Package ‘BullsEyeR’

December 21, 2017

Type Package

Title Topic Modelling

Version 0.2.0

Date 2017-12-11

Author Krishna Harsha

Maintainer Krishna Harsha <khkrishnaharsha123@gmail.com>

Depends tm, NLP, topicmodels, Matrix, slam

Description Helps in initial processing like converting text to lower case, removing punctuation, numbers, stop words, stemming, sparsity control and term frequency inverse document frequency processing. Helps in recognizing domain or corpus specific stop words. Makes use of 'ldatunig' output to pick optimal number of topics for topic modelling. Helps in topic modelling the content.

License GPL

Encoding UTF-8

LazyData true

RoxygenNote 6.0.1

NeedsCompilation no

Repository CRAN

Date/Publication 2017-12-21 11:15:41 UTC

R topics documented:

BullsEye *Topic Modelling*

Description

BullsEye runs intial preprocessing, removes custom stop words and runs LDA with selected number of topics.

Usage

```
BullsEye(ds, spvar = 0.99, myStopWords = NULL, tno = 20, seedno = 12345,
  stemvar = 0)
```

Arguments

`ds` a character vector of text documents

`spvar` a sparsity variable which defaults to 0.99

`myStopWords` a character vector of custom stop words which defaults to NULL

`tno` a number of topics to be used to model text using LDA approach which defaults to 20

`seedno` seed which defaults to 12345

`stemvar` a variable indicating stemming to be performed or not which defaults to ’0’ meaning no stemming

Value

A dataframe with index of empty rows and topic terms.

See Also

`FindTopicsNumber`

Examples

```
## Not run:
# Run it and see for yourself

## End(Not run)
data.tmp<-read.csv(system.file("ext", "testdata.csv", package="BullsEyeR"))
ds<-as.character(data.tmp$Story[1:2])
stopwords<-c("sallin","hannah","company","number","started","unlike")
BullsEye(ds=ds,spvar=0.99,myStopWords=stopwords,tno=20,seedno=12345,stemvar=0)
```

BullsEyeR *Topic Modelling for Content curation Cognizant CDB-AIM-BAI-Business Analytics*

Description

This Package provides three categories of important functions: frequency Analysis of word tokens, Creation of Document Term Matrix and Topic Modelling using LDA.

FreqAnalysis()

Frequency Analysis of word tokens - returns dataframe with words and their frequencies after initial preprocessing, sparsity control and TFIDF analysis is performed.we can pick some words from the high frequency list as custom stop words

createDTM()

Creation of Document Term Matrix -repeats first step, now including the custom stop words as well, removes empty documents if any and returns a Document term matrix. This DTM is used for finding optimal number of topics for LDA modelling using 'FindTopicsNumber' from 'ldatuning' package

BullsEye()

Topic Modelling- Performs preprocessing along with removal of custom stop words,Uses topic number selected using 'ldatuning' and builds unigram topic model with/without stemming. Returns,

EmptyRows

A list of zero length documents after preprocessing

Topics

A data frame with top 20 terms in all the topics discovered by LDA.

createDTM *Create Document term Matrix*

Description

The function createDTM creates a document term matrix after preprocessing and removal of stop words.

Usage

```
createDTM(ds, spvar = 0.99, myStopWords = NULL, stemvar = 0)
```

Arguments

`ds` a character vector of text documents

`spvar` a sparsity variable which defaults to 0.99

`myStopWords` a character vector of custom stop words which defaults to NULL

`stemvar` a variable indicating stemming to be performed or not which defaults to ’0’ meaning no stemming

Value

A Document Term Matrix.

Examples

```
## Not run:
# Run it and see for yourself

## End(Not run)
data.tmp<-read.csv(system.file(”ext”, ”testdata.csv”, package=”BullsEyeR”))
ds<-as.character(data.tmp$Story[1:2])
stopwords<-c(”sallin”,”hannah”,”company”,”number”,”started”,”unlike”)
createDTM(ds=ds,spvar=0.99,myStopWords=stopwords,stemvar=0)
```

`freqAnalysis` *Functions Frequency Analysis*

Description

The function freqAnalysis does a frequency analysis of retained words after initial preprocessing.

Usage

```
freqAnalysis(ds, spvar = 0.99, stemvar = 0)
```

Arguments

`ds` a character vector of text documents

`spvar` a sparsity variable which defaults to 0.99

`stemvar` a variable indicating stemming to be performed or not which defaults to ’0’ meaning no stemming

Value

A dataframe with words and their frequencies after listed preprocessing.

Examples

```
## Not run:
# Run it and see for yourself

## End(Not run)
data.tmp<-read.csv(system.file("ext", "testdata.csv", package="BullsEyeR"))
ds<-as.character(data.tmp$Story[1:2])
freqAnalysis(ds=ds,spvar=0.99,stemvar=0)
```

testdata *Sample text data*

Description

A collection of four articles with two columns- Article and Story namely.

Usage

```
testdata
```

Format

A csv file of text data

Index