Package 'N2H4'

July 15, 2022

Type Package

Title Handling Methods for Naver News Text Crawling

Version 0.6.5

Date 2022-07-15

Description Provides some functions to get Korean text sample from news articles in Naver which is popular news portal service <https://news.naver.com/> in Korea.

License MIT + file LICENSE

URL https://github.com/forkonlp/N2H4

BugReports https://github.com/forkonlp/N2H4/issues

RoxygenNote 7.1.2

Depends R (>= 3.5.0)

Encoding UTF-8

Suggests testthat

Imports rvest, httr, jsonlite, tibble, dplyr, urltools

NeedsCompilation no

Author Chanyub Park [aut, cre] (<https://orcid.org/0000-0001-6474-2570>)

Maintainer Chanyub Park <mrchypark@gmail.com>

Repository CRAN

Date/Publication 2022-07-15 08:00:02 UTC

R topics documented:

`getAllComment` *Get All Comment*

Description

Get all comments from the provided news article url on naver

Usage

```
getAllComment(turl = url, ...)
```

Arguments

`turl`	character. News article on 'Naver' such as <http://news.naver.com/main/read.nhn?mode=LSD&mid=shm News articl url that is not on Naver.com domain will generate an error.
`...`	parameter in getComment function.

Details

Works just like getComment, but this function executed in a fashion where it finds and extracts all comments from the given url.

Value

a [tibble][tibble::tibble-package]

Examples

```
## Not run:
  getAllComment("https://n.news.naver.com/mnews/article/214/0001195110?sid=103")

## End(Not run)
```

getComment *Get Comment*

Description

Get naver news comments if you want to get data only comment, enter command like below. getComment(url)$result$commentList[[1]]

Usage

```
getComment(
  turl = url,
  pageSize = 10,
  page = 1,
  sort = c("favorite", "reply", "old", "new", "best"),
  type = c("df", "list")
)
```

Arguments

turl	like <https://news.naver.com/main/read.nhn?mode=LSD&mid=shm&sid1=100&oid=056&aid=0010335
pageSize	is a number of comments per page. defult is 10. max is 100.
page	is defult is 1.
sort	you can select favorite, reply, old, new. favorite is defult.
type	type return df or list. Defult is df. df return part of data not all.

Value

a [tibble][tibble::tibble-package]

Examples

```
## Not run:
  getComment("https://n.news.naver.com/mnews/article/421/0002484966?sid=100")

## End(Not run)
```

getContent *Get Content*

Description

Get naver news content from links.

Usage

```
getContent(
  turl,
 col = c("url", "section", "datetime", "edittime", "press", "title", "body", "value")
)
```

Arguments

turl	is naver news link.
col	is what you want to get from news. Defualt is all.

Value

a [tibble][tibble::tibble-package]

Examples

```
## Not run:
  getContent("https://n.news.naver.com/mnews/article/214/0001195110?sid=103")

## End(Not run)
```

getContentBody *Get Content body name.*

Description

Get naver news body from link.

Usage

```
getContentBody(html_obj, body_node_info = "div#dic_area", body_attr = "")
```

Arguments

html_obj	"xml_document" "xml_node" using read_html function.
body_node_info	Information about node names like tag with class or id. Default is "div.article_info h3" for naver news title.
body_attr	if you want to get attribution text, please write down here.

Value

Get character body content.

Examples

```
## Not run:
 hobj <- rvest::read_html("https://n.news.naver.com/mnews/article/214/0001195110?sid=103")
  getContentBody(hobj)

## End(Not run)
```

`getContentDatetime` *Get Content datetime*

Description

Get naver news published datetime from link.

Usage

```
getContentDatetime(
  html_obj,
  datetime_node_info = "span._ARTICLE_DATE_TIME",
  datetime_attr = "data-date-time"
)
```

Arguments

`html_obj`	"xml_document" "xml_node" using read_html function.
`datetime_node_info`	Information about node names like tag with class or id. Default is "div.article_info h3" for naver news title.
`datetime_attr`	if you want to get attribution text, please write down here.

Value

Get POSIXlt type datetime.

Examples

```
## Not run:
 hobj <- rvest::read_html("https://n.news.naver.com/mnews/article/214/0001195110?sid=103")
  getContentDatetime(hobj)

## End(Not run)
```

getContentEditDatetime *Get Content Edit datetime*

Description

Get naver news edited datetime from link.

Usage

```
getContentEditDatetime(
  html_obj,
  datetime_node_info = "span._ARTICLE_MODIFY_DATE_TIME",
  datetime_attr = "data-modify-date-time"
)
```

Arguments

html_obj	"xml_document" "xml_node" using read_html function.
datetime_node_info	Information about node names like tag with class or id. Default is "div.article_info h3" for naver news title.
datetime_attr	if you want to get attribution text, please write down here.

