Package 'KoNLP'

July 2, 2014

Maintainer Heewon Jeon <madjakarta@gmail.com>

License GPL-3

Title Korean NLP Package

Author Heewon Jeon

Description Korean language processing package. Morphological
analyzer, POS tagger, Keystroke converter, Hangul automata,Concordance, Mutual Information..

SystemRequirements Java (>= 1.6)

URL https://github.com/haven-jeon/KoNLP

Version 0.76.9

Repository CRAN

Date/Publication 2013-05-31 14:23:46

Date 2011-10-12

Encoding UTF-8

Depends R (>= 2.15.0), rJava (>= 0.9-0), utils (>= 2.14.0), stringr
(>= 0.6.2), hash (>= 2.2.6), tau (>= 0.0-15), Sejong (>= 0.01)

Collate 'onLoad.R' 'manageDic.R' 'hangulUtils.R' 'koAnalyzerRun.R' 'tagdata.R' 'Concordances.R'

NeedsCompilation no

R topics documented:

`backupUsrDic` *use for backup current dic_user.txt*

Description

Utility function for backup dic_user.txt file to backup directory.

Usage

```
backupUsrDic(ask = TRUE)
```

Arguments

`ask`	ask to confirm backup

Examples

```
## Not run:
## This codes can not be run
##  if you don't have encoding system which can en/decode
##  Hangul(ex) CP949, EUC-KR, UTF-8).
dicpath <- file.path(system.file(package="Sejong"), "dics", "handic.zip")
conn <- unz(dicpath, file.path("data","kE","dic_user2.txt"))
newdic <- read.table(conn, sep="\t", header=FALSE, fileEncoding="UTF-8", stringsAsFactors=FALSE)
mergeUserDic(newdic)
```

```
## backup merged new dictionary
backupUsrDic(ask=FALSE)
## restore from backup directory
restoreUsrDic(ask=FALSE)
## reloading new dictionary
reloadAllDic()
## End(Not run)
```

---

concordance_file *concordance for input text file*

Description

returns concordance text for input file.

Usage

```
concordance_file(filename, pattern,
  encoding = getOption("encoding"), span = 5)
```

Arguments

filename	file name
pattern	patterns of central words
span	how many character will be produced around input pattern
encoding	filename's encoding

Author(s)

Heewon Jeon

References

Church, K. W. and Mercer, R. L. (1993). Introduction to the special issue on computational linguistics using large corpora. Computational Linguistics, 19(1):1-24.

concordance_str *concordance for input text vector*

Description

returns concordance text for input pattern and span.

Usage

```
concordance_str(string, pattern, span = 5)
```

Arguments

string	input text as character vector or single character
pattern	patterns of central words
span	how many character will be produced around input pattern

Author(s)

Heewon Jeon

References

Church, K. W. and Mercer, R. L. (1993). Introduction to the special issue on computational linguistics using large corpora. Computational Linguistics, 19(1):1-24.

convertHangulStringToJamos *convertion function Hangul string to Jamos*

Description

convert Hangul sentence to Jamos. Example will be shown in github wiki.

Usage

```
convertHangulStringToJamos(hangul)
```

Arguments

hangul	Hangul string

Value

Jamo sequences

convertHangulStringToKeyStrokes *convertion function Hangul string to keyStrokes*

Description

Function can convert Hangul string to Keystrokes. Example will be shown in github wiki.

Usage

```
convertHangulStringToKeyStrokes(hangul,
  isFullwidth = TRUE)
```

Arguments

hangul	Hangul sentence
isFullwidth	specify returned character will be Fullwidth ASCII or Halfwidth ASCII

Value

Keystroke sequence

convertTag *tag name converter*

Description

only suppport tag convertion between KAIST and Sejong tag set.

Usage

```
convertTag(fromTag, toTag, tag)
```

Arguments

fromTag	tag set name to convert from
toTag	desired tag set name
tag	tag name to search

editweights *Keystroke misspell cost table*

Description

Keystroke misspell cost table

Author(s)

Heewon Jeon <madjakarta@gmail..com>

extractNoun *Noun extractor for Hangul*

Description

extract Nouns from Korean sentence uses Hannanum analyzer. see detail in Hannanum. Example will be shown in github wiki.

