Package 'StandardizeText'

February 19, 2015

Type Package

Title Standardize Text

Version 1.0

Date 2013-3-4

Author David Nepomechie

Maintainer ``Nepomechie, David Israel'' <d.nepomechie@umiami.edu>

Description Standardizes text according to a template; particularly useful for country names.

License GPL-3

NeedsCompilation no

Repository CRAN

Date/Publication 2013-03-04 17:08:54

R topics documented:

`StandardizeText-package`
Standardize Text

Description

Standardizes text according to a template; particularly useful for country names.

Details

Package:	StandardizeText
Type:	Package
Version:	1.0
Date:	2013-3-4
License:	GPL-3

This package contains two main functions: standardize.text() standardizes generic text, and standardize.countrynames() is optimized for standardizing country names.

Author(s)

David Nepomechie; Maintainer: "Nepomechie, David Israel" <d.nepomechie@umiami.edu>

Examples

```
library(StandardizeText)
sample.text <- c("blue car","STOP","email","tree")
sample.std <- c("the tree","car","e-mail","stop")
sample.df <- data.frame(foo=2:5,bar=sample.text, baz=7:4, qux=sample.std)
out.a <- standardize.text(sample.text,standard=sample.std,suggest="auto")
out.b <- standardize.text(sample.df,2,sample.df,"qux",suggest="auto")
```

---

country.names *Country Names*

Description

A list containing the country names used by various organizations, including the International Monetary Fund (IMF), International Standards Organization (ISO), Penn World Tables (PWT), World Bank (WB), and World Health Organization (WHO). The default names are mostly derived from the ISO name set.

Usage

```
data(country.names)
```

Format

The format is: List of 6 $ default: chr [1:260] "Afghanistan" "Aland Islands" "Albania" "Algeria" ... $ imf : chr [1:184] "Islamic Republic of Afghanistan" "Albania" "Algeria" "Angola" ... $ iso : chr [1:249] "Afghanistan" "Aland Islands" "Albania" "Algeria" ... $ pwt : chr [1:189] "Afghanistan" "Albania" "Algeria" "Angola" ... $ wb : chr [1:217] "Afghanistan" "Albania" "Algeria" "American Samoa" ... $ who : chr [1:193] "Afghanistan" "Albania" "Algeria" "Andorra" ...

Examples

```
data(country.names)
```

country.regex *Country Name Regular Expressions*

Description

A data frame containing regular expressions for matching country names, where each country is identified by the ISO 3166-1 Alpha-2 code. Used internally by the standardize.countrynames function.

Usage

```
data(country.regex)
```

Format

A data frame with 260 observations on the following 2 variables.

`regex` a character vector

`code` a character vector

Examples

```
data(country.regex)
```

standardize.countrynames *Standardize Country Names*

Description

Takes in a dataframe or vector containing a column of country names and returns the data structure with the names standardized.

Usage

```
standardize.countrynames(input, input.column = NULL, standard = "default", standard.column = NULL, on
```

Arguments

`input`	A dataframe or vector containing a column of country names
`input.column`	The column containing country names if input is a dataframe, identified by name or number; ignored if input a vector
`standard`	The name of an included name set (see details), or a dataframe or vector containing a column of standard names

standard.column	The column containing standard names if standard is a dataframe, identified by name or number; ignored if standard a vector or an included name set
only.names	Only return a vector of standardized names
na.rm	Remove any countries not contained in the standard set
suggest	Suggestions for inexact matches; "prompt" allows user to select desired suggestions, "auto" applies all, "none" applies none
print.changes	Print which names changed
verbose	Print full output, including names of nonidentified countries

Details

Included name sets "default": Naming convention based on the ISO "imf": International Monetary Fund names "iso": International Standards Organization names "pwt": Penn World Tables names "wb": World Bank names "who: World Health Organization names

Value

If input a dataframe, returns the identical dataframe with the country names column standardized; if input a vector of country names, returns the standardized vector

Examples

```
library(StandardizeText)
sample.names <- c("Aland Is.","Brunei Daru.","Ivory Coast","The Gambia")
sample.std <- c("brunei","aland is","gambia, the","cote divoire")
sample.df <- data.frame(foo=2:5,bar=sample.names, baz=7:4, qux=sample.std)

#Standardize vector using iso names
out.a <- standardize.countrynames(sample.names,standard="iso",suggest="auto")
#Standardize vector using provided names
out.b <- standardize.countrynames(sample.names,standard=sample.std,suggest="auto")
#Standardize dataframe using wb names
out.c <- standardize.countrynames(sample.df,2,standard="wb",suggest="auto",verbose=TRUE)
#Standardize dataframe using provided names without suggestions
out.d <- standardize.countrynames(sample.df,"bar",sample.df,"qux",suggest="none",verbose=TRUE)
```

---

standardize.text *Standardize Text*

Description

Takes in a dataframe or vector containing a column of text and returns the data structure with the text standardized.

Usage

```
standardize.text(input, input.column = NULL, standard, standard.column = NULL, regex = NULL, codes = N
```

Arguments

input	A dataframe or vector containing a column of text
input.column	The column containing text if input is a dataframe, identified by name or number; ignored if input a vector
standard	A dataframe or vector containing a column of standard text
standard.column	The column containing standard text if standard is a dataframe, identified by name or number; ignored if standard a vector
regex	An optional vector of regular expressions; if NULL regex will be generated from standard
codes	An optional vector of identified codes; if NULL codes will be generated automatically
match	Mark true if there is a one-to-one correspondence between provided standard and provided regex
only.names	Only return a vector of standardized names
na.rm	Remove any entries whose text does not appear in the standard set
suggest	Suggestions for inexact matches; "prompt" allows user to select desired suggestions, "auto" applies all, "none" applies none
print.changes	Print which text entries changed
verbose	Print full output, including unidentified text

Value

If input a dataframe, returns the identical dataframe with the text column standardized; if input a vector of text, returns the standardized vector

Examples

```
library(StandardizeText)
sample.text <- c("blue car","STOP","email","tree")
sample.std <- c("the tree","car","e-mail","stop")
sample.df <- data.frame(foo=2:5,bar=sample.text, baz=7:4, qux=sample.std)
out.a <- standardize.text(sample.text,standard=sample.std,suggest="auto")
out.b <- standardize.text(sample.df,2,sample.df,"qux",suggest="auto")
```

Index