Package 'datacollector'

August 29, 2017

Title R package for automatic data collection

Version 0.1

Author Med Dhaker Abdeljawed

Maintainer Med Dhaker Abdeljawed <med.dhaker.abdeljawed@gmail.com>

Description A suite of functions and tools made to automate data collection from the internet, social media networks and webpages.

License Free

Imports rJava, stringr, xlsx

Depends R (>= 3.2.3)

Encoding UTF-8

LazyData true

RoxygenNote 6.0.1

Repository github

Prerequisite

1- Setup java environment including JAVA_HOME and jre/jdk path.

2- Setup Apache Zookeeper Download Zookeeper from
https://zookeeper.apache.org/releases.html#download, Start
server instance :
https://zookeeper.apache.org/doc/r3.3.3/zookeeperStarted.html#sc_InstallingSingleMode

3- Setup Apache Drill download drill from
https://drill.apache.org/download/ and start server instance :
https://drill.apache.org/docs/starting-drill-in-distributed-mode/
Note : the package should run in the same node with drillbit
instance

R topics documented:

`collectContentByKey` *Collect by keyword links from a web page*

Description

This function extract a webpage content by keyword, and return the blocs containing that keyword

Usage

```
collectContentByKey(url, key)
```

Arguments

`url`	url of the target web page
`key`	keyword to search for

Value

list of bloc content containing that keyword

`collectEmails` *Collect emails from a web page*

Description

this function extract all emails from a web page and return it as an R list

Usage

```
collectEmails(url)
```

Arguments

url	url of the target web page

Value

list of emails

---

collectFacebookNodes *Collect facebook data*

Description

This function collect data from a facebook node that could be a page, group, event or else based on a given edge which represents the type of data to collect such, posts and feed then parse it and save into an R dataframe and spredsheet excel file.

Usage

```
collectFacebookNodes(nodeid, edge = "feed", user = FALSE, rootPath)
```

Arguments

nodeid	Id of the target facebook node
user	FALSE to use an application based call, TRUE to use an Autenticated call (used to fetch private groups and events)
rootPath	path to save the generated results

---

collectFromYoutube *Collect youtube data*

Description

This function collect data about a specific youtube element attached to videos and channels including activities, playlists, subscriptions, comments and video captions tracks then parse it and save it into an R dataframe and an excel spredsheet file.

Usage

```
collectFromYoutube(type, id, rootPath)
```

Arguments

type	element type to fetch ("activity","playlist","comment","subscription","caption")
id	a video id or a channel id, video id for ("caption","comment") and channel id for the rest
rootPath	path to save the generated results

Value

R dataframe representing the result node

`collectMedias` *Collect media links from a web page*

Description

this function extract all medias such images, videos, audio, documents.. from a web page

Usage

```
collectMedias(url, media)
```

Arguments

`url`	url of the target web page
`media`	media type could be : image,video,document,audio

Value

list of medias

`collectPhones` *Collect phone numbers from a web page*

Description

this function extract all phone numbers from a web page and return it as an R list

Usage

```
collectPhones(url)
```

Arguments

`url`	url of the target web page

Value

list of phone numbers

collectSocialLinks *Collect SocialLinks*

Description

this function extract all social medias links from a web page and return it as an R list

Usage

```
collectSocialLinks(url)
```

Arguments

url	url of the target web page

Value

list of social media links

collectWebPageData *Collect data from a web page*

Description

This function scraps data from a web page and save it into an excel sheet.

Usage

```
collectWebPageData(excelpath, url)
```

Arguments

excelpath	Path to the excel file
url	url of the target web page

Details

Data such : Email, phones numbers, medias, links, social media links, and html tables

Value

void

collectWebSiteData *Collect data from a whole website*

Description

collectWebSiteData crawls a whole website for the requested data and save each scraped page into an excel sheet, and download the availble media files in a sepcific directory

Usage

```
collectWebSiteData(url, excelpath, downloadpath)
```

Arguments

url	url of the target web page
excelpath	Path to the excel file
downloadpath	path to the downloaded media files

Value

void

searchFacebookNodes *Search facebook data*

Description

This function searchs facebook nodes including groups, events and pages; then parse the results and save into an R dataframe and an excel spredsheet file.

Usage

```
searchFacebookNodes(query, rootPath)
```

Arguments

query	text query to use for the search
rootPath	path to save the generated results

searchFromFlickr *Search Flickr data*

Description

This function search for photos from Flickr based on a text query, then parse the result into an R dataframe and an excel spreedSheet

Usage

```
searchFromFlickr(query, rootPath)
```

Arguments

query	text query to use for the search
rootPath	path to save the generated results

searchFromGooglePlus *Search google plus data*

Description

This function searchs google plus activities using a text query, then parse the result into an R dataframe and spreedSheet xls

Usage

```
searchFromGooglePlus(query, rootPath)
```

Arguments

query	text query to use for the search
rootPath	path to save the generated results

searchFromTumblr *Search Tumblr data*

Description

This function search for blog posts from Tumblr using a text query, then parse the result into an R dataframe and spreedSheet xls

Usage

```
searchFromTumblr(query, rootPath)
```

Arguments

query	text query to use for the search
rootPath	path to save the generated results

searchFromTwitter *Search Twitter data*

Description

This function searchs Tweets of the past seven days (week) from twitter using a text query, then parse the result into an R dataframe and also creates an excel spreedSheet

Usage

```
searchFromTwitter(query, rootPath)
```

Arguments

query	text query to use for the search
rootPath	path to save the generated results

searchFromYouTube *Search Youtube data*

Description

This function search for videos and channels from youtube using a text query, then parse the result into an R dataframe and also an excel spreedSheet

Usage

```
searchFromYouTube(query, rootPath)
```

Arguments

query	text query to use for the search
rootPath	path to save the generated results

Index