Usage

```
extractNoun(sentence)
```

Arguments

sentence	input

Value

Noun of sentence

References

Sangwon Park et al(2010). A Plug-In Component-based Korean Morphological Analyzer

HangulAutomata *do Hangul automata*

Description

function to be used for converting to complete Hangul syllables from Jamo or Keystrokes. Example will be shown in github wiki.

Usage

```
HangulAutomata(input, isKeystroke = F, isForceConv = F)
```

Arguments

input	to be processed mostly Jamo sequences
isKeystroke	boolean parameter to check input is keystroke or Jamo sequences
isForceConv	boolean parameter to force converting if input is not valid Jamo or keystroke sequences.

Value

complete Hangul syllable

is.ascii *check if sentence is all ASCII*

Description

Function checks with each charactor is ASCII

Usage

```
is.ascii(sentence)
```

Arguments

sentence	input charactors

Value

TRUE or FALSE

is.hangul *check if sentence is all Hangul*

Description

Function checks if each charactor is Hangul or Jamo. Example will be shown in github wiki.

Usage

```
is.hangul(sentence)
```

Arguments

sentence	input charactors

Value

TRUE or FALSE

is.jaeum *check if sentence is all Jaeum*

Description

Function checks with each charactor is Jaeum

Usage

```
is.jaeum(sentence)
```

Arguments

sentence	input charactors

Value

TRUE or FALSE

is.jamo *check if sentence is all Jamo*

Description

Function checks with each charactor is Jamo. Example will be shown in github wiki.

Usage

```
is.jamo(sentence)
```

Arguments

sentence	input charactors

Value

TRUE or FALSE

is.moeum *check if sentence is all Moeum*

Description

Function checks with each charactor is Moeum

Usage

```
is.moeum(sentence)
```

Arguments

sentence	input charactors

Value

TRUE or FALSE

KtoS *KAIST tag to Sejong tag*

Description

KAIST tag to Sejong tag

Author(s)

Heewon Jeon <madjakarta@gmail..com>

mergeUserDic *appending or replacing with new data.frame*

Description

appending new dictionary to current dictionary. replaceing current dictionary with new dictionary.

Usage

```
mergeUserDic(newUserDic, append = TRUE, verbose = FALSE,
  ask = FALSE)
```

Arguments

newUserDic	new user dictionary as data.frame
append	append or replacing
verbose	see detail error logs
ask	ask to backup

Examples

```
## Not run:
## This codes can not be run
##  if you don't have encoding system which can en/decode
##  Hangul(ex) CP949, EUC-KR, UTF-8).
dicpath <- file.path(system.file(package="Sejong"), "dics", "handic.zip")
conn <- unz(dicpath, file.path("data","kE","dic_user2.txt"))
newdic <- read.table(conn, sep="\t", header=FALSE, fileEncoding="UTF-8", stringsAsFactors=FALSE)
mergeUserDic(newdic)
## backup merged new dictionary
backupUsrDic(ask=FALSE)
## restore from backup directory
restoreUsrDic(ask=FALSE)
## reloading new dictionary
reloadAllDic()
## End(Not run)
```

MorphAnalyzer *Hannanum morphological analyzer interface function*

Description

Do the morphological analysis, not doing pos tagging uses Hannanum analyzer. see details in Hannanum. Example will be shown in github wiki.

Usage

```
MorphAnalyzer(sentence)
```

Arguments

sentence	input

Value

result of analysis

References

Sangwon Park et al(2010). A Plug-In Component-based Korean Morphological Analyzer

mutualinformation *mutual information for input text*

Description

returns mutual information or t-scores for input text

Usage

```
mutualinformation(text, query = "",
  method = c("mutual", "tscores"))
```

Arguments

text	input character vector
method	for calculations('mutual' or 't-scores')
query	term to get information

Author(s)

Heewon Jeon

References

Church, K. W. and Hanks, P. (1990). Word association norms, mutual information,and lexicography.Computational Linguistics, 16(1):22-29.

