Package ‘stopwords’

July 24, 2019

Type Package
Title Multilingual Stopword Lists
Version 1.0
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 covr, lintr, quanteda, spelling, testthat
URL https://github.com/quanteda/stopwords
BugReports https://github.com/quanteda/stopwords/issues
Encoding UTF-8
LazyData true
Language en-GB
RoxygenNote 6.1.1
NeedsCompilation no
Author Kenneth Benoit [aut, cre],
    David Muhr [aut],
    Kohei Watanabe [aut]
Maintainer Kenneth Benoit <kbenoit@lse.ac.uk>
Repository CRAN
Date/Publication 2019-07-24 10:40:02 UTC

R topics documented:

  stopwords-package ............................................. 2
  data_stopwords_misc ......................................... 2
  data_stopwords_smart ......................................... 3
  data_stopwords_snowball ...................................... 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 5.

Usage

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/quanteda/quanteda/issues/282).
The Gujarati stopwords are taken from https://github.com/gujarati-ir/Gujarati-Stop-Words and modified by Chandrakant Bhogayata.
The Chinese stopwords are taken from the Baidu stopword list¹.

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 11² of Lewis et. al. (2004).

References


¹http://www.baiduguide.com/baidu-stopwords/
**data_stopwords_snowball**

*snowball stopword list*

**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**

```r
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/

Description

This function returns character vectors of stopwords for different languages, using the ISO-639-1 language codes, and allows for different sources of stopwords to be defined.

The default source is the Snowball stopwords collection but other sources are also available.

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**


**Usage**

```r
stopwords_getlanguages(source)
```

**Arguments**

- `source` the source of the stopwords

stopwords_getsources

*list available stopwords sources*

**Description**

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

**Usage**

```r
stopwords_getsources()
```