Package ‘rapidraker’

January 5, 2018

Type Package

Title Rapid Automatic Keyword Extraction (RAKE) Algorithm

Version 0.1.0

Description A 'Java' 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/articles/rapidraker.html`

BugReports `https://github.com/crew102/rapidraker/issues`

License MIT + file LICENSE

Encoding UTF-8

LazyData true

Depends R (>= 3.1)

Imports rJava, openNLPdata, slowraker, utils

Suggests knitr, rmarkdown, testthat

SystemRequirements Java (>= 7.0)

RoxygenNote 6.0.1.9000

NeedsCompilation no

Author Christopher Baker [aut, cre]

Maintainer Christopher Baker <`chriscrewbaker@gmail.com`>

Repository CRAN

Date/Publication 2018-01-05 19:14:49 UTC

R topics documented:

rapidrake 2

Index **3**

rapidrake *Rapid RAKE*

Description

A relatively fast version of the Rapid Automatic Keyword Extraction (RAKE) algorithm. See Automatic keyword extraction from individual documents for details on how RAKE works.

Usage

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

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 characters that will be used as phrase delimiters

Examples

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

Index

pos_tags, *2*

rapidrake, 2