Church, K. W. and Mercer, R. L. (1993). Introduction to the special issue on computational linguistics using large corpora. Computational Linguistics, 19(1):1-24.

`reloadAllDic` *reload all Hannanum analyzer dictionary*

Description

Mainly, user dictionary reloading for Hannanum Analyzer. If you want to update user dictionary on KoNLP_dic/current/dic_user.txt, need to execute this function after editing dictionary.

Usage

```
reloadAllDic()
```

Examples

```
## Not run:
## This codes can not be run
##  if you don't have encoding system which can en/decode
##  Hangul(ex) CP949, EUC-KR, UTF-8).
dicpath <- file.path(system.file(package="Sejong"), "dics", "handic.zip")
conn <- unz(dicpath, file.path("data","kE","dic_user2.txt"))
newdic <- read.table(conn, sep="\t", header=FALSE, fileEncoding="UTF-8", stringsAsFactors=FALSE)
mergeUserDic(newdic)
## backup merged new dictionary
backupUsrDic(ask=FALSE)
## restore from backup directory
restoreUsrDic(ask=FALSE)
## reloading new dictionary
reloadAllDic()
## End(Not run)
```

`restoreUsrDic` *use for restoring backuped dic_user.txt*

Description

Utility function for restoring dic_user.txt file to dictionary directory.

Usage

```
restoreUsrDic(ask = TRUE)
```

Arguments

ask	ask to confirm backup

Examples

```
## Not run:
## This codes can not be run
##  if you don't have encoding system which can en/decode
##  Hangul(ex) CP949, EUC-KR, UTF-8).
dicpath <- file.path(system.file(package="Sejong"), "dics", "handic.zip")
conn <- unz(dicpath, file.path("data","kE","dic_user2.txt"))
newdic <- read.table(conn, sep="\t", header=FALSE, fileEncoding="UTF-8", stringsAsFactors=FALSE)
mergeUserDic(newdic)
## backup merged new dictionary
backupUsrDic(ask=FALSE)
## restore from backup directory
restoreUsrDic(ask=FALSE)
## reloading new dictionary
reloadAllDic()
## End(Not run)
```

SimplePos09 *POS tagging by using 9 KAIST tags*

Description

Do pos tagging using 9 tags uses Hannanum analyzer. see details in Hannanum. Example will be shown in github wiki.

Usage

```
SimplePos09(sentence)
```

Arguments

sentence	input

Value

results of tagged analysis

References

Sangwon Park et al(2010). A Plug-In Component-based Korean Morphological Analyzer

SimplePos22 *POS tagging by using 22 KAIST tags*

Description

Do POS tagging using 22 tags uses Hannanum analyzer. see details in Hannanum. Example will be shown in github wiki.

Usage

```
SimplePos22(sentence)
```

Arguments

sentence	input

Value

results of tagged analysis

References

Sangwon Park et al(2010). A Plug-In Component-based Korean Morphological Analyzer

statDic *summary of dictionaries*

Description

show summary, head and tail of current or backup dictionaries

Usage

```
statDic(which = "current", n = 6)
```

Arguments

which	"current" or "backup" dictionary
n	a single integer. Size for the resulting object to view

Examples

```
## show current dictionary's summary, head, tail
statDic("current", 10)
```

StoK *Sejong tag to KAIST tag*

Description

Sejong tag to KAIST tag

Author(s)

Heewon Jeon <madjakarta@gmail..com>

tags *tag names*

Description

tag names

Author(s)

Heewon Jeon <madjakarta@gmail..com>

useSejongDic *use Sejong noun dictionary*

Description

Retrive Sejong dictionary to use in KoNLP

Usage

```
useSejongDic(backup = T)
```

Arguments

backup	will backup current dictionary?

References

http://www.sejong.or.kr/

useSystemDic *use system default dictionary*

Description

Retrive system default dictionary to use in KoNLP

Usage

```
useSystemDic(backup = T)
```

Arguments

backup	will backup current dictionary?

Index