Package ‘spell.replacer’

September 3, 2025

Title Probabilistic Spelling Correction in a Character Vector

Version 1.0.1

Description Automatically replaces ``misspelled'' words in a character vector based on their string distance from a list of words sorted by their frequency in a corpus. The default word list provided in the package comes from the Corpus of Contemporary American English. Uses the Jaro-Winkler distance metric for string similarity as implemented in van der Loo (2014) <doi:10.32614/RJ-2014-011>. The word frequency data is derived from Davies (2008-) ``The Corpus of Contemporary American English (COCA)'' <https://www.english-corpora.org/coca/>.

License MIT + file LICENSE

Encoding UTF-8

LazyData true

RoxygenNote 7.3.2

Depends R (>= 2.10)

Imports hunspell, stringr, stringdist, textclean

Suggests rmarkdown, knitr, testthat (>= 3.0.0)

VignetteBuilder knitr

Config/testthat/edition 3

NeedsCompilation no

Author David Brown [aut, cre] (ORCID: <https://orcid.org/0000-0001-7745-6354>)

Maintainer David Brown <dwb2@andrew.cmu.edu>

Repository CRAN

Date/Publication 2025-09-03 21:10:02 UTC

Contents

coca_list *COCA Word List*

Description

A character vector containing the 100,000 most frequent words from the Corpus of Contemporary American English (COCA), sorted by frequency from most to least frequent. This word list serves as the default reference for spelling correction in the spell_replace function.

Usage

```
coca_list
```

Format

A character vector with 100,000 elements:

> Each element is a word from COCA, with the first element being the most frequent word ("the") and subsequent elements decreasing in frequency.

Source

Corpus of Contemporary American English (COCA) https://www.english-corpora.org/coca/

Examples

```
# View the first 10 most frequent words
head(coca_list, 10)

# Check if a word is in the list
"hello" %in% coca_list

# Find the rank of a specific word
which(coca_list == "hello")
```

correct *Correct a Single Misspelled Word*

Description

Finds the best correction for a single misspelled word using string distance and frequency-based ranking from a sorted word list.

Usage

```
correct(word, sorted_words, ignore_punct = FALSE, threshold = 0.12)
```

Arguments

`word` A character string representing the misspelled word

`sorted_words` A character vector of correctly spelled words sorted by frequency

`ignore_punct` Logical. If TRUE, ignores punctuation when calculating string distance

`threshold` Numeric. Maximum string distance threshold for considering a word as a correction candidate

Value

A character string with the corrected word

`spell_replace` *Probabilistic Spelling Correction*

Description

Automatically replaces misspelled words in a character vector based on their string distance from a list of words sorted by frequency in a corpus.

Usage

```
spell_replace(
  txt,
  word_list = coca_list,
  ignore_names = TRUE,
  threshold = 0.12,
  ignore_punct = FALSE
)
```

Arguments

`txt` A character vector containing text to be spell-checked

`word_list` A character vector of correctly spelled words sorted by frequency (default: coca_list)

`ignore_names` Logical. If TRUE, ignores potential proper names (capitalized words that appear multiple times)

`threshold` Numeric. Maximum string distance threshold for considering a word as a correction candidate (default: 0.12)

`ignore_punct` Logical. If TRUE, ignores punctuation when calculating string distance

Value

A character vector with corrected spellings

Index