Package 'LSX'

February 26, 2022

Type Package

Title Semisupervised Document Scaling by Word-Embedding Models

Date 2022-02-26

Version 1.1.1

Description A word embeddings-based semisupervised model for document scaling Watanabe (2020) <doi:10.1080/19312458.2020.1832976>.
LSS allows users to analyze large and complex corpora on arbitrary dimensions with seed words exploiting efficiency of word embeddings (SVD, Glove).
It can generate word vectors on a users-provided corpus or incorporate a pre-trained word vectors.

License GPL-3

LazyData TRUE

Encoding UTF-8

Depends methods, R (>= 3.5.0)

Imports quanteda (>= 2.0), quanteda.textstats, stringi, digest, Matrix, RSpectra, irlba, rsvd, rsparse, proxyC, stats, ggplot2, ggrepel, reshape2, locfit

Suggests testthat

RoxygenNote 7.1.2

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

NeedsCompilation no

Author Kohei Watanabe [aut, cre, cph]

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

Repository CRAN

Date/Publication 2022-02-26 07:30:07 UTC

R topics documented:

`as.seedwords` *Convenient function to convert a list to seed words*

Description

Convenient function to convert a list to seed words

Usage

```
as.seedwords(x, upper = 1, lower = 2, concatenator = "_")
```

Arguments

`x`	a list of characters vectors or a dictionary object.
`upper`	numeric index or key for seed words for higher scores.
`lower`	numeric index or key for seed words for lower scores.
`concatenator`	character to replace separators of multi-word seed words.

Value

named numeric vector for seed words with polarity scores

`coef.textmodel_lss` *Extract model coefficients from a fitted textmodel_lss object*

Description

`coef()` extract model coefficients from a fitted `textmodel_lss` object. `coefficients()` is an alias.

Usage

```
## S3 method for class 'textmodel_lss'
coef(object, ...)

coefficients.textmodel_lss(object, ...)
```

Arguments

`object`	a fitted textmodel_lss object.
`...`	not used.

`data_dictionary_ideology` *Seed words for analysis of left-right political ideology*

Description

Seed words for analysis of left-right political ideology

Examples

```
as.seedwords(data_dictionary_ideology)
```

`data_dictionary_sentiment` *Seed words for analysis of positive-negative sentiment*

Description

Seed words for analysis of positive-negative sentiment

References

Turney, P. D., & Littman, M. L. (2003). Measuring Praise and Criticism: Inference of Semantic Orientation from Association. ACM Trans. Inf. Syst., 21(4), 315–346. doi: 10.1145/944012.944013

Examples

```
as.seedwords(data_dictionary_sentiment)
```

data_textmodel_lss_russianprotests *A fitted LSS model on street protest in Russia*

Description

This model was trained on a Russian media corpus (newspapers, TV transcripts and newswires) to analyze framing of street protests. The scale is protests as "freedom of expression" (high) vs "social disorder" (low). Although some slots are missing in this object (because the model was imported from the original Python implementation), it allows you to scale texts using `predict`.

References

Lankina, Tomila, and Kohei Watanabe. "'Russian Spring' or 'Spring Betrayal'? The Media as a Mirror of Putin's Evolving Strategy in Ukraine." Europe-Asia Studies 69, no. 10 (2017): 1526–56. doi: 10.1080/09668136.2017.1397603.

predict.textmodel_lss *Prediction method for textmodel_lss*

Description

Prediction method for textmodel_lss

Usage

```
## S3 method for class 'textmodel_lss'
predict(
  object,
  newdata = NULL,
  se_fit = FALSE,
  density = FALSE,
  rescaling = TRUE,
  min_n = 0L,
  ...
)
```

Arguments

`object`	a fitted LSS textmodel.
`newdata`	a dfm on which prediction should be made.
`se_fit`	if `TRUE`, returns standard error of document scores.
`density`	if `TRUE`, returns frequency of polarity words in documents.
`rescaling`	if `TRUE`, normalizes polarity scores using `scale()`.
`min_n`	set the minimum number of polarity words in documents.
`...`	not used

Details

Polarity scores of documents are the means of polarity scores of words weighted by their frequency. When `se_fit = TRUE`, this function returns the weighted means, their standard errors, and the number of polarity words in the documents. When `rescaling = TRUE`, it converts the raw polarity scores to z sores for easier interpretation.

