

# Package ‘stopwords’

April 14, 2020

**Type** Package

**Title** Multilingual Stopword Lists

**Version** 2.0

**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, usethis, desc

**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** 7.1.0

**NeedsCompilation** no

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

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

**Repository** CRAN

**Date/Publication** 2020-04-14 20:10:02 UTC

## R topics documented:

stopwords-package . . . . .	2
data_stopwords_ancient . . . . .	2
data_stopwords_marimo . . . . .	3
data_stopwords_misc . . . . .	4

data_stopwords_nltk . . . . .	4
data_stopwords_smart . . . . .	5
data_stopwords_snowball . . . . .	5
data_stopwords_stopwordsiso . . . . .	6
stopwords . . . . .	7
stopwords_getlanguages . . . . .	7
stopwords_getsources . . . . .	8
use_stopwords . . . . .	8

## Index 9

---

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

**Source**

The [Digital Classicist Wiki](http://wiki.digitalclassicist.org/Main_Page). See [http://wiki.digitalclassicist.org/Main\\_Page](http://wiki.digitalclassicist.org/Main_Page).

---

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 6.

**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](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](#).

---

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](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\_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 stopwords lists are taken from the Snowball stemmer project in different languages (see <http://snowballstem.org/projects.html>).

The stopwords lists can be found in [http://snowball.tartarus.org/dist/snowball\\_all.tgz](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

<code>language</code>	specify language of stopwords by ISO 639-1 code
<code>source</code>	specify a stopwords source. To list the currently available options, use <code>stopwords_getsources()</code> .
<code>simplify</code>	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 stopwords list use the two-letter ISO code from [https://en.wikipedia.org/wiki/List\\_of\\_ISO\\_639-1\\_codes](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](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()
```

---

```
use_stopwords            Use stopwords in your package
```

---

**Description**

Sets up your package to import and re-export the `stopwords()` function in your package. This function requires the use of the **roxygen2** package.

- Adds `stopwords`` to "Imports" in DESCRIPTION'.
- Imports `stopwords()`, which is necessary for internal use.
- Exports `stopwords()`, if `export = TRUE`, which is necessary to make the function available to the users of your package.

**Usage**

```
use_stopwords(export = TRUE)
```

**Arguments**

export            If TRUE, the file R/stopwords.R is added, which provides the roxygen template to import and re-export `stopwords()`. If FALSE, the necessary roxygen directive is added, if possible, or otherwise instructions are given.

**Examples**

```
## Not run:
use_stopwords()

## End(Not run)
```



# Index

## \*Topic **datasets**

- data\_stopwords\_ancient, 2
- data\_stopwords\_marimo, 3
- data\_stopwords\_misc, 4
- data\_stopwords\_nltk, 4
- data\_stopwords\_smart, 5
- data\_stopwords\_snowball, 5
- data\_stopwords\_stopwordsiso, 6

- data\_stopwords\_ancient, 2
- data\_stopwords\_marimo, 3
- data\_stopwords\_misc, 4
- data\_stopwords\_nltk, 4
- data\_stopwords\_smart, 5
- data\_stopwords\_snowball, 5
- data\_stopwords\_stopwordsiso, 6

- marimo, 2
- misc, 2

- other(), 7

- smart, 2
- snowball, 2
- Snowball(), 7
- stopwords, 7
- stopwords(), 6, 8
- stopwords-iso, 2
- stopwords-package, 2
- stopwords\_getlanguages, 7
- stopwords\_getlanguages(), 6
- stopwords\_getsources, 8
- stopwords\_getsources(), 7

- use\_stopwords, 8