Package 'stopwords'

October 28, 2021

Type Package

Title Multilingual Stopword Lists

Version 2.3

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, 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 7.1.1

NeedsCompilation no

Author Kenneth Benoit [aut, cre],
David Muhr [aut],
Kohei Watanabe [aut]

Maintainer Kenneth Benoit <kbenoit@lse.ac.uk>

Repository CRAN

Date/Publication 2021-10-28 11:20:02 UTC

R topics documented:

`stopwords-package` *stopwords: one-stop shopping for stopwords in R*

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).

`stopwords-iso` The collection taken from `https://github.com/stopwords-iso/stopwords-iso/`.

`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.

`marimo` Stopword lists compiled by Kohei Watanabe.

Author(s)

Kenneth Benoit, David Muhr, and Kohei Watanabe

`data_stopwords_ancient`

stopword lists for ancient languages

Description

Stopword lists for ancient Greek and Latin. These lists are far more extensive than the Perseus lists for ancient Greek and Latin from the Perseus Digital Library.

Format

An object of class `list` of length 2.

Details

As there is no 2-letter code for ancient Greek in ISO-639-1, we use "grc" to denote Greek (as per ISO-639-3).

Usage

```
stopwords(language = "grc",source = "ancient")
stopwords(language = "la",source = "ancient")
```

Source

Aurélien Berra, Ancient Greek and Latin stopwords, doi: 10.5281/zenodo.1165205. See https://github.com/aurelberra/stopwords/blob/master/rationale.md.

See Also

data_stopwords_perseus

---

data_stopwords_marimo *stopword lists including parts-of-speech*

Description

Stopword lists that include specific parts of speech, maintained by Kohei Watanabe.

Format

An object of class `list` of length 8.

Details

These are multi-level lists, in the original data. If you wish to use them as lists, please access the data object directly.

Usage

```
stopwords(language = "en",source = "marimo")
```

Source

The English version was adopted from the Snowball collection, and then extended and translated into other languages by contributors. Names of contributors are in the header of the original YAML files.

Examples

```
# access English pronouns directly
stopwords::data_stopwords_marimo$en$pronoun
```

data_stopwords_misc *miscellaneous stopword lists*

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 Catalan stopwords come from http://latel.upf.edu/morgana/altres/pub/ca_stop.htm.

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 list (see http://www.baiduguide.com/baidu-stopwords/).

data_stopwords_nltk *stopword lists from the Python NLTK library*

Description

Stopword lists for 23 languages from the Python NLTK library.

Format

An object of class `list` of length 23.

Usage

```
stopwords(language = "en",source = "nltk")
```

Source

https://github.com/nltk/nltk_data/blob/gh-pages/packages/corpora/stopwords.zip

References

Bird, Steven, Edward Loper and Ewan Klein (2009). Natural Language Processing with Python. O'Reilly Media Inc.

`data_stopwords_perseus` *stopword lists for ancient languages - Perseus Digital Library*

Description

Stopword lists for ancient Greek and Latin. As there is no 2-letter code for ancient Greek in ISO-639-1, we use "grc" to denote Greek (as per ISO-639-3).

Format

An object of class `list` of length 2.

Usage

```
stopwords(language = "grc",source = "perseus")
stopwords(language = "la",source = "perseus")
```

Source

The Perseus Digital Library. See https://wiki.digitalclassicist.org/Stopwords_for_Greek_and_Latin and https://wiki.digitalclassicist.org/Perseus_Digital_Library.

`data_stopwords_smart` *stopword lists from the SMART system*

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

Lewis, David D., et al. (2004) "Rcv1: A new benchmark collection for text categorization research." *Journal of machine learning research* 5: 361-397.

`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

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

Source

The main stopword lists are taken from the Snowball stemmer project in different languages (see https://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 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", simplify = TRUE)
```

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()`.
`simplify`	logical; if `TRUE` return a simple vector, if `FALSE` return a list if the original word list was nested

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, or a list of characters `simplify = FALSE`

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 *list available stopwords sources*

Description

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

Usage

```
stopwords_getsources()
```

Index