Package 'ANLP'

July 11, 2016

Type Package

Title Build Text Prediction Model

Version 1.3

Author Achal Shah

Maintainer Achal Shah <achalshah20@gmail.com>

Description Library to sample and clean text data, build N-gram model, Backoff algorithm etc.

License GPL-3

LazyData True

RoxygenNote 5.0.1

Depends tm,qdap,RWeka,dplyr, R (>= 2.10)

Suggests knitr, rmarkdown, testthat

VignetteBuilder knitr

NeedsCompilation no

Repository CRAN

Date/Publication 2016-07-11 08:44:37

R topics documented:

ANLP-package *Build Text Prediction Model*

Description

Library to sample and clean text data, build N-gram model, Backoff algorithm etc.

Author(s)

Achal Shah

Maintainer: Achal Shah <achalshah20@gmail.com>

See Also

~~ sampleTextData generateTDM predict_Backoff ~~

buildNgramModel *Build N gram model*

Description

This function is an abstract function used by generateTDM

Usage

```
buildNgramModel(N)
```

Arguments

N	size of n-gram model

Value

function which can be used to build N-gram model

See Also

NGramTokenizer Weka_control

cleanTextData *Clean and tokenize string data*

Description

This function applies different cleaning techniques to clean corpus data.

Usage

```
cleanTextData(data)
```

Arguments

data	Data read by readTextFile

Details

This function removes non english characters, numbers, white spaces, brackets, punctuation. It also handles cases like abbreviation, contraction. It converts entire text to lower case.

Value

a list having sampled text data

See Also

tm_map iconv content_transformer removeNumbers replace_contraction replace_abbreviation bracketX removePunctuation tolower stripWhitespace

generateTDM *Generate term document frequency table from corpus*

Description

This function builds term documement sparse matrix

Usage

```
generateTDM(data, N, isTrace = F)
```

Arguments

data	It can be text corpus/data cleaned by cleanTextData
N	size of n-gram model
isTrace	for debugging purpose, use this if you want to track time to build model.

Details

This function generates terms with N number of words specified in argument. This can be used in many tasks like information retrival, document similarity etc.

Value

term document matrix for terms having N words

See Also

`TermDocumentMatrix` `buildNgramModel`

---

predict_Backoff *Predict next word using backoff method*

Description

This function predicts next word using back-off algorithm.

Usage

```
predict_Backoff(testline, modelsList, isDebugMode = F)
```

Arguments

`testline`	Line on which we are performing algorithm to predict next word
`modelsList`	List having all Ngram models generated by `generateTDM`
`isDebugMode`	for debugging purpose, this will print out debug statements

Details

This function predicts next word based on previous N number of words using N-gram models generated by `generateTDM`.

Value

next predicted word

See Also

`generateTDM` `TermDocumentMatrix`

readTextFile *Read text files in binary mode*

Description

This function reads text files in the binary mode

Usage

```
readTextFile(fileName, encoding)
```

Arguments

fileName	A path to the text file
encoding	Read text data with encoding

Value

a list having all the text data from file

Examples

```
## Not run: sampleTextData("data/twitter.txt","UTF-8")
```

sampleTextData *Sample text data*

Description

This function reads text files in the binary mode

Usage

```
sampleTextData(data, proportion)
```

Arguments

data	Data read by readTextFile
proportion	Value between 0 to 1 which represents portion of data

Value

a list having sampled text data

See Also

rbinom

`twitter.data` *Twitter dataset*

Description

This dataset is having actual tweets. It contains more than 100k tweets.

Usage

```
twitter.data
```

Format

An object of class `character` of length 109091.

Index