Package 'tidystopwords'

October 27, 2021

Type Package

Title Customisable Stop-Words in 110 Languages

Version 0.9.1

Date 2021-10-24

Author Silvie Cinkova [aut],
Maciej Eder [aut, cre]

Maintainer Maciej Eder <maciejeder@gmail.com>

Depends R (>= 3.5.0)

Imports dplyr

Description Functions to generate stop-word lists in 110 languages, in a way consistent across all the languages supported. The generated lists are based on the morphological tagset from the Universal Dependencies.

License GPL (>= 3)

Encoding UTF-8

Suggests knitr, rmarkdown

VignetteBuilder knitr

NeedsCompilation no

Repository CRAN

Date/Publication 2021-10-27 12:10:02 UTC

R topics documented:

`generate_stoplist` *Listing of stop words in different languages.*

Description

Generate a vector of stop words in one or several languages.

Usage

```
generate_stoplist(language = NULL, output_form = 1)
```

Arguments

`language`	single string or a character vector. `NULL` by default. The strings can be language names or ISO-639 language codes as listed by the `list_supported_languages()`, freely combined, case-sensitive. When no language is recognized, the following error message appears: "The language name or language id you have selected is not supported. (Or you didn't specify a language at all). Check out the supported languages by calling 'list_supported_languages'.".
`output_form`	default `1`, alternatively `2` or `3`. Option `1` returns a character vector of unique stopwords word forms. Option `2` returns a named vector whose elements are the stopwords word forms and names are the associated stop classes. One word form can occur with different stop classes; hence the word forms in this vector are not unique, unlike Option `1`. Option `3` returns a data frame filtered according to the language selection.

Value

The function comes with three output options.

- Option '1' outputs a character vector of unique word forms.
- Option '2' outputs a named character vector of word forms. The names denote 'stop classes' roughly corresponding to parts of speech. Note that, in this output, the word forms are not unique. For instance, in English stopwords, *that* would occur as a subordinating conjunction as well as as a pronoun.
- Option '3' (the default) outputs a data frame, where each row represents a combination of language (columns 'lang_name' and 'lang_id'), word form and word lemma (columns 'form' and 'lemma'), and several other columns explained below.

All outputs are encoded in UTF-8.

Warning

- The function stops when no language is selected.
- The stop classes (pre-defined linguistic filters) are not mutually exclusive. Their overlap varies among languages.

- The stoplists are fully data-driven. We have set a threshold of 3 occurrences of a combination of language, form, lemma, and upos to remove obvious noise, but some noise is bound to have come through anyway. It is mainly foreign words that were given a regular upos tag (e.g. the English "and" has sneaked in among the German coordinating conjunctions). Another known case is the `contraction` stop class in English, which, among well-suited instances such as *ain't* includes uses of the so-called Saxonic genitive (e.g. *world's*). Many languages are represented by balanced and large corpora of standard written texts, but some are not; e.g. based mainly on a Bible translation or Wikipedia. Hence also their stopwords can be biased.

Author(s)

Silvie Cinková, Maciej Eder

References

The underlying data frame 'multilingual_stoplist' is based on the official release of Version 2.8 of Universal Dependencies.

https://universaldependencies.org

Zeman, Daniel; et al., 2021, Universal Dependencies 2.8.1, LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, http://hdl.handle.net/11234/1-3687.

See Also

`list_supported_languages`, `multilingual_stoplist`

Examples

```
generate_stoplist(language = "English", output_form = 1)

generate_stoplist(language = "English", output_form = 2)

generate_stoplist(language = "English", output_form = 3)
```

`list_supported_languages`

Listing of languages supported by `list_supported_languages` *by their names and ISO-639 codes in a data frame.*

Description

Generate a data frame containing language names and their corresponding ISO-639 codes, with numbers of stop words for the given language

Usage

```
list_supported_languages()
```

Arguments

No arguments.

Value

A grouped tibble (data frame) with three columns:

Author(s)

Silvie Cinková, Maciej Eder

References

The underlying data frame ‘multilingual_stoplist‘ is based on the official release of Version 2.8 of Universal Dependencies.

https://universaldependencies.org

Zeman, Daniel; et al., 2021, Universal Dependencies 2.8.1, LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, http://hdl.handle.net/11234/1-3687.

See Also

generate_stoplist, multilingual_stoplist

Examples

```
list_supported_languages()
```

---

multilingual_stoplist *Multilingual Stop-Word List*

Description

This dataset contains a dataframe with individual word forms in rows. You can control the part of speech and various frequency counts of your desired stop-word list.

Format

A data frame encoded in UTF-8, with the following columns:

- abbreviation: common abbreviations acting as adverbs or adjectives, for instance *e.g., etc., cf.*;
- adposition: prepositions or postpositions (e.g. *in*, *ago*);
- auxiliary_verb: auxiliary or modal verb (e.g. *would*);
- conjunction_subordinator: coordinating or subordinating conjunctions(e.g. *and*, *because*);

- `contractions`: contracted forms (e.g. *'n'* or *she'd*);
- `determiner_quantifier`: pronouns, articles, pronominal adverbs, and some numerals not written as digits - all acting as adjectives or adverbs, not nouns (e.g. *yours*, *the*, *both* ,*where*, *twofold*. Cf. `pronominals`;
- `interjection`: words denoting sounds and performative words like *yes*, *no*, *please*, *thanks*;
- `particle`: either preposition-like words in phrasal verbs (e.g. in English) or diverse words indicating the speaker's attitude to the statement (e.g. *fortunately*);
- `pronominal`: pronouns acting as nouns (e.g. *we* - cf. `determiner_quantifier`)

Details

This data frame has been derived from an official release of the Universal Dependencies (UD) treebanks. Treebanks are text corpora with linguistic annotation. The UD syntactic annotation follows the principles of dependency syntax. The annotation encompasses for each text token:

- relevant morphological categories;
- lemma (the vocabulary form; e.g. active present infinitive in verbs)
- a reference to its syntactically governing word in the clause; e.g. "house" governs "old" in "old house".
- the type of the syntactic dependency between the word and its governing word; e.g. "attribute".

Source

The data set is based on the official release of Version 2.8.1 of the Universal Dependencies stored in the LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, Czech Republic, http://hdl.handle.net/11234/1-3687.

References

https://universaldependencies.org

Zeman, Daniel; et al., 2021, Universal Dependencies 2.8.1, LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (UFAL), Faculty of Mathematics and Physics, Charles University, http://hdl.handle.net/11234/1-3687.

`tidystopwords` *Customisable Lists of Stop-Words in 110 Languages*

Description

The idea behind this package is to give the user control over the stop-word selection.

Details

The idea behind this package is to give the user control over the stop-word selection. The core `generate_stoplist` function relies on `multilingual_stopwords`, a large data frame derived from the current release of the Universal Dependencies Treebanks. We have included all languages whose corpora totalled above 10,000 tokens – large enough to cover all common closed-class words, such as prepositions, conjunctions, and auxiliary verbs. The data comes encoded in UTF-8.

Author(s)

Silvie Cinková, Maciej Eder

References

The data set is based on the official release of Version 2.1 of Universal Dependencies.

https://universaldependencies.org

Nivre, Joakim; Agić, Željko; Ahrenberg, Lars; et al., 2017, Universal Dependencies 2.1, LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, http://hdl.handle.net/11234/1-2515.

See Also

`list_supported_languages`, `multilingual_stoplist`

Index