Package 'moranajp'

July 12, 2022

Title Morphological Analysis for Japanese

Version 0.9.5

Description Supports morphological analysis for Japanese by using 'MeCab'. Can input data.frame and obtain all results of 'MeCab' and row number of original data.frame as a text id.

License MIT + file LICENSE

Depends R (>= 3.5.0)

URL https://github.com/matutosi/moranajp

https://github.com/matutosi/moranajp/tree/develop (devel)

BugReports https://github.com/matutosi/moranajp/issues

Imports dplyr, ggplot2, ggraph, igraph, magrittr, purrr, rlang, stats, stringr, tibble, tidyr

Suggests knitr, rmarkdown, stringi, testthat (>= 3.0.0), tidyverse

VignetteBuilder knitr

Config/testthat/edition 3

Encoding UTF-8

LazyData true

RoxygenNote 7.2.0

NeedsCompilation no

Author Toshikazu Matsumura [aut, cre]

Maintainer Toshikazu Matsumura <matutosi@gmail.com>

Repository CRAN

Date/Publication 2022-07-12 02:20:02 UTC

R topics documented:

`add_series_no` *Add series no col according to match condition.*

Description

Internal function for moranajp_all(). 'EOS' means breaks of text in this package (and most of morphological analysis). add_text_id() add `text_id` column when there is 'EOS'.

Usage

```
add_series_no(tbl, cond = "", end_sep = TRUE, new_col = "series_no")

add_text_id(tbl)
```

Arguments

`tbl`	A tibble or data.frame.
`cond`	Condition to split series no.
`end_sep`	A logical. TRUE: condition indicate the end of separation.
`new_col`	A string name of new column.

Value

`A tibble, which include new_col as series no.`

A tibble.

Examples

```
## Not run:
  tbl <- tibble::tibble(col=c(rep("a", 2), "sep", rep("b", 3), "sep", rep("c", 4), "sep"))
  cond <- ".$col == 'sep'"   # Use ".$'colname'" to identify column
    # when separator indicate the end
  add_series_no(tbl, cond = cond, end_sep = TRUE,  new_col = "series_no")
    # when separator indicate the begining
  add_series_no(tbl, cond = cond, end_sep = FALSE, new_col = "series_no")

## End(Not run)
```

---

clean_up *Clean up result of morphological analyzed data frame*

Description

Clean up result of morphological analyzed data frame

Usage

```
clean_mecab_local(df, ...)

clean_chamame(df, ...)

pos_filter_mecab_local(df)

pos_filter_chamame(df)

delete_stop_words(df, use_common_data = TRUE, add_stop_words = NULL, ...)

replace_words(
  df,
  synonym_df = NULL,
  synonym_from = NULL,
  synonym_to = NULL,
  ...
)
```

Arguments

df	A dataframe including result of morphological analysis.
...	Extra arguments to internal fuctions.
use_common_data	A logical. TRUE: use data(stop_words).
add_stop_words	A string vector adding into stop words. When use_common_data is TRUE and add_stop_words are given, both of them will be used as stop_words.

synonym_df A datarame including synonym word pairs. The first column: replace from, the second: replace to.

synonym_from, synonym_to
A string vector. Length of synonym_from and synonym_to should be the same. When synonym_df and synonym pairs (synonym_from and synonym_to) are given, both of them will be used as synonym.

Value

A dataframe.

Examples

```
data(neko_mecab)
data(synonym)
synonym <-
  synonym %>% dplyr::mutate_all(stringi::stri_unescape_unicode)

neko_mecab %>%
  dplyr::select(-text_id) %>%
  dplyr::mutate_all(stringi::stri_unescape_unicode) %>%
  magrittr::set_colnames(stringi::stri_unescape_unicode(colnames(.))) %>%
  clean_mecab_local(
    use_common_data = TRUE,
    synonym_df = synonym)
```

---

draw_bigram_network *Draw bigram network using morphological analysis data.*

Description

Draw bigram network using morphological analysis data.

