Package 'N2H4'

March 20, 2020

Type Package

Title Handling Methods for Naver News Text Crawling

Version 0.5.7

Date 2020-03-20

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 http://github.com/forkonlp/N2H4

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

RoxygenNote 7.0.2

Depends R (>= 3.2)

Encoding UTF-8

LazyData true

Suggests testthat

Imports xml2, rvest, httr, jsonlite, tidyr, tibble, dplyr, lubridate

NeedsCompilation no

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

Maintainer Chanyub Park <mrchypark@gmail.com>

Repository CRAN

Date/Publication 2020-03-20 11:10:02 UTC

R topics documented:

`cate_list_url_ex` *Category Page Url Example*

Description

Category Page Url Example

Usage

```
cate_list_url_ex
```

Format

example url character

`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

```
print(news_url_ex)
getAllComment(news_url_ex)
```

---

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

```
print(news_url_ex)
getComment(news_url_ex)
```

`getContent` *Get Content*

Description

Get naver news content from links.

Usage

```
getContent(
  turl,
  col = c("url", "datetime", "edittime", "press", "title", "body"),
  try_cnt = 3
)
```

Arguments

`turl`	is naver news link.
`col`	is what you want to get from news. Defualt is all.
`try_cnt`	is how many you want to try again if error. Default is 3.

Value

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

Examples

```
print(news_url_ex)
getContent(news_url_ex)
```

`getContentBody` *Get Content body name.*

Description

Get naver news body from link.

Usage

```
getContentBody(
  html_obj,
  body_node_info = "div#articleBodyContents",
  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

```
print(news_url_ex)
hobj <- xml2::read_html(news_url_ex)
getContentBody(hobj)
```

`getContentDatetime` *Get Content datetime*

Description

Get naver news published datetime from link.

Usage

```
getContentDatetime(
  html_obj,
  datetime_node_info = "span.t11",
  datetime_attr = "",
  getEdittime = TRUE
)
```

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.
getEdittime	if TRUE, can get POSIXlt type datetime length 2 means published time and final edited time. if FALSE, get Date length 1.

Value

Get POSIXlt type datetime.

Examples

```
print(news_url_ex)
hobj <- xml2::read_html(news_url_ex)
getContentDatetime(hobj)
```

---

getContentPress *Get Content Press name.*

Description

Get naver news press name from link.

Usage

```
getContentPress(
  html_obj,
  press_node_info = "div.article_header div 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

```
print(news_url_ex)
hobj <- xml2::read_html(news_url_ex)
getContentPress(hobj)
```

---

getContentTitle *Get Content Title*

Description

Get naver news Title from link.

Usage

```
getContentTitle(
  html_obj,
  title_node_info = "div.article_info h3",
  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

```
print(news_url_ex)
hobj <- xml2::read_html(news_url_ex)
getContentTitle(hobj)
```

`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

```
print(news_url_ex)
getLike(news_url_ex)
```

`getMainCategory` *Get News Main Categories*

Description

Get naver news main category names and ids recently.

Usage

```
getMainCategory()
```

Value

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

Examples

```
getMainCategory()
```

`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

```
print(cate_list_url_ex)
getMaxPageNum(cate_list_url_ex)
```

`getNewsTrend` *Get Query news trend by date.*

Description

Get number of query volume in naver news. Params depend on getQueryUrl function.

Usage

```
getNewsTrend(query, startDate, endDate)
```

Arguments

query	requred.
startDate	requred form YYYY-MM-DD.
endDate	requred form YYYY-MM-DD.

Value

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

Examples

```
getNewsTrend("endgame", "2019-03-03", "2019-03-04")
```

---

getQueryUrl *Get Query page url*

Description

Get naver news query page url without pageNum.

Usage

```
getQueryUrl(
  query,
  startDate = as.Date(Sys.time()) - 3,
  endDate = as.Date(Sys.time())
)
```

Arguments

query	requred.
startDate	Dfault is 3 days before today.
endDate	Default is today.

Value

url.

Examples

```
getQueryUrl("endgame")
```

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

```
getSubCategory(100)
getSubCategory(100, FALSE)
```

getUrlListByCategory *Get Url List By Category*

Description

Get naver news titles and links from target url.

Usage

```
getUrlListByCategory(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

```
print(cate_list_url_ex)
getUrlListByCategory(cate_list_url_ex)
```

`getUrlListByQuery` *Get Url List By Query*

Description

Get naver news(only not other sites links) titles and links from target url.

Usage

```
getUrlListByQuery(turl = url)
```

Arguments

`turl`	is target url naver news.

Value

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

Examples

```
print(query_list_url_ex)
getUrlListByQuery(query_list_url_ex)
```

getVideoUrl *Get video clip download url in news*

Description

Get naver news video url

Usage

```
getVideoUrl(turl = url)
```

Arguments

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

Value

Get character url.

Examples

```
print(video_url_ex)
getVideoUrl(video_url_ex)
```

news_url_ex *News Url Example*

Description

News Url Example

Usage

```
news_url_ex
```

Format

example url character

query_list_url_ex *Query Result Page Url Example*

Description

Query Result Page Url Example

Usage

```
query_list_url_ex
```

Format

example url character

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")
```

`video_url_ex` *Video Url Example*

Description

Video Url Example

Usage

```
video_url_ex
```

Format

example url character

Index