Package 'Rwordseg'

August 23, 2019

License LGPL

Title Chinese Word Segmentation

Type Package

LazyLoad yes

Author Jian Li

Maintainer Jian Li <rweibo@sina.com>

Description Provides interfaces and useful tools for Chinese word segmentation. Implements a segmentation algorithm based on Hidden Markov Model (HMM) in native R codes. Methods for HHMM-Based Chinese lexical analyzer are as described in : Hua-Ping Zhang et al., (2003) <doi:10.3115/1119250.1119280>.

Version 0.3-2

Date 2019-08-21

Depends R (>= 3.0.0), utils, stats, tmcn, HMM

Suggests jiebaR, coreNLP

RoxygenNote 6.1.0

NeedsCompilation no

Repository CRAN

Date/Publication 2019-08-23 09:00:06 UTC

R topics documented:

`createDict` *Create a dictionary file from corpus.*

Description

Read a corpus vector and generate the dictionary data frame.

Usage

```
createDict(trainvec, dicfile = NULL, wordsplit = "\\s+",
  natruesplit = "/")
```

Arguments

`trainvec`	A character vector of corpus.
`dicfile`	The path of output file. Defult is NULL.
`wordsplit`	Character containing regular expression to use for splitting words.
`natruesplit`	Character containing regular expression to use for splitting nature.

Value

A data frame of:

`word`	Word.
`freq`	Frequency.
`nature`	Nature.

Author(s)

Jian Li <<rweibo@sina.com>>

Examples

```
data(PD980105)
d1 <- createDict(PD980105[1:10])
head(d1)
```

createHMM *Create a HMM model from corpus.*

Description

Read a corpus vector and generate a HMM model file.

Usage

```
createHMM(trainvec, outputfolder = NULL, sensplit = "/w",
  wordsplit = "\\s+", natruesplit = "/", removestr = "^.*?/m")
```

Arguments

trainvec	A character vector of corpus.
outputfolder	The folder of output file. Defult is NULL.
sensplit	Character containing regular expression to use for splitting sentence.
wordsplit	Character containing regular expression to use for splitting words.
natruesplit	Character containing regular expression to use for splitting nature.
removestr	Character containing regular expression to use for removing string.

Value

a list from initHMM.

Examples

```
data(PD980105)
m1 <- createHMM(PD980105[1:10])
names(m1)
```

importSogouScel *Import a Sogou dictionary.*

Description

Import a scel file of Sogou dictionary.

Usage

```
importSogouScel(strpaths)
```

Arguments

strpaths	The path of .scel file.

Value

A list of:

desc	A data frame of the description.
dict	A data frame of the dictionary.

Author(s)

Jian Li <<rweibo@sina.com>>

References

https://pinyin.sogou.com/dict/

---

insertWords *Insert new words into analyzer.*

Description

When you restart R, all of the wordes will be removed. If you want to keep them please try installDict.

Usage

```
insertWords(strwords)
```

Arguments

strwords	Vector of words.

Value

No results.

Author(s)

Jian Li <<rweibo@sina.com>>

installDict *Install a new dictionary.*

Description

Install a new dictionary from a Sogou scel file or text file. Make sure the file encoding is in UTF-8.

Usage

```
installDict(dictpath, dictname = "", dictdesc = "")
```

Arguments

dictpath	Path of dictionary.
dictname	Name of the dictionary. Sogou scel file don't need this input.
dictdesc	Description of the dictionary. Default is empty string.

Value

No results.

Author(s)

Jian Li <<rweibo@sina.com>>

listDict *List the installed dictionaries.*

Description

List all of the installed user-defined dictionaries.

Usage

```
listDict()
```

Value

A data frame of:

id	ID of the dictionary.
dict	Name of the dictionary.
time	Created time.
size	Word counts of the dictionary.
example	Example words.
desc	Description of the dictionary.

Author(s)

Jian Li <<rweibo@sina.com>>

Examples

```
listDict()
```

`parseSentence` *Parse a string of text.*

Description

Runs the CoreNLP annotators to parse a string of text.

Usage

```
parseSentence(text)
```

Arguments

`text`	A vector of strings for parsing.

Value

A list of:

`parse`	A data frame of the results of syntactic parsing tree.
`token`	A data frame of the results of word segmentation.
`depenencies`	A data frame of the results of dependency parsing.

Author(s)

Jian Li <<rweibo@sina.com>>

PD980105 *Corpus of Multi-level Processing for Modern Chinese*

Description

Corpus from The People's Daily from 1998-01-01 to 1998-01-05.

Usage

```
data(PD980105)
```

Format

A charactor vector in UTF-8.

References

http://klcl.pku.edu.cn/gxzy/231686.htm

segmentCN *Sengment Chinese text.*

Description

A function to segment Chinese text into words.

Usage

```
segmentCN(strwords, analyzer = c("default", "hmm", "jiebaR", "fmm",
  "coreNLP"), nature = FALSE, nosymbol = TRUE,
  returnType = c("vector", "tm"), ...)
```

Arguments

strwords	A charactor vector of Chinese sentence.
analyzer	One of 'default', 'jiebaR', 'hmm', 'fmm' and 'coreNLP'. Default is 'hmm'.
nature	Whether to recognise the nature of the words.
nosymbol	Whether to keep symbols in the sentence. Default is TRUE, means no symbols kept.
returnType	Default is a string vector but we also can choose 'tm' to output a single string separated by space so that it can be used by Corpus directly.
...	Other arguments.

Value

a vector of words (list if input is vecter) which have been segmented.

Author(s)

Jian Li <<rweibo@sina.com>>

Examples

```
segmentCN("hello world!")
```

`setAnalyzer` *Set the default analyzer.*

Description

The default analyzer is 'hmm', which is implemented by native R codes and still in development. You can use 'jiebaR' instead. Or 'coreNLP' to invoke Stanford CoreNLP. Or choose 'fmm' to try the forward maximum matching algorithm.

Usage

```
setAnalyzer(analyzer = c("hmm", "jiebaR", "fmm", "coreNLP"),
  coreNLPdir = "")
```

Arguments

`analyzer`	One of 'jiebaR', 'hmm', 'fmm' and 'coreNLP'.
`coreNLPdir,`	Set the coreNLP file path, only use for 'coreNLP'.

Value

No results.

Examples

```
setAnalyzer("hmm")
```

setAppDir *Set the application path.*

Description

The directory path of the application folder will contain the dictionaries and setting files. You can set a user-defined folder permanently. We suggest setting the folder of 'APPDATA' environment variable by running 'setAppDir(\"APPDATA\")'.

Usage

```
setAppDir(appdir)
```

Arguments

appdir	The directory path of the application folder. Default is 'tempdir()'.

Value

No results.

uninstallDict *Uninstall a dictionary.*

Description

Uninstall a user-defined dictionary.

Usage

```
uninstallDict(dictid)
```

Arguments

dictid	The ID of the dictionary, which is shown in the result of listDict.

Value

No results.

Author(s)

Jian Li <<rweibo@sina.com>>

Index