Package: CTM (via r-universe)

August 28, 2024

Type Package

Title A Text Mining Toolkit for Chinese Document

Version 0.2

Date 2016-11-28

Author Jim Liu, Quan Gu

Maintainer Jim Liu <jimliu741523@gmail.com>

Description The CTM package is designed to solve problems of text mining and is specific for Chinese document.

License GPL-3

LazyData TRUE

RoxygenNote 5.0.1

Imports jiebaR, plyr

NeedsCompilation no

Date/Publication 2016-11-28 08:20:59

Repository https://jimliu741523.r-universe.dev

RemoteUrl https://github.com/cran/CTM

RemoteRef HEAD

RemoteSha fd7a9320ee5008c6d1e2ac265fc02014cdf7dc37

Contents

CDTM   *Document Term Matrix*

Description

Constructs Document-Term Matrix from Chinese Text Documents.

Usage

```
CDTM(doc, weighting, EngTermDeleted = TRUE, NumTermDeleted = TRUE,
  shortTermDeleted = TRUE)
```

Arguments

doc   The Chinese text document. A vector of Chinese strings.

weighting   Available weighting function with matrix are binary, count, tf, tfidf. See details.

EngTermDeleted   remove English from text documents.

NumTermDeleted   remove Numbers from text documents.
shortTermDeleted
                    Deltected short word when nchar <2.

Details

This function run a Chinese word segmentation by jiebeR and build document-term matrix, and there is four weighting function with matrix, and "binary" means value can only be 1 if the term occurs, "count" means how many times the term occurs in a doc, "tf" means term frequency and "tfidf" means term frequency inverse document frequency.

Author(s)

Jim Liu, Quan Gu

Examples

```
library(CTM)
a1 <- "hello taiwan"
b1 <- "world of tank"
c1 <- "taiwan weather"
d1 <- "local weather"
text1 <- t(data.frame(a1,b1,c1,d1))
dtm1 <- CTDM(doc = text1, weighting = "tfidf",EngTermDeleted = FALSE, shortTermDeleted = FALSE)
```

CTDM   *Term Document Matrix*

Description

Constructs Term-Document Matrix from Chinese Text Documents.

Usage

```
CTDM(doc, weighting, EngTermDeleted = TRUE, NumTermDeleted = TRUE,
  shortTermDeleted = TRUE)
```

Arguments

doc   The Chinese text document. A vector of Chinese strings.

weighting   Available weighting function with matrix are binary, count, tf, tfidf. See details.

EngTermDeleted remove English from text documents.

NumTermDeleted remove Numbers from text documents.

shortTermDeleted

                    Deltected short word when nchar <2.

Details

This function run a Chinese word segmentation by jiebeR and build term-document matrix, and there is four weighting function with matrix, and "binary" means value can only be 1 if the term occurs, "count" means how many times the term occurs in a doc, "tf" means term frequency and "tfidf" means term frequency inverse document frequency.

Author(s)

Jim Liu, Quan Gu

Examples

```
library(CTM)
a1 <- "hello taiwan"
b1 <- "world of tank"
c1 <- "taiwan weather"
d1 <- "local weather"
text1 <- t(data.frame(a1,b1,c1,d1))
tdm1 <- CTDM(doc = text1, weighting = "tfidf", EngTermDeleted = FALSE, shortTermDeleted = FALSE)
```

---

termCount *Term Count*

Description

Computing term count from text documents

Usage

```
termCount(doc, EngTermDeleted = TRUE, NumTermDeleted = TRUE,
  shortTermDeleted = TRUE)
```

Arguments

doc   The Chinese text document.

EngTermDeleted remove English from text documents.

NumTermDeleted remove Numbers from text documents.
shortTermDeleted
                    Deltected short word when nchar <2.

Details

This function run a Chinese word segmentation by jiebeR and compute term count from all these text document.

Author(s)

Jim Liu

Examples

```
library(CTM)
a1 <- "hello taiwan"
b1 <- "world of tank"
c1 <- "taiwan weather"
d1 <- "local weather"
text1 <- t(data.frame(a1,b1,c1,d1))
count1 <- termCount(doc = text1, EngTermDeleted=FALSE, shortTermDeleted = FALSE)
```

Index