documents with zero sentiment have been ignored in the within-document aggregation.
attribWeights	a list of document and time weights used in the retrieve_attributions function. Serves further no direct purpose.

Author(s)

Samuel Borms, Keven Bluteau

See Also

compute_sentiment, perform_agg

Examples

```
data("usnews")
data("lexicons")
data("valence")

# construct a sentomeasures object to start with
corpus <- sento_corpus(corpusdf = usnews)
corpusSample <- quanteda::corpus_sample(corpus, size = 750)
l <- setup_lexicons(lexicons[c("LM_eng", "HENRY_eng")], valence[["valence_eng"]])
ctr <- ctr_agg(howWithin = "tf-idf",
               howDocs = "proportional",
               howTime = c("equal_weight", "linear", "almon"),
               by = "month",
               lag = 3,
               ordersAlm = 1:3,
               do.inverseAlm = TRUE,
               do.normalizeAlm = TRUE)
sentomeasures <- sento_measures(corpusSample, l, ctr)
summary(sentomeasures)
```

---

sento_model *Optimized and automated sparse regression*

Description

Linear or nonlinear penalized regression of any dependent variable on the wide number of sentiment measures and potentially other explanatory variables. Either performs a regression given the provided variables at once, or computes regressions sequentially for a given sample size over a longer time horizon, with associated one-step ahead prediction performance metrics.

Usage

```
sento_model(sentomeasures, y, x = NULL, ctr)
```

Arguments

sentomeasures	a sentomeasures object created using sento_measures. There should be at least two explanatory variables including the ones provided through the x argument.
y	a one-column data.frame or a numeric vector capturing the dependent (response) variable. In case of a logistic regression, the response variable is either a factor or a matrix with the factors represented by the columns as binary indicators, with the second factor level or column as the reference class in case of a binomial regression. No NA values are allowed.
x	a named data.frame with other explanatory variables as numeric, by default set to NULL.
ctr	output from a ctr_model call.

Details

Models are computed using the elastic net regularization as implemented in the **glmnet** package, to account for the multidimensionality of the sentiment measures. Additional explanatory variables are not subject to shrinkage. Independent variables are normalized in the regression process, but coefficients are returned in their original space. For a helpful introduction to **glmnet**, we refer to the vignette. The optimal elastic net parameters lambda and alpha are calibrated either through a to specify information criterion or through cross-validation (based on the "rolling forecasting origin" principle, using the train function). In the latter case, the training metric is automatically set to "RMSE" for a linear model and to "Accuracy" for a logistic model. We suppress many of the details that can be supplied to the glmnet and train functions we rely on, for the sake of user-friendliness.

Value

If ctr$do.iter = FALSE, a sentomodel object which is a list containing:

reg	optimized regression, i.e. a model-specific glmnet object.
model	the input argument ctr$model, to indicate the type of model estimated.
x	a matrix of the values used in the regression for all explanatory variables.
alpha	calibrated alpha.
lambda	calibrated lambda.
trained	output from train call (if ctr$type = "cv").
ic	a list composed of two elements: the information criterion used in the calibration under "criterion", and a vector of all minimum information criterion values for each value in alphas under "opts" (if ctr$type != "cv").
dates	sample reference dates as a two-element character vector, being the earliest and most recent date from the sentomeasures object accounted for in the estimation window.

nVar	the sum of the number of sentiment measures and other explanatory variables inputted.
discarded	a named logical vector of length equal to the number of sentiment measures, in which TRUE indicates that the particular sentiment measure has not been considered in the regression process. A sentiment measure is not considered when it is a duplicate of another, or when at least 25% of the observations are equal to zero.

If ctr$do.iter = TRUE, a sentomodeliter object which is a list containing:

models	all sparse regressions, i.e. separate sentomodel objects as above, as a list with as names the dates from the perspective of the sentiment measures at which predictions for performance measurement are carried out (i.e. one date step beyond the date sentomodel$dates[2]).
alphas	calibrated alphas.
lambdas	calibrated lambdas.
performance	a data.frame with performance-related measures, being "RMSFE" (root mean squared forecasting error), "MAD" (mean absolute deviation), "MDA" (mean directional accuracy, in which's calculation zero is considered as a positive; in percentage points), "accuracy" (proportion of correctly predicted classes in case of a logistic regression; in percentage points), and each's respective individual values in the sample. Directional accuracy is measured by comparing the change in the realized response with the change in the prediction between two consecutive time points (omitting the very first prediction, resulting in NA). Only the relevant performance statistics are given depending on the type of regression. Dates are as in the "models" output element, i.e. from the perspective of the sentiment measures.