Documents tend to receive extreme polarity scores when they have only few polarity word. This is problematic when LSS is applied to short documents (e.g. social media posts) or individual sentences, but we can alleviate this problem by adding zero polarity words to short documents using `min_n`. This setting does not affect empty documents.

`seedwords` *Seed words for Latent Semantic Analysis*

Description

Seed words for Latent Semantic Analysis

Usage

```
seedwords(type)
```

Arguments

`type`	type of seed words currently only for sentiment (`sentiment`) or political ideology (`ideology`).

References

Turney, P. D., & Littman, M. L. (2003). Measuring Praise and Criticism: Inference of Semantic Orientation from Association. ACM Trans. Inf. Syst., 21(4), 315–346. doi: 10.1145/944012.944013

Examples

```
seedwords('sentiment')
```

smooth_lss *Smooth predicted LSS scores by local polynomial regression*

Description

Smooth predicted LSS scores by local polynomial regression

Usage

```
smooth_lss(
  x,
  lss_var = "fit",
  date_var = "date",
  span = 0.1,
  from = NULL,
  to = NULL,
  engine = c("loess", "locfit"),
  ...
)
```

Arguments

x	a data.frame containing LSS scores and dates.
lss_var	the name of the column for LSS scores.
date_var	the name of the columns for dates.
span	determines the level of smoothing.
from	start of the time period.
to	end of the time period.
engine	specifies the function to smooth LSS scores: loess() or locfit(). The latter should be used when n > 10000.
...	extra arguments passed to loess() or lp()

textmodel_lss *A word embeddings-based semisupervised model for document scaling*

Description

A word embeddings-based semisupervised model for document scaling

Usage

```
textmodel_lss(x, ...)

## S3 method for class 'dfm'
textmodel_lss(
  x,
  seeds,
  terms = NULL,
  k = 300,
  slice = NULL,
  weight = "count",
  cache = FALSE,
  simil_method = "cosine",
  engine = c("RSpectra", "irlba", "rsvd"),
  auto_weight = FALSE,
  include_data = FALSE,
  verbose = FALSE,
  ...
)

## S3 method for class 'fcm'
textmodel_lss(
  x,
  seeds,
  terms = NULL,
  w = 50,
  max_count = 10,
  weight = "count",
  cache = FALSE,
  simil_method = "cosine",
  engine = c("rsparse"),
  auto_weight = FALSE,
  verbose = FALSE,
  ...
)
```

Arguments

x	a dfm or fcm created by quanteda::dfm() or quanteda::fcm()
...	additional arguments passed to the underlying engine.
seeds	a character vector or named numeric vector that contains seed words. If seed words contain "*", they are interpreted as glob patterns. See quanteda::valuetype.
terms	a character vector or named numeric vector that specify words for which polarity scores will be computed; if a numeric vector, words' polarity scores will be weighted accordingly; if NULL, all the features of quanteda::dfm() or quanteda::fcm() will be used.

k	the number of singular values requested to the SVD engine. Only used when x is a dfm.
slice	a number or indices of the components of word vectors used to compute similarity; slice < k to further truncate word vectors; useful for diagnosys and simulation.
weight	weighting scheme passed to quanteda::dfm_weight(). Ignored when engine is "rsparse".
cache	if TRUE, save result of SVD for next execution with identical x and settings. Use the base::options(lss_cache_dir) to change the location cache files to be save.
simil_method	specifies method to compute similarity between features. The value is passed to quanteda.textstats::textstat_simil(), "cosine" is used otherwise.
engine	select the engine to factorize x to generate word vectors. Choose from RSpectra::svds(), irlba::irlba(), rsvd::rsvd(), and rsparse::GloVe().
auto_weight	automatically determine weights to approximate the polarity of terms to seed words. See details.
include_data	if TRUE, fitted model include the dfm supplied as x.
verbose	show messages if TRUE.
w	the size of word vectors. Used only when x is a fcm.
max_count	passed to x_max in rsparse::GloVe$new() where cooccurrence counts are ceiled to this threshold. It should be changed according to the size of the corpus. Used only when x is a fcm.

