Package 'textfeatures'

November 29, 2018

Type Package

Title Extracts Features from Text

Version 0.3.0

Description A tool for extracting some generic features (e.g., number of words, line breaks, characters per word, URLs, lower case, upper case, commas, periods, exclamation points, etc.) from strings of text.

License MIT + file LICENSE

URL https://github.com/mkearney/textfeatures

BugReports https://github.com/mkearney/textfeatures/issues

Depends R (>= 3.1.0)

Imports dplyr, purrr, rlang, syuzhet, text2vec, tfse, tibble, tokenizers, utils, stats

Suggests knitr, roxygen2, testthat, covr

Encoding UTF-8

LazyData yes

RoxygenNote 6.1.1

NeedsCompilation no

Author Michael Wayne Kearney [aut, cre] (<https://orcid.org/0000-0002-0730-4694>)

Maintainer Michael Wayne Kearney <kearneymw@missouri.edu>

Repository CRAN

Date/Publication 2018-11-29 15:50:03 UTC

R topics documented:

min_var *Select columns with minimum amount of variance*

Description

Filters numeric columns by requiring a minimum amount of variance

Usage

```
min_var(x, min = 1)
```

Arguments

x	Input data, which should be either a data frame or matrix.
min	Minimum amount of variance to require per column.

Details

This function omits missing values.

Value

Returns data frame (or matrix, depending on input class) with all non-numeric columns and only those numeric columns that meet the minimum amount of variance.

scale_count *Apply various transformations to numeric (and non-id) data*

Description

scale_count: Transforms integer and integerlike columns using log

scale_log: Transforms numeric columns using log

scale_normal: Transforms numeric columns using mean centering and dividing by standard deviation

scale_standard: Transforms numeric columns onto 0-1 scales with 0 and 1 set empirically

scale_sqrt: Transforms numeric columns using sqrt

Usage

```
scale_count(x)

scale_log(x)

scale_normal(x)

scale_standard(x)

scale_sqrt(x)
```

Arguments

x	Input data frame containing numeric columns.

Details

Scale transformations are applied only to numeric (or in the case of `scale_count` only integer or integerish) columns that are not named `"id"` or `"(\.|_)?id"`.

Value

A data frame with the same dimensions but with the numeric/relevant variables transformed.

---

textfeatures *textfeatures*

Description

Extracts features from text vector.

Usage

```
textfeatures(x, sentiment = TRUE, word_dims = NULL, threads = 1,
  normalize = TRUE, export = FALSE)
```

Arguments

x	Input data. Should be character vector or data frame with character variable of interest named "text". If a data frame then the first "idl*_id" variable, if found, is assumed to be an ID variable.
sentiment	Logical, indicating whether to return sentiment analysis features, the variables `sent_afinn` and `sent_bing`. Defaults to FALSE. Setting this to true will speed things up a bit.

word_dims	Integer indicating the desired number of word2vec dimension estimates. When NULL, the default, this function will pick a reasonable number of dimensions (ranging from 2 to 200) based on size of input. To disable word2vec estimates, set this to 0 or FALSE.
threads	Integer, specifying the number of threads to use when generating word2vec estimates. Defaults to 1. Ignored if `word_dims = 0`.
normalize	Logical indicating whether to normalize (mean center, sd = 1) features. Defaults to TRUE.
export	Logical indicating whether to store sufficient information for exporting the feature extraction process (stores the means, standard deviations, and the word2vec reference object, which can then be used to process new data).

Value

A tibble data frame with extracted features as columns.

Examples

```
## the text of five of Trump's most retweeted tweets
trump_tweets <- c(
  "#FraudNewsCNN #FNN https://t.co/WYUnHjjUjg",
  "TODAY WE MAKE AMERICA GREAT AGAIN!",
  paste("Why would Kim Jong-un insult me by calling me \"old,\" when I would",
    "NEVER call him \"short and fat?\" Oh well, I try so hard to be his",
    "friend - and maybe someday that will happen!"),
  paste("Such a beautiful and important evening! The forgotten man and woman",
    "will never be forgotten again. We will all come together as never before"),
  paste("North Korean Leader Kim Jong Un just stated that the \"Nuclear",
    "Button is on his desk at all times.\" Will someone from his depleted and",
    "food starved regime please inform him that I too have a Nuclear Button,",
    "but it is a much bigger &amp;amp; more powerful one than his, and my Button",
    "works!")
)

## get the text features of a character vector
textfeatures(trump_tweets)

## data frame with a character vector named "text"
df <- data.frame(
  id = c(1, 2, 3),
  text = c("this is A!\t sEntence https://github.com about #rstats @github",
    "and another sentence here",
    "The following list:\n- one\n- two\n- three\nOkay!?!"),
  stringsAsFactors = FALSE
)

## get text features of a data frame with "text" variable
textfeatures(df)
```

Index