Usage

```
draw_bigram_network(df, ...)

bigram(df, text_id = "text_id", ...)

bigram_net(bigram, rand_seed = 12, threshold = 100, ...)

word_freq(df, bigram_net)

bigram_network_plot(
  bigram_net,
  freq,
  ...,
```

```
  arrow_size = 5,
  circle_size = 5,
  text_size = 5,
  font_family = "",
  arrow_col = "darkgreen",
  circle_col = "skyblue",
  x_limits = NULL,
  y_limits = NULL,
  no_scale = FALSE
)
```

Arguments

df	A dataframe including result of morphological analysis.
...	Extra arguments to internal fuctions.
text_id	A dstring to specify text.
bigram	A result of bigram().
rand_seed	A numeric.
threshold	A numeric used as threshold for frequency of bigram.
bigram_net	A result of bigram_net().
freq	A numeric of word frequency in bigram_net. Can be got using word_freq().
arrow_size, circle_size, text_size,	A numeric.
font_family	A string.
arrow_col, circle_col	A string to specicy arrow and circle color in bigram network.
x_limits, y_limits	A Pair of numeric to specify range.
no_scale	A logical. FALSE: Not draw x and y axis.

Value

A gg object of bigram network plot.

Examples

```
library(tidyverse)
data(neko_mecab)
data(synonym)
synonym <-
  synonym %>% dplyr::mutate_all(stringi::stri_unescape_unicode)

bigram_neko <-
  neko_mecab %>%
  dplyr::select(-text_id) %>%
  dplyr::mutate_all(stringi::stri_unescape_unicode) %>%
  magrittr::set_colnames(stringi::stri_unescape_unicode(colnames(.))) %>%
```

```
  clean_mecab_local(
    use_common_data = TRUE,
    synonym_df = synonym) %>%
  draw_bigram_network()

add_stop_words <-
  c("\\u3042\\u308b", "\\u3059\\u308b", "\\u3066\\u308b",
    "\\u3044\\u308b","\\u306e", "\\u306a\\u308b", "\\u304a\\u308b",
    "\\u3093", "\\u308c\\u308b", "*") %>%
  stringi::stri_unescape_unicode()

bigram_review <-
  review_chamame %>%
  dplyr::mutate_all(stringi::stri_unescape_unicode) %>%
  magrittr::set_colnames(stringi::stri_unescape_unicode(colnames(.))) %>%
  clean_chamame(add_stop_words = add_stop_words) %>%
  draw_bigram_network()
```

---

make_groups *Make groups by splitting string length*

Description

Using 'MeCab' for morphological analysis. Keep other colnames in dataframe.

Usage

```
make_groups(
  tbl,
  text_col = "text",
  length = 8000,
  group = "tmp_group",
  str_length = "str_length"
)

make_groups_sub(tbl, text_col, n_group, group, str_length)

max_sum_str_length(tbl, group, str_length)
```

Arguments

tbl	A tibble or data.frame.
text_col	A text. Colnames for morphological analysis.
length	A numeric.
group, str_length	A string to use temporary.
n_group	A numeric.

Value

A tibble. Output of 'MeCab' and added column "text_id".

Examples

```
## Not run:
  library(tidyverse)
  data(neko)
  neko <-
      neko %>%
      dplyr::mutate(text=stringi::stri_unescape_unicode(text)) %>%
      dplyr::mutate(cols=1:nrow(.))
  bin_dir <- "d:/pf/mecab/bin"
  moranajp_all(neko, text_col = "text", bin_dir = bin_dir) %>%
      print(n=100)

## End(Not run)
```

---

moranajp_all *Morphological analysis for a specific column in dataframe*

Description

Using 'MeCab' for morphological analysis. Keep other colnames in dataframe.

Usage

```
moranajp_all(
  tbl,
  bin_dir,
  text_col = "text",
  option = "",
  iconv = "CP932_UTF-8"
)

moranajp(tbl, bin_dir, option = "", iconv = "")

make_cmd_mecab(tbl, bin_dir, option = "")

out_cols_mecab()

mecab_all(tbl, text_col = "text", bin_dir = "")

mecab(tbl, bin_dir)
```

Arguments

tbl	A tibble or data.frame.
bin_dir	A text. Directory of mecab.
text_col	A text. Colnames for morphological analysis.
option	A text. Options for mecab. "-b" option is already set by moranajp. See by "mecab -h".
iconv	A text. Convert encoding of MeCab output. Default (""): don't convert. "CP932_UTF-8": iconv(output, from = "Shift-JIS" to = "UTF-8") "EUC_UTF-8" : iconv(output, from = "eucjp", to = "UTF-8")

Value

A tibble. Output of 'MeCab' and added column "text_id".

