Package ‘SnowballC’

January 15, 2019

Type Package
Version 0.6.0
Date 2019-01-07
Title Snowball Stemmers Based on the C 'libstemmer' UTF-8 Library
Description An R interface to the C 'libstemmer' library that implements
Porter's word stemming algorithm for collapsing words to a common
root to aid comparison of vocabulary. Currently supported languages are
Danish, Dutch, English, Finnish, French, German, Hungarian, Italian,
Norwegian, Portuguese, Romanian, Russian, Spanish, Swedish
and Turkish.
License BSD_3_clause + file LICENSE
Copyright Dr Martin Porter (2001) and Richard Boulton (2004, 2005) for
the 'libstemmer' C library, and Milan Bouchet-Valat (2013) for
the R package contents.
URL https://r-forge.r-project.org/projects/r-temis/
BugReports https://r-forge.r-project.org/tracker/?group_id=1437
NeedsCompilation yes
Author Milan Bouchet-Valat [aut, cre]
Maintainer Milan Bouchet-Valat <nalimilan@club.fr>
Repository CRAN
Date/Publication 2019-01-15 16:10:03 UTC

R topics documented:

getStemLanguages ................. 2
wordStem .................................. 3

Index 5
getStemLanguages  Query the list of supported languages

Description

This dynamically determines the names of the languages for which stemming is currently supported by this package.

Usage

getStemLanguages()

Details

The language names in lower case are returned, though please note that two- and three-letter ISO-639 codes are also accepted by wordStem (see references for the list of codes).

This queries the C code for the list of languages that were compiled when the package was installed which in turn is determined by the code that was included in the distributed package itself.

Value

A character vector giving the names of the languages.

Author(s)

Milan Bouchet-Valat

References

http://snowball.tartarus.org/

See Also

wordStem

Examples

getStemLanguages()
**wordStem**

*Get the stem of words*

---

**Description**

This function extracts the stems of each of the given words in the vector.

**Usage**

```r
wordStem(words, language = "porter")
```

**Arguments**

- **words**: a character vector of words whose stems are to be extracted.
- **language**: the name of a recognized language, as returned by `getStemLanguages`, or a two- or three-letter ISO-639 code corresponding to one of these languages (see references for the list of codes).

**Details**

This uses Dr. Martin Porter’s stemming algorithm and the C libstemmer library generated by Snowball.

**Value**

A character vector with as many elements as there are in the input vector with the corresponding elements being the stem of the word. Elements of the vector are converted to UTF-8 encoding before the stemming is performed, and the returned elements are marked as such when they contain non-ASCII characters.

**Author(s)**

Milan Bouchet-Valat

**References**

[http://snowball.tartarus.org/](http://snowball.tartarus.org/)


**Examples**

```r
# Simple example
wordStem(c("win", "winning", "winner"))

# Test the supplied vocabulary
for(lang in getStemLanguages()) {
  load(system.file("words", paste0(lang, ".RData"), package="SnowballC"))
}
stopifnot(all(wordStem(voc[[1]], lang) == voc[[2]]))

stopifnot(is.na(wordStem(NA)))
Index

getStemLanguages, 2, 3

wordStem, 2, 3