Author(s)

Samuel Borms, Keven Bluteau

See Also

ctr_model, glmnet, train, retrieve_attributions

Examples

```
data("usnews")
data("lexicons")
data("valence")
data("epu")

# construct a sentomeasures object to start with
corpusAll <- sento_corpus(corpusdf = usnews)
corpus <- quanteda::corpus_subset(corpusAll, date >= "2004-01-01" & date < "2014-10-01")
l <- setup_lexicons(lexicons[c("LM_eng", "HENRY_eng")], valence[["valence_eng"]])
ctr <- ctr_agg(howWithin = "tf-idf", howDocs = "proportional",
               howTime = c("equal_weight", "almon"),
               by = "month", lag = 3, ordersAlm = 1:2,
```

```
                do.inverseAlm = TRUE, do.normalizeAlm = TRUE)
sentomeasures <- sento_measures(corpus, l, ctr)

# prepare y and other x variables
y <- epu[epu$date >= sentomeasures$measures$date[1], ]$index
length(y) == nrow(sentomeasures$measures) # TRUE
x <- data.frame(runif(length(y)), rnorm(length(y))) # two other (random) x variables
colnames(x) <- c("x1", "x2")

# a linear model based on the Akaike information criterion
ctrIC <- ctr_model(model = "gaussian", type = "AIC", do.iter = FALSE, h = 0)
out1 <- sento_model(sentomeasures, y, x = x, ctr = ctrIC)

# some post-analysis (attribution)
attributions1 <- retrieve_attributions(out1, sentomeasures,
                                       refDates = sentomeasures$measures$date[20:40])

## Not run:
# a cross-validation based model
cl <- makeCluster(detectCores() - 2)
registerDoParallel(cl)
ctrCV <- ctr_model(model = "gaussian", type = "cv", do.iter = FALSE,
                   h = 0, alphas = c(0.10, 0.50, 0.90), trainWindow = 70,
                   testWindow = 10, oos = 0, do.progress = TRUE)
out2 <- sento_model(sentomeasures, y, x = x, ctr = ctrCV)
stopCluster(cl)
summary(out2)
## End(Not run)

## Not run:
# a cross-validation based model but for a binomial target
yb <- epu[epu$date >= sentomeasures$measures$date[1], ]$above
ctrCVb <- ctr_model(model = "binomial", type = "cv", do.iter = FALSE,
                    h = 0, alphas = c(0.10, 0.50, 0.90), trainWindow = 70,
                    testWindow = 10, oos = 0, do.progress = TRUE)
out3 <- sento_model(sentomeasures, yb, x = x, ctr = ctrCVb)
summary(out3)
## End(Not run)

# an example of an iterative analysis
ctrIter <- ctr_model(model = "gaussian", type = "BIC", do.iter = TRUE,
                     alphas = c(0.25, 0.75), h = 0, nSample = 100, start = 21)
out4 <- sento_model(sentomeasures, y, x = x, ctr = ctrIter)
summary(out4)

## Not run:
# a similar iterative analysis, parallelized
cl <- makeCluster(detectCores() - 2)
registerDoParallel(cl)
ctrIter <- ctr_model(model = "gaussian", type = "Cp", do.iter = TRUE,
                     h = 0, nSample = 100, do.parallel = TRUE)
out5 <- sento_model(sentomeasures, y, x = x, ctr = ctrIter)
stopCluster(cl)
```

```
summary(out5)
## End(Not run)

# some more post-analysis (attribution and prediction)
attributions2 <- retrieve_attributions(out4, sentomeasures)
plot_attributions(attributions2, "features")

nx <- ncol(sentomeasures$measures) - 1 + ncol(x) # don't count date column
newx <- runif(nx) * cbind(sentomeasures$measures[, -1], x)[30:50, ]
preds <- predict(out1, newx = as.matrix(newx), type = "link")
```

`setup_lexicons` *Set up lexicons (and valence word list) for use in sentiment analysis*

Description

Structures provided lexicons and potentially integrates valence words. One can also provide (part of) the built-in lexicons from `data("lexicons")` or a valence word list from `data("valence")` as an argument. Makes use of `as_key` from the **sentimentr** package to make the output coherent and check for duplicates.

Usage

```
setup_lexicons(lexiconsIn, valenceIn = NULL, do.split = FALSE)
```

Arguments

`lexiconsIn`	a list of (raw) lexicons, each element being a `data.table` or a `data.frame` with respectively a words column and a polarity score column. The lexicons should be appropriately named for clarity in terms of subsequently obtained sentiment measures. Alternatively, a subset of the already formatted built-in lexicons accessible via `lexicons` can be declared too, as part of the same list input. If only (some of) the package built-in lexicons want to be used (with *no* valence shifters), one can simply supply `lexicons[c(...)]` as an argument to either `sento_measures` or `compute_sentiment`. However, it is strongly recommended to pass all lexicons (and a valence word list) to this function first, in any case.
`valenceIn`	a single valence word list as a `data.table` or a `data.frame` with respectively a words column, a type column (`1` for negators, `2` for amplifiers/intensifiers, and `3` for deamplifiers/downtoners) and a score column. Suggested scores are -1, 2, and 0.5 respectively, and should be the same within each type. This argument can also be one of the already formatted built-in valence word lists accessible via `valence`. If `NULL`, no valence word list is part of this function's output, nor will it applied in the sentiment analysis.
`do.split`	a `logical` that if `TRUE` splits every lexicon into a separate positive polarity and negative polarity lexicon.

