Package: slowraker (via r-universe)

September 2, 2024

Type Package

Title A Slow Version of the Rapid Automatic Keyword Extraction (RAKE) Algorithm

Version 0.1.1

Description A mostly pure-R implementation of the RAKE algorithm (Rose, S., Engel, D., Cramer, N. and Cowley, W. (2010) <doi:10.1002/9780470689646.ch1>), which can be used to extract keywords from documents without any training data.

URL https://crew102.github.io/slowraker/index.html

BugReports https://github.com/crew102/slowraker/issues

License MIT + file LICENSE

Encoding UTF-8

LazyData TRUE

Depends R (>= 3.1)

Imports SnowballC, NLP, openNLP, utils

Suggests testthat, knitr, rmarkdown, covr

SystemRequirements Java (>= 5.0)

RoxygenNote 7.3.2

VignetteBuilder knitr

Repository https://crew102.r-universe.dev

RemoteUrl https://github.com/crew102/slowraker

RemoteRef HEAD

RemoteSha 294bb63410a07dc5644151b418c21665a7d3f569

Contents

`dog_pubs` *Dog publications*

Description

A data frame containing PLOS publication data for publications related to dogs. The purpose of this data frame is to provide an example of some text to extract keywords from.

Usage

```
dog_pubs
```

Format

A data frame with 30 rows and 3 variables:

doi The publication's DOI

title The publication's title

abstract The publication's abstract

`pos_tags` *Part-of-speech (POS) tags*

Description

A data frame containing all possible parts-of-speech, as per the openNLP package. This list was taken from Part-Of-Speech Tagging with R. `pos_tags` contains the following two columns:

tag The abbreviation for the part-of-speech (i.e., its tag)

description A short description of the part-of-speech

Usage

```
pos_tags
```

Format

An object of class `data.frame` with 36 rows and 2 columns.

rbind_rakelist *rbind a rakelist*

Description

rbind a rakelist

Usage

```
rbind_rakelist(rakelist, doc_id = NULL)
```

Arguments

rakelist	An object of class rakelist, which you create by calling slowrake.
doc_id	An optional vector of document IDs, which should be the same length as rakelist. These IDs will be added to the resulting data frame.

Value

A single data frame which contains all documents' keywords. The doc_id column tells you which document a keyword was found in.

Examples

```
rakelist <- slowrake(txt = dog_pubs$abstract[1:2])

# Without specifying doc_id:
head(rbind_rakelist(rakelist))

# With specifying doc_id:
head(rbind_rakelist(rakelist, doc_id = dog_pubs$doi[1:2]))
```

slowrake *Slow RAKE*

Description

A relatively slow version of the Rapid Automatic Keyword Extraction (RAKE) algorithm. See Automatic keyword extraction from individual documents for details on how RAKE works or read the "Getting started" vignette (vignette("getting-started")).

Usage

```
slowrake(
  txt,
  stop_words = smart_words,
  stop_pos = c("VB", "VBD", "VBG", "VBN", "VBP", "VBZ"),
  word_min_char = 3,
  stem = TRUE,
  phrase_delims = "[[:space:]]-[[:space:]]|[,.?():;\"!/]|]|\\["
)
```

Arguments

txt	A character vector, where each element of the vector contains the text for one document.
stop_words	A vector of stop words which will be removed from your documents. The default value (smart_words) contains the 'SMART' stop words (equivalent to tm::stopwords('SMART')) . Set stop_words = NULL if you don't want to remove stop words.
stop_pos	All words that have a part-of-speech (POS) that appears in stop_pos will be considered a stop word. stop_pos should be a vector of POS tags. All possible POS tags along with their definitions are in the pos_tags data frame (View(slowraker::pos_tags)). The default value is to remove all words that have a verb-based POS (i.e., stop_pos = c("VB", "VBD", "VBG", "VBN", "VBP", "VBZ")). Set stop_pos = NULL if you don't want a word's POS to matter during keyword extraction.
word_min_char	The minimum number of characters that a word must have to remain in the corpus. Words with fewer than word_min_char characters will be removed before the RAKE algorithm is applied. Note that removing words based on word_min_char happens before stemming, so you should consider the full length of the word and not the length of its stem when choosing word_min_char.
stem	Do you want to stem the words before running RAKE?
phrase_delims	A regular expression containing the patterns that will be used as phrase delimiters.

Value

An object of class rakelist, which is just a list of data frames (one data frame for each element of txt). Each data frame will have the following columns:

keyword A keyword that was identified by RAKE.

freq The number of times the keyword appears in the document.

score The keyword's score, as per the RAKE algorithm. Keywords with higher scores are considered to be higher quality than those with lower scores.

stem If you specified stem = TRUE, you will get the stemmed versions of the keywords in this column. When you choose stemming, the keyword's score (score) will be based off its stem, but the reported number of times that the keyword appears (freq) will still be based off of the raw, unstemmed version of the keyword.

Examples

```
slowrake(txt = "some text that has great keywords")

slowrake(txt = dog_pubs$title[1:2], stem = FALSE)
```

`smart_words` *SMART stop words*

Description

A vector containing the SMART information retrieval system stop words. See tm::stopwords('SMART') for more details.

Usage

```
smart_words
```

Format

An object of class `character` of length 571.

Index