Package 'ngramrr'

April 6, 2015

Title A Simple General Purpose N-Gram Tokenizer

Version 0.1.1

Date 2015-04-05

Author Chung-hong Chan <chainsawtiney@gmail.com>

Maintainer Chung-hong Chan <chainsawtiney@gmail.com>

Description A simple n-gram (contiguous sequences of n items from a given sequence of text) tokenizer to be used with the 'tm' package with no 'rJava'/'RWeka' dependency.

URL https://github.com/chainsawriot/ngramrr

Depends R (>= 3.0.0)

License GPL-2

LazyData true

Imports tau

Suggests tm, testthat, magrittr

NeedsCompilation no

Repository CRAN

Date/Publication 2015-04-06 07:36:48

R topics documented:

ngramrr *General purpose n-gram tokenizer*

Description

A non-Java based n-gram tokenizer to be used with the tm package. Support both character and word n-gram.

Usage

```
ngramrr(x, char = FALSE, ngmin = 1, ngmax = 2, rmEOL = TRUE)
```

Arguments

x	input string.
char	logical, using character n-gram. char = FALSE denotes word n-gram.
ngmin	integer, minimun order of n-gram, ignore when char = TRUE
ngmax	integer, maximun order of n-gram
rmEOL	logical, remove ngrams wih EOL character

Value

vector of n-grams

Examples

```
require(tm)

nirvana <- c("hello hello hello how low", "hello hello hello how low",
"hello hello hello how low", "hello hello hello",
"with the lights out", "it's less dangerous", "here we are now", "entertain us",
"i feel stupid", "and contagious", "here we are now", "entertain us",
"a mulatto", "an albino", "a mosquito", "my libido", "yeah", "hey yay")

ngramrr(nirvana[1], ngmax = 3)
ngramrr(nirvana[1], ngmax = 3, char = TRUE)
nirvanacor <- Corpus(VectorSource(nirvana))
TermDocumentMatrix(nirvanacor, control = list(tokenize = function(x) ngramrr(x, ngmax =3)))

# Character ngram

TermDocumentMatrix(nirvanacor, control = list(tokenize =
function(x) ngramrr(x, char = TRUE, ngmax =3), wordLengths = c(1, Inf)))
```

Index