Value

A `list` with each lexicon as a separate element according to its name, as a `data.table`, and optionally an element named `valence` that comprises the valence words. Every `x` column contains the words, every `y` column contains the polarity score, and for the valence word list, `t` contains the word type. If a valence word list is provided, all lexicons are expanded by copying the respective lexicon, and changing the words and scores according to the valence word type: "NOT_" is added for negators, "VERY_" is added for amplifiers and "HARDLY_" is added for deamplifiers. Lexicon scores are multiplied by -1, 2 and 0.5 by default, respectively, or the first value of the scores column of the valence word list.

Author(s)

Samuel Borms

See Also

`as_key`

Examples

```
data("lexicons")
data("valence")

# sets up output list straight from built-in word lists including valence words
l1 <- c(lexicons[c("LM_eng", "HENRY_eng")], valence[["eng"]])

# including a self-made lexicon, with and without valence shifters
lexIn <- c(list(myLexicon = data.table(w = c("nice", "boring"), s = c(2, -1))),
           lexicons[c("GI_eng")])
valIn <- valence[["valence_eng"]]
l2 <- setup_lexicons(lexIn)
l3 <- setup_lexicons(lexIn, valIn)
l4 <- setup_lexicons(lexIn, valIn, do.split = TRUE)
```

---

to_global *Merge sentiment measures into one global sentiment measure*

Description

Merges all sentiment measures into one global textual sentiment measure based on a set of weights to indicate the importance of each component in the `lexicons`, `features`, and `time` vectors as specified in the input `sentomeasures` object. Every measure receives a weight in the global measure equal to the multiplication of the supplied weights of the components it is contained of. The global sentiment measure then corresponds to a weighted average of these weights times the sentiment scores, per date.

Usage

```
to_global(sentomeasures, lexicons = 1, features = 1, time = 1)
```

Arguments

sentomeasures	a sentomeasures object created using sento_measures.
lexicons	a numeric vector of weights, of size length(sentomeasures$lexicons), in the same order and summing to one. By default set to 1, which means equally weighted.
features	a numeric vector of weights, of size length(sentomeasures$features), in the same order and summing to one. By default set to 1, which means equally weighted.
time	a numeric vector of weights, of size length(sentomeasures$time), in the same order and summing to one. By default set to 1, which means equally weighted.

Details

This function returns no sentomeasures object, however the global sentiment measure as outputted can be added to regressions as an additional variable using the x argument in the sento_model function.

Value

A data.frame with the values for the global sentiment measure under the global column and dates as row names.

Author(s)

Samuel Borms

See Also

sento_model

Examples

```
data("usnews")
data("lexicons")
data("valence")

# construct a sentomeasures object to start with
corpus <- sento_corpus(corpusdf = usnews)
corpusSample <- quanteda::corpus_sample(corpus, size = 1250)
l <- setup_lexicons(lexicons[c("LM_eng", "HENRY_eng")], valence[["valence_eng"]])
ctr <- ctr_agg(howTime = c("equal_weight", "linear"), by = "year", lag = 3)
sentomeasures <- sento_measures(corpusSample, l, ctr)

# merge into one global sentiment measure, with specified weighting for lexicons and features
```

```
global <- to_global(sentomeasures, lexicons = c(0.40, 0.60),
                                   features = c(0.10, 0.20, 0.30, 0.40),
                                   time = 1)
```

---

usnews *Texts (not) relevant to the U.S. economy*

Description

A collection of texts annotated by humans in terms of relevance to the U.S. economy or not. The texts come from two major journals in the U.S. (The Wall Street Journal and The Washington Post) and cover 4145 documents between 1995 and 2014. It contains following information:

- id. A `character` ID identifier.
- date. Date as `"yyyy-mm-dd"`.
- text. Texts in `character` format.
- wsj. Equals 1 if the article comes from The Wall Street Journal.
- wapo. Equals 1 if the article comes from The Washington Post.
- economy. Equals 1 if the article is relevant to the U.S. economy.
- noneconomy. Equals 1 if the article is not relevant to the U.S. economy.

Usage

```
data("usnews")
```

Format

A `data.frame`, formatted as required to be an input for `sento_corpus`.

Source

Economic News Article Tone and Relevance

Examples

```
data("usnews")
usnews[3192, "text"]
usnews[1:5, c("id", "date", "text")]
```

`valence` *Built-in valence word lists*

Description

A list containing all built-in valence word lists, a `data.table` with three columns: a `x` column with the words, a `t` column with the type of valence words, and a `y` column with the values associated to each word and type of valence shifter. The list element names incorporate the language of the valence word list. All non-English word lists are translated. The valence word lists are in the form required for further sentiment analysis. The built-in valence word lists are the following:

- valence_eng
- valence_fr
- valence_nl

Usage

```
data("valence")
```

Format

A list with all built-in valence word lists, appropriately named.

Source

`hash_valence_shifters` (negators)

Index