Package 'Rcrawler'

November 8, 2016

Type Package

Title Crawl a website and extract its contents

Version 0.1

Date 2016-10-1

Author Salim KHALIL

Maintainer Salim KHALIL <khalilsalim1@gmail.com>

Description Rcrawler supports data collection of web content under R environment. It is designed to crawl, parse, store web pages and produce data that can be directly used for web content mining applications.

License What license is it under?

LazyData TRUE

Depends httr, rJava, xml2, data.table, foreach, doParallel, parallel, data.table

R topics documented:

`contentscraper` *contentscraper*

Description

A function for extracting content matching a specific XPath patterns from a web page, the function takes a web page text as _character_ and a set of named patterns then it return a list contining content matching patterns

Usage

```
contentscraper(webpage, patterns, patnames, excludepat, astext = TRUE, encod)
```

Arguments

webpage	character, a web page as text
patterns	character vector, one or more XPath patterns to extract from the web page
patnames	character vector, names of patterns

Details

contentscraper(x ,c("//head/title","//body/div/article"),c("title", "article"))

Value

return a named list of extracted content

Author(s)

salim khalil

Getencoding *Getencoding*

Description

A function that parse a web page content and retreive it's encoding charset based on content and HTTP header

Usage

```
Getencoding(url)
```

Arguments

url	character, url to

Details

xxx

Value

return the encoding charset as character

Author(s)

salim khalil

See Also

other function

getsimHash *getsimHash*

Description

A function that take a _charachter_ as input, and generate it's simhash.

Usage

```
getsimHash(string, hashbits)
```

Arguments

string	character, the content to hash.
hashbits	numeric, specify the hash bits 64 or 128

Details

This funcion call an external java class

Value

return the simhash as a nmeric value

Author(s)

salim khalil

LinkExtractor *LinkExtractor*

Description

A function that take a -charachter_ url as input, fetch the web page, gets its detail, and extract all links following a set of rules .

Usage

```
LinkExtractor(url, id, lev, IndexErrPages, Useragent, Timeout = 5,
  URLlenlimit = 255, urlExtfilter, statslinks = FALSE, encod, urlbotfiler,
  removeparams)
```

Arguments

url	character, url to fetch and extract links.
id	numeric, an id to identify a specific web page in a website collection, it's auto-generated by default
lev	numeric, the depth level of the web page, auto-generated by the Rcrawler function.
IndexErrPages	a vector of html error code-statut page to index, by default it's c(200), to include 404 and 403 pages c(404,403)
Useragent	, default to "Rcrawler"
Timeout	,default to 5s
URLlenlimit	interger, the url character length limit to index, default to 255 characters (to avoid spider traps)
urlExtfilter	character vector, the list of file extensions to exclude from indexing, by dfault a large list is defined (html pages only are permitted) in order to prevent large files downloading; To define your own use c(ext1,ext2,ext3 ...)
statslinks	boolean, specifies if input and output links shoud be counted, work only when the function is called from the main function scrawler
urlbotfiler	character vector , directories/files restricted by robot.txt
encoding	character, specify the encoding of th web page

Details

xxx

Value

return a list of two elements, the first is a list containing the page detail (url, encoding-type, content-type, content ... etc) and the second is a character-vector containing the list of urls discovered in that page

Author(s)

salim khalil

See Also

other function

LinkNormalization *Link Normalization*

Description

A function that take a URL _charachter_ as input, and transforms it into a canonical form.

Usage

```
LinkNormalization(links, current)
```

Arguments

current	character, The URL of the current page source of the link.
link	character, the URL to Normalize.

Details

This funcion call an external java class

Value

return the simhash as a nmeric value

Author(s)

salim khalil

Linkparameters *Link parameters*

Description

A function that take a URL _charachter_ as input, and extract the parameters and values from this URL .

Usage

```
Linkparameters(URL)
```

Arguments

URL	character, the URL to extract

Details

This function extract the link parameters and values (Up to 10 parameters)

Value

return the URL paremeters=values

Author(s)

salim khalil

Linkparamsfilter *Link parameters filter*

Description

This function remove a given set of parameters from a specific URL

Usage

```
Linkparamsfilter(URL, params)
```

Arguments

URL	character, the URL from which params and values have to be removed
params	character vector, parameters to be removed

Details

This function exclude given parameters from the urls,

Value

return a URL wihtout given parameters

Author(s)

salim khalil

Rcrawler *Rcrawler*

Description

A function that take a _charachter_ as input, and generate it's simhash.

Usage

```
Rcrawler(Website, no_cores, nbcon, MaxDepth, DIR, RequestsDelay = 0,
  duplicatedetect = FALSE, Obeyrobots = FALSE, IndexErrPages, Useragent,
  Timeout = 5, URLlenlimit = 255, urlExtfilter, urlregexfilter,
  ignoreUrlParams, statslinks = FALSE, Encod, patterns, excludepattern,
  Backup = FALSE)
```

Arguments

Website	character, the root URL of the web site to crawl
no_cores	integer, specify the number of clusters for parallel crawling, by default is the numbers of available cores
MaxDepth	integer, repsent the max deph level for the crawler, this is not the file depth in a directory structure, but 1+ number of links between this document and root document, default to 10
DIR	character, correspond to the path of the local repository the crawler will use to store crawled data ex, "C:/collection" , by default R working directory

Details

This funcion call an external java class for arabic use Sys.setlocale("LC_CTYPE","Arabic_Saudi Arabia.1256") and set encoding of the web page

Value

return the simhash as a nmeric value

Author(s)

salim khalil

RobotParser *RobotParser*

Description

This function fetch and parse robots.txt file of the website which is specified in the first argument and return the list of correspending rules .

Usage

```
RobotParser(website, useragent)
```

Arguments

website	character, url of the website which rules have to be extracted .
Useragent	character, the useragent of the crawler

Details

xxx

Value

return a list of three elements, the first is a character vector of Disallowed directories, the third is a Boolean value which is TRUE if the user agent of the crawler is blocked.

See Also

other function

Index