Examples

```
## Not run:
  library(tidyverse)
  data(neko)
  neko <-
      neko %>%
      dplyr::mutate(text=stringi::stri_unescape_unicode(text)) %>%
      dplyr::mutate(cols=1:nrow(.))
  bin_dir <- "d:/pf/mecab/bin"
  moranajp_all(neko, text_col = "text", bin_dir = bin_dir) %>%
      print(n=100)

## End(Not run)
```

---

neko *The first part of 'I Am a Cat' by Soseki Natsume*

Description

The first part of 'I Am a Cat' by Soseki Natsume

Usage

```
neko
```

Format

A data frame with 9 rows and 1 variable:

text Body text. Escaped by stringi::stri_escape_unicode().

Examples

```
data(neko)
neko %>%
  dplyr::mutate_all(stringi::stri_unescape_unicode)
```

---

neko_chamame *Analyzed data of neko by chamame*

Description

chamame: https://chamame.ninjal.ac.jp/index.html

Usage

```
neko_chamame
```

Format

A data frame with 2965 rows and 14 variable: (column names are escaped by stringi::stri_escape_unicode(), stringi::stri_unescape_unicode() will show Japanese)

\u8f9e\u66f8 result of chamame

\u6587\u5883\u754c result of chamame

\u66f8\u5b57\u5f62\uff08\uff1d\u8868\u5c64\u5f62\uff09 result of chamame

\u8a9e\u5f59\u7d20 result of chamame

\u8a9e\u5f59\u7d20\u8aad\u307f result of chamame

\u54c1\u8a5e result of chamame

\u6d3b\u7528\u578b result of chamame

\u6d3b\u7528\u5f62 result of chamame

\u767a\u97f3\u5f62\u51fa\u73fe\u5f62 result of chamame

\u4eee\u540d\u5f62\u51fa\u73fe\u5f62 result of chamame

\u8a9e\u7a2e result of chamame

\u66f8\u5b57\u5f62(\u57fa\u672c\u5f62) result of chamame

\u8a9e\u5f62(\u57fa\u672c\u5f62) result of chamame

...14 result of chamame

Examples

```
data(neko_chamame)
neko_chamame %>%
  dplyr::mutate_all(stringi::stri_unescape_unicode) %>%
  magrittr::set_colnames(stringi::stri_unescape_unicode(colnames(.)))
```

neko_mecab *Analyzed data of neko by MeCab*

Description

MeCab: https://taku910.github.io/mecab/

Usage

```
neko_mecab
```

Format

A data frame with 2893 rows and 11 variable: (column names are escaped by stringi::stri_escape_unicode(), stringi::stri_unescape_unicode() will show Japanese)

text_id result of Mecab

\u8868\u5c64\u5f62 result of Mecab

\u54c1\u8a5e result of Mecab

\u54c1\u8a5e\u7d30\u5206\u985e1 result of Mecab

\u54c1\u8a5e\u7d30\u5206\u985e2 result of Mecab

\u54c1\u8a5e\u7d30\u5206\u985e3 result of Mecab

\u6d3b\u7528\u578b result of Mecab

\u6d3b\u7528\u5f62 result of Mecab

\u539f\u5f62 result of Mecab

\u8aad\u307f result of Mecab

\u767a\u97f3 result of Mecab

Examples

```
data(neko_mecab)
neko_mecab %>%
  dplyr::mutate_all(stringi::stri_unescape_unicode) %>%
  magrittr::set_colnames(stringi::stri_unescape_unicode(colnames(.)))
```

review *Full text of review article*

Description

Full text of review article

Usage

```
review
```

Format

A data frame with 457 rows and 2 variables:

text Body text. Escaped by stringi::stri_escape_unicode(). Citation is as below. Matsumura et al. 2014. Conditions and conservation for biodiversity of the semi-natural grassland vegetation on rice paddy levees. Vegetation Science, 31, 193-218. doi = 10.15031/vegsci.31.193 https://www.jstage.jst.go.jp/article/vegsci/31/2/31_193/_article/-char/en

chap Dammy number of chapter.