Value

Get POSIXlt type datetime.

Examples

```
## Not run:
 hobj <- rvest::read_html("https://n.news.naver.com/mnews/article/214/0001195110?sid=103")
  getContentEditDatetime(hobj)

## End(Not run)
```

getContentPress *Get Content Press name.*

Description

Get naver news press name from link.

Usage

```
getContentPress(
  html_obj,
  press_node_info = "div.media_end_head_top a img",
  press_attr = "title"
)
```

Arguments

html_obj	"xml_document" "xml_node" using read_html function.
press_node_info	Information about node names like tag with class or id. Default is "div.article_info h3" for naver news title.
press_attr	if you want to get attribution text, please write down here. Defalt is "title".

Value

Get character press.

Examples

```
## Not run:
 hobj <- rvest::read_html("https://n.news.naver.com/mnews/article/214/0001195110?sid=103")
  getContentPress(hobj)

## End(Not run)
```

---

getContentTitle *Get Content Title*

Description

Get naver news Title from link.

Usage

```
getContentTitle(
  html_obj,
  title_node_info = "h2.media_end_head_headline",
  title_attr = ""
)
```

Arguments

html_obj	"xml_document" "xml_node" using read_html function.
title_node_info	Information about node names like tag with class or id. Default is "div.article_info h3" for naver news title.
title_attr	if you want to get attribution text, please write down here.

Value

Get character title.

Examples

```
## Not run:
 hobj <- rvest::read_html("https://n.news.naver.com/mnews/article/214/0001195110?sid=103")
  getContentTitle(hobj)

## End(Not run)
```

`getLike` *Get like Count*

Description

Get naver news like Count

Usage

```
getLike(turl = url)
```

Arguments

`turl`	like <https://news.naver.com/main/read.nhn?mode=LSD&mid=shm&sid1=100&oid=056&aid=0010335

Value

a [tibble][tibble::tibble-package]

Examples

```
## Not run:
  getLike("https://n.news.naver.com/mnews/article/214/0001195110?sid=103")

## End(Not run)
```

getMainCategory *Get News Main Categories*

Description

Get naver news main category names and ids recently.

Usage

```
getMainCategory()
```

Value

a [tibble][tibble::tibble-package]

Examples

```
## Not run:
  getMainCategory()

## End(Not run)
```

getMaxPageNum *Get Max Page Number*

Description

Get Max Page Number

Usage

```
getMaxPageNum(turl = url, max = 100)
```

Arguments

turl	is target url include sid1, sid2, date like below. <http://news.naver.com/main/list.nhn?sid2=265&sid1=100
max	is also interval to try max page number is numeric. Default is 100.

Value

Get numeric

Examples

```
## Not run:
 getMaxPageNum("https://news.naver.com/main/list.naver?mode=LS2D&mid=shm&sid1=103&sid2=376")

## End(Not run)
```

getSubCategory *Get News Sub Categories*

Description

Get naver news sub category names and urls recently.

Usage

```
getSubCategory(sid1 = 100, onlySid2 = TRUE)
```

Arguments

sid1	Main category id in naver news url. Only 1 value is passible. Default is 100 means Politics.
onlySid2	sid2 is sub category id. some sub categories don't have id. If TRUE, functions return data.frame(chr:sub_cate_naem, char:sid2). Defaults is TRUE.

Value

a [tibble][tibble::tibble-package]

Examples

```
## Not run:
  getSubCategory(100)
  getSubCategory(100, FALSE)

## End(Not run)
```

getUrlList *Get Url List By Category*

Description

Get naver news titles and links from target url.

Usage

```
getUrlList(turl = url, col = c("titles", "links"))
```

Arguments

turl	is target url naver news.
col	is what you want to get from news. Defualt is all.

Value

a [tibble][tibble::tibble-package]

Examples

```
 ## Not run:
 getUrlList("https://news.naver.com/main/list.naver?mode=LS2D&mid=shm&sid1=103&sid2=376")

## End(Not run)
```

`setUrls` *Set url for crawling*

Description

Set naver news links with sid, date, etc. sid1, sid2, page can use vectors. sid1, sid2, start Date, end Date is requred.

Usage

```
setUrls(
  sid1_vec,
  sid2_vec,
  strDate,
  endDate,
  page_vec = NA,
  return_type = c("list", "df")
)
```

Arguments

`sid1_vec`	is news code in naver news url
`sid2_vec`	is news code in naver news url.
`strDate`	target date of start.
`endDate`	target date of end.
`page_vec`	pageNum default is NA.
`return_type`	list or data.frame. default is list.

Value

Get data.frame(sid1,sid2,date,pageNum,pageUrl) or list(sid1,sid2,date,pageNum,pageUrl)

Examples

```
setUrls(105, 227, "20180101", "20180102")
```

Index