Package 'datacollector'

July 23, 2017

Title RdataCollector

Version 0.1

Description this package collects data from web pages and social media networks

License Free

Depends R (>= 3.2.3)

Encoding UTF-8

LazyData true

RoxygenNote 6.0.1

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

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 facebook (nodes/edge) couple such posts and feed in pages, groups and events, then parse it and save into an R dataframe and spredsheet excel file.

Usage

```
collectFacebookNodes(nodeid, edge = "feed", user = FALSE,
  rootPath = "/home/dhaker/Desktop/Ghandi/")
```

Arguments

nodeid	Id if the target facebook node
user	FALSE to use an application based call, TRUE t o use Autenticated calls (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 track then parse it and save it into an R dataframe and an excel spredsheet file.

Usage

```
collectFromYoutube(type, id, rootPath = "/home/dhaker/Desktop/Ghandi/")
```

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 exctract 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 exctract all phone numbers from a web page

Usage

```
collectPhones(url)
```

Arguments

url	url of the target web page

Value

list of phone numbers

collectSocialLinks *Collect SocialLinks*

Description

this function exctract all social medias links from a web page

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

Usage

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

Arguments

`excelpath`	Path to the excel file
`url`	url of the target web page11

Details

Example of usage : collectWebPageData(<path.xls>,"https://www.apec.fr/")

Value

void

`collectWebSiteData` *Collect data from a whole website*

Description

this function crawls a whole website for the requested data and save each scraped page into an excel sheet, and download the availble media 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 download directory

Value

void

countData *Count data rows in a json file*

Description

This function excutes an SQL select count(*) query over a json file and return data row count'

Usage

```
countData(path, item = "*", flatten = FALSE, rootarray)
```

Arguments

path	Path to the json file
item	columns to count

Value

number of rows in the target file

downloadDocuments *Download documents from a target page*

Description

This functions extract all documents files into a provided directory path

Usage

```
downloadDocuments(path, url)
```

Arguments

path	Path to the download directory
url	url of the target web page

Value

void

downloadImages *Download images from a target page*

Description

This function exctract all image files from a target webpage and download it into a provided directory path Example of usage : downloadImages(<path>,"https://www.wallpaper.com/art")

Usage

```
downloadImages(path, url)
```

Arguments

path	Path to the download directory
url	url of the target web page

Value

void

downloadVideos *Download videos from a target page*

Description

This functions extract all video files and download it into a provided directory path

Usage

```
downloadVideos(path, url)
```

Arguments

path	Path to the download directory
url	url of the target web page

Value

void

jsonLoadByListFromYoutube *Collect List of Data from youtube*

Description

this function calls youtube Google api to retreive json data of a iDs list and save into a file

Usage

```
jsonLoadByListFromYoutube(path, type, ids)
```

Arguments

path	Path to save the json file
type	type of element to retreive (channel, video, playlist)
ids	array of ids

Value

void

jsonLoadFromFacebook *Get facebook graph edge of a specific node*

Description

this function calls facebook graph API endpoints to collect public data using nodes and edges and save it into a json file

Usage

```
jsonLoadFromFacebook(path, node, edge, user = FALSE,
  params = rJava::.jarray(c("")))
```

Arguments

path	Path to save the data
node	Facebook graph node id, eg(Page id, User id, Post id)...
edge	Graph edge name, eg(comment,posts,feed,likes) ...
user	TRUE to use user token for the edge call, FALSE to use the App token

jsonLoadFromYoutube *Collect Data from youtube*

Description

this function calls youtube Google api to retreive json data of a single given element and save into a file

Usage

```
jsonLoadFromYoutube(path, type, id)
```

Arguments

path	Path to save the json file
type	type of element to retreive (channel, video, playlist)
id	id of the element

Value

void

jsonLoadTopics *Collect topics from Social medias such twitter and google+*

Description

this function calls twitter and google+ api endpoints to retreive activities and tweets then save into json files

Usage

```
jsonLoadTopics(path, query)
```

Arguments

path	path of the json file to save
query	Search query

Value

void

jsonLoadUsers *Collect data about users of facebook and google+*

Description

this function collect and fetch data about facebook and google+ users and save it into a json file

Usage

```
jsonLoadUsers(path, query)
```

Arguments

path	Path to save file
query	Search query

Value

void

jsonSearchFomGooglePlus *Get google plus data*

Description

this function fetch activities from google plus API by submitting a filtered search in save it to a json file in the directory path provided

Usage

```
jsonSearchFomGooglePlus(path, query = "", lang = "en")
```

Arguments

path	path to save the data
query	text query to search for activities
lang	language code of the activities ...

Value

void

jsonSearchFromFacebook *Search for facebook nodes with query*

Description

this function call the search endpoint for facebook nodes including users, pages, groups and events by calling the facebook graph search endpoint, and save the result data into a json file

Usage

```
jsonSearchFromFacebook(path, node = "page", query)
```

Arguments

path	Path to save the data
node	Facebook node type : user, page, event, group
query	Text query of the search request

jsonSearchFromYoutube *search Data from youtube*

Description

this function calls youtube search endpoint api to retreive json data of a given query or type and save into a file

Usage

```
jsonSearchFromYoutube(path, type = "", query)
```

Arguments

path	Path to save the json file
type	type of element to retreive (channel, video, playlist)
query	query of the request

Value

void

`jsonSearchTweetsByFilter` *Get Twitter data by filter*

Description

this function target the twitter search endpoint of twitter API by filter and save the reponse data into a json file in the directory path provided

Usage

```
jsonSearchTweetsByFilter(path, lang = "en", exact = "", allword = "",
  hashtags = "", noneWords = "", oneOf = "", accounts = "",
  attitude = TRUE, question = FALSE)
```

Arguments

`path`	directory path to save the json data file
`lang`	language code of the tweets ...
`attitude`	postive attitude : True, negative attitude : False
`question`	question exist in tweet : True else false.
`allwords`	filter tweets by all words in the string seperated by space
`exactphrase`	filter tweets by exact string
`hashtag`	filter tweets by hashtags in this string seperated by space ...
`oneof`	filter tweets by one of the words in this string seperated by space ...
`noneof`	filter tweets by excluding all the words in this string seperated by space ...

Value

void

`jsonSearchTweetsByQuery` *Get Twitter data by text query*

Description

this function target the twitter search endpoint of twitter API by text query and save the reponse data into a json file in the directory path provided

Usage

```
jsonSearchTweetsByQuery(path, query)
```

Arguments

path	directory path to save the json data file
query	fetch tweets by the provided text query

Value

void

---

searchFacebookNodes *Search facebook data*

Description

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

Usage

```
searchFacebookNodes(query, rootPath = "/home/dhaker/Desktop/Ghandi/")
```

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 = "/home/dhaker/Desktop/Ghandi/")
```

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 last 7 days from twitter from using a text query, then parse the result into an R dataframe and spreedSheet xls

Usage

```
searchFromTwitter(query, rootPath = "/home/dhaker/Desktop/Ghandi/")
```

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 spreedSheet xls

Usage

```
searchFromYouTube(query, rootPath = "/home/dhaker/Desktop/Ghandi/")
```

Arguments

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

`selectData` *Filter data from json file*

Description

This function excutes an SQL select query over a json file and return filtered data'

Usage

```
selectData(path, item = "id", where = "", file = "", append = FALSE,
  flatten = FALSE)
```

Arguments

`path`	Path to the json file
`item`	column to select
`where`	an sql where clause.
`file`	an optional argument to save the result into a provided file path
`append`	TRUE to append file, FALSE to overwrite file

Value

array of result query

Index