data(neko) review %>% dplyr::mutate_all(stringi::stri_unescape_unicode)

review_chamame *Analyzed data of review by chamame*

Description

chamame: https://chamame.ninjal.ac.jp/index.html

Usage

```
review_chamame
```

Format

A data frame with 21013 rows and 14 variable (column names are escaped by stringi::stri_escape_unicode(), stringi::stri_unescape_unicode() will show Japanese)

\u8f9e\u66f8 result of chamame

\u6587\u5883\u754c result of chamame

\u66f8\u5b57\u5f62\uff08\uff1d\u8868\u5c64\u5f62\uff09 result of chamame

\u8a9e\u5f59\u7d20 result of chamame

\u8a9e\u5f59\u7d20\u8aad\u307f result of chamame

\u54c1\u8a5e result of chamame

\u6d3b\u7528\u578b result of chamame

\u6d3b\u7528\u5f62 result of chamame

\u767a\u97f3\u5f62\u51fa\u73fe\u5f62 result of chamame

\u4eee\u540d\u5f62\u51fa\u73fe\u5f62 result of chamame

\u8a9e\u7a2e result of chamame

\u66f8\u5b57\u5f62(\u57fa\u672c\u5f62) result of chamame

\u8a9e\u5f62(\u57fa\u672c\u5f62) result of chamame

...14 result of chamame

Examples

```
data(review_chamame)
review_chamame %>%
  dplyr::mutate_all(stringi::stri_unescape_unicode) %>%
  magrittr::set_colnames(stringi::stri_unescape_unicode(colnames(.)))
```

---

review_mecab *Analyzed data of neko by MeCab*

Description

MeCab: https://taku910.github.io/mecab/

Usage

```
review_mecab
```

Format

A data frame with 20523 rows and 11 variable: (column names are escaped by stringi::stri_escape_unicode(), stringi::stri_unescape_unicode() will show Japanese)

text_id result of Mecab

\u8868\u5c64\u5f62 result of Mecab

\u54c1\u8a5e result of Mecab

\u54c1\u8a5e\u7d30\u5206\u985e1 result of Mecab

\u54c1\u8a5e\u7d30\u5206\u985e2 result of Mecab

\u54c1\u8a5e\u7d30\u5206\u985e3 result of Mecab

\u6d3b\u7528\u578b result of Mecab

\u6d3b\u7528\u5f62 result of Mecab

\u539f\u5f62 result of Mecab

\u8aad\u307f result of Mecab

\u767a\u97f3 result of Mecab

Examples

```
data(review_mecab)
review_mecab %>%
  dplyr::mutate_all(stringi::stri_unescape_unicode) %>%
  magrittr::set_colnames(stringi::stri_unescape_unicode(colnames(.)))
```

---

stop_words *Stop words for morphological analysis*

Description

Stop words for morphological analysis

Usage

```
stop_words
```

Format

A data frame with 310 rows and 1 variable:

stop_word Stop words can be used with delete_stop_words(). Escaped by stringi::stri_escape_unicode(). Downloaded from http://svn.sourceforge.jp/svnroot/slothlib/CSharp/Version1/SlothLib/NLP/Filter/StopWord/word/Jap

Examples

```
data(stop_words)
stop_words %>%
  dplyr::mutate_all(stringi::stri_unescape_unicode)
```

---

synonym *An example of synonym word pairs*

Description

An example of synonym word pairs

Usage

```
synonym
```

Format

A data frame with 25 rows and 2 variables:

from Words to be replaced from. Escaped by stringi::stri_escape_unicode().

to Words to be replaced to.

Examples

```
data(synonym)
synonym %>%
  dplyr::mutate_all(stringi::stri_unescape_unicode)
```

`text_id_with_break` *Add ids.*

Description

Add ids.

Usage

```
text_id_with_break(x, brk, end_with_brk = TRUE)

add_text_id_df(df, col, brk, end_with_brk = TRUE)
```

Arguments

`x`	A string vector.
`brk`	A string to specify the break between ids.
`end_with_brk`	A logical. TRUE: brk means the end of groups. FALSE: brk means the begining of groups.
`df`	A dataframe.
`col`	A string to specify the column.

Value

id_with_break() returns id vector, add_id_df() returns dataframe.

Examples

```
tmp <- c("a", "brk", "b", "brk", "c")
brk <- "brk"
text_id_with_break(tmp, brk)
add_text_id_df(tibble::tibble(tmp), col = "tmp", "brk")
```

Index