Package ‘stopwords’

December 14, 2017

Type Package
Title Multilingual Stopword Lists
Version 0.9.0
Description Provides multiple sources of stopwords, for use in text analysis and natural language processing.
License MIT + file LICENSE
Depends R (>= 2.10)
Imports ISOcodes
Suggests quanteda, testthat, covr
URL https://github.com/davnn/stopwords
BugReports https://github.com/davnn/stopwords/issues
Encoding UTF-8
LazyData true
RoxygenNote 6.0.1
NeedsCompilation no
Author Kenneth Benoit [aut],
David Muhr [aut, cre],
Kohei Watanabe [aut]
Maintainer David Muhr <muhrdavid+github@gmail.com>
Repository CRAN
Date/Publication 2017-12-14 14:16:27 UTC

R topics documented:

  stopwords-package .................................................. 2
  data_stopwords_misc .............................................. 2
  data_stopwords_smart ............................................. 3
  data_stopwords_snowball ......................................... 3
  data_stopwords_stopwordsiso ................................... 4
Description

Provides a `stopwords()` function to return character vectors of stopwords for different languages, using the ISO-639-1 language codes, and allows for different sources of stopwords to be defined.

Currently available sources

- **snowball** The Snowball stopword lists sources for multiple languages. Most of these have been ported from the `quanteda` stopword lists (in versions <1.0 of that package).
- **smart** The English-language stopword list from the SMART information retrieval system.
- **misc** A few additional stopword lists, including the non-Snowball word lists from `quanteda` versions < 1.0.

Author(s)

Kenneth Benoit, David Muhr, and Kohei Watanabe

Description

Other, miscellaneous stopword lists.

Format

An object of class `list` of length 4.

Usage

```r
stopwords(language, source = "misc")
```
Source

The Arabic stopwords come from https://sites.google.com/site/kevinbouge/stopwords-lists.
The Greek stopwords were supplied by Carsten Schwemmer (see https://github.com/kbenoit/quanteda/issues/282).
The Chinese stopwords are taken from the Baidu stopword list.

Description

The stopword lists based on the SMART (System for the Mechanical Analysis and Retrieval of Text) Information Retrieval System, an information retrieval system developed at Cornell University in the 1960s.

Format

An object of class list of length 1.

Usage

stopwords(language = "en", source = "smart")

Source

The English stopword list is taken from the online appendix 11 of Lewis et. al. (2004).

References


Description

snowball stopword list

Format

An object of class list of length 15.
Details
Provides stopword lists in multiple languages, based on the Snowball stemmer’s word lists.

Usage
stopwords(language, source = "snowball")

Source
The main stopword lists are taken from the Snowball stemmer project in different languages (see http://snowballstem.org/projects.html).

The stopword lists can be found in http://snowball.tartarus.org/dist/snowball_all.tgz.

See Also
stopwords

data_stopwords_stopwordsiso
multilingual stopwords from https://github.com/stopwords-iso/stopwords-iso

Description
The Stopwords ISO Dataset is the most comprehensive collection of stopwords for multiple languages. The collection follows the ISO 639-1 language code.

Format
A named list of length 57, of character vectors that represent stopwords in 57 languages. To see the languages available, use stopwords_getlanguages.

Usage
stopwords(language, source = "stopwords-iso")

Source
https://github.com/stopwords-iso/stopwords-iso/
stopwords

Collection of stopwords in multiple languages

Description

This function returns stopwords collated for Stopwords ISO library (https://github.com/stopwords-iso/stopwords-iso).

Usage

stopwords(language = "en", source = "snowball")

Arguments

language specify language of stopwords by ISO 639-1 code
source specify a stopwords source. To list the currently available options, use stopwords_getsources.

Details

The language codes for each stopword list use the two-letter ISO code from https://en.wikipedia.org/wiki/List_of_ISO_639-1_codes. For backwards compatibility, the full English names of the stopwords from the quanteda package may also be used, although these are deprecated.

Value

a character vector containing the stopwords

Examples

stopwords('en')
stopwords('de')

stopwords_getlanguages

list available stopwords country codes

Description

Lists the available stopwords country codes for a given stopwords source. See https://en.wikipedia.org/wiki/ISO_639-1 for details of the language code.

Usage

stopwords_getlanguages(source)

Arguments

source the source of the stopwords
stopwords_getsources

Description

Returns a character vector of the stopword sources available from the `stopwords` package.

Usage

```r
stopwords_getsources()
```
Index

*Topic datasets
  data_stopwords_misc, 2
  data_stopwords_smart, 3
  data_stopwords_snowball, 3
  data_stopwords_stopwordsiso, 4

data_stopwords_misc, 2
data_stopwords_smart, 3
data_stopwords_snowball, 3
data_stopwords_stopwordsiso, 4

misc, 2
smart, 2
snowball, 2
stopwords, 4, 5
stopwords-iso, 2
stopwords-package, 2
stopwords_getlanguages, 4, 5
stopwords_getsources, 5, 6