Details

Latent Semantic Scaling (LSS) is a semisupervised document scaling method. textmodel_lss() constructs word vectors from use-provided documents (x) and weights words (terms) based on their semantic proximity to seed words (seeds). Seed words are any known polarity words (e.g. sentiment words) that users should manually choose. The required number of seed words are usually 5 to 10 for each end of the scale.

If seeds is a named numeric vector with positive and negative values, a bipolar LSS model is construct; if seeds is a character vector, a unipolar LSS model. Usually bipolar models perform better in document scaling because both ends of the scale are defined by the user.

A seed word's polarity score computed by textmodel_lss() tends to diverge from its original score given by the user because it's score is affected not only by its original score but also by the original scores of all other seed words. If auto_weight = TRUE, the original scores are weighted automatically using stats::optim() to minimize the squared difference between seed words' computed and original scores. Weighted scores are saved in seed_weighted in the object.

References

Watanabe, Kohei. 2020. "Latent Semantic Scaling: A Semisupervised Text Analysis Technique for New Domains and Languages", Communication Methods and Measures. doi: 10.1080/19312458.2020.1832976.

Watanabe, Kohei. 2017. "Measuring News Bias: Russia's Official News Agency ITAR-TASS' Coverage of the Ukraine Crisis" European Journal of Communication. doi: 10.1177/0267323117695735.

Examples

```
library("quanteda")

# download corpus
tryCatch({
   con <- url("https://bit.ly/2GZwLcN", "rb")
   corp <- readRDS(con)
   close(con)
   },
   error = function(e) e,
   warning = function(w) w
)

if (exists("corp")) {
toks <- corpus_reshape(corp, "sentences") %>%
        tokens(remove_punct = TRUE) %>%
        tokens_remove(stopwords("en")) %>%
        tokens_select("^[\\p{L}]+$", valuetype = "regex", padding = TRUE)
dfmt <- dfm(toks) %>%
        dfm_trim(min_termfreq = 10)

seed <- as.seedwords(data_dictionary_sentiment)

# SVD
lss_svd <- textmodel_lss(dfmt, seed, include_data = TRUE)
head(coef(lss_svd), 20)
head(predict(lss_svd))
head(predict(lss_svd, min_n = 10)) # more robust

dfmt_grp <- dfm_group(dfmt) # group sentences

# sentiment model on economy
eco <- head(textstat_context(toks, 'econom*'), 500)
lss_svd_eco <- textmodel_lss(dfmt, seed, terms = eco)
head(predict(lss_svd_eco, newdata = dfmt_grp))

# sentiment model on politics
pol <- head(textstat_context(toks, 'politi*'), 500)
lss_svd_pol <- textmodel_lss(dfmt, seed, terms = pol)
head(predict(lss_svd_pol, newdata = dfmt_grp))

# modify hyper-parameters of existing model
lss_svd_pol2 <- as.textmodel_lss(lss_svd_pol, seed[c(1, 8)], terms = pol, slice = 200)
head(predict(lss_svd_pol2, newdata = dfmt_grp))

# GloVe
fcmt <- fcm(toks, context = "window", count = "weighted", weights = 1 / 1:5, tri = TRUE)
lss_glov <- textmodel_lss(fcmt, seed)
head(predict(lss_glov, newdata = dfmt_grp))
}
```

textplot_simil *Plot similarity between seed words*

Description

Plot similarity between seed words

Usage

```
textplot_simil(x)
```

Arguments

x	fitted textmodel_lss object.

textplot_terms *Plot polarity scores of words*

Description

Plot polarity scores of words

Usage

```
textplot_terms(x, highlighted = NULL, max_words = 10000)
```

Arguments

x	a fitted textmodel_lss object.
highlighted	quanteda::pattern to select words to highlight.
max_words	the maximum number of words to plot. Words are randomly sampled to keep the number below the limit.

Index