Package 'morestopwords'

June 12, 2023

Type Package

Title All Stop Words in One Place

Version 0.2.0

Maintainer Fabio Ashtar Telarico <Fabio-Ashtar.Telarico@fdv.uni-lj.si>

Description A standalone package combining several stop-word lists for 65 languages with a median of 329 stop words for language and over 1,000 entries for English, Breton, Latin, Slovenian, and Ancient Greek! The user automatically gets access to all the unique stop words contained in: the 'StopwordISO' repository; the 'Natural Language Toolkit' for 'python'; the 'Snowball' stop-word list; the R package 'quanteda'; the 'marimo' repository; the 'Perseus' project; and A. Berra's list of stop words for Ancient Greek and Latin.

License MIT + file LICENSE

URL https://fatelarico.github.io/morestopwords.html

BugReports https://github.com/FATelarico/morestopwords/issues

Encoding UTF-8

Depends R (>= 2.10)

LazyData no

RoxygenNote 7.2.3

Suggests cld2

NeedsCompilation no

Author Fabio Ashtar Telarico [aut, cre]
(<https://orcid.org/0000-0002-8740-7078>),
Kohei Watanabe [aut]

Repository CRAN

Date/Publication 2023-06-12 09:30:02 UTC

R topics documented:

`languages` *Returns ISO codes and names for all language or only those available in this package*

Description

See the relevant Wikipedia article for details on the language codes.

Usage

```
languages(available = TRUE)
```

Arguments

`available`	*logical*, whether to return only the languages supported in this package.

Details

Note that:

- the ISO 639-1 code for mainland Chinese was changed to `zh-cn`.
- A list of stop words in the variety of Chinese spoken in the island of Taiwan is accessible using the ISO 639-1 `zh-tw` or the name `'Chinese Taiwan'`.
- Ancient Greek has been assigned an artifact ISO 639-1 code (`gr`) because it had none. Its ISO 639-2 and 639-3 codes are both `grc`.

Value

A data frame with a row for each languages (only those supported if `available` is `TRUE`) and columns for the several ISO codes (639-2, 639-3, 639-1) and the name.

Examples

```
# Return all languages in the ISO 639-2/3 standard
languages()
```

`remove.stopwords` *Removes stop words for a string the language of which is known*

Description

Removes stop words for a string the language of which is known

Usage

```
remove.stopwords(str, lang = "auto", fallback = "English")
```

Arguments

`str`	A string or a vector of strings which to delete the stop words from
`lang`	Either: • `'auto'` in which case `cld2` is used to perform language detection; or • A string (or a vector of strings, depending on `str`) representing an ISO 639-2/3 or a language name from which to derive a ISO 639-2 code (for language names, string matching is performed)
`fallback`	Fallback language in case `cld2` fails to detect the language of the manually-specified string does not match a supported language. Default to `'English'`.

Value

A strings (or a vector, depending on `str`) corresponding to the string/s `str` without stop words for the language/s `lang`.

Examples

```
# Multiple strings in different languages
remove.stopwords(str = c(Gibberish = 'dadas',
                         Catalan = 'Adeu amic meu',
                         Irish = 'Slan a chara',
                         French = 'Je suis en Allemagne',
                         German = 'Eich liebe Deutschland'),
                 # Various ways of indicating the language
                 lang = c(NA, 'cata', 'Iris', 'fr', 'deu'),
                 # Yet another way
                 fallback = 'english'
                 )
```

stopwords *Collection of stopwords in multiple languages*

Description

This function returns stop words contained in the StopwordsISO repository.

Usage

```
stopwords(lang = "en")
```

Arguments

lang	Language for which to retrieve the stop word among those supported. This parameters supports: • three-letter ISO 639-2/3 codes (e.g., 'eng'); • two-letter ISO639-1 codes ('en'); • names based ISO 639-2 codes ('English' or 'english') and their unambiguous substrings ('engl', 'engli', etc.).

Value

A character vector containing the stop words from the selected language as listed in the StopwordISO repository.

Examples

```
# They all return the correct list of stop words!

stopwords('German')
stopwords('germ')
stopwords('de')
stopwords('deu')
```

stopwordsISO *Combined stop words for all languages*

Description

A list of stop words in each of the supported languages

Usage

```
stopwordsISO
```

Format

An object of class `list` of length 65.

Details

Note: All Unicode characters are escaped. To un-escape them, consider using:

```
library(AllStopwords)
if(!requireNamespace('stringi')){
  install.packages('stringi')
}
data('stopwordsISO')
stopwords_unescaped <- lapply(stopwordsISO,
                              stringi::stri_unescape_unicode)
```

Author(s)

Each stop-word list's Authors

Source

All unique stopwords in the following databases:

- the StopwordISO repository;
- python's Natural Language Toolkit (nltk);
- the Snowball stop-word list;
- the R package quanteda;
- the marimo repository;
- the Perseus project; and
- Aurélien Berra's list of stop words for Ancient Greek and Latin (doi:10.5281/zenodo.3860343).

Index