Package ‘GuardianR’

July 1, 2018

Type Package

Title The Guardian API Wrapper

Version 0.9

Date 2018-05-30

Maintainer Marco Bastos <marco@toledobastos.com>

Depends R (>= 3.2.0)

Imports RCurl, RJSONIO

Description Provides an interface to the Open Platform’s Content API of the Guardian Media Group. It retrieves content from news outlets The Observer, The Guardian, and guardian.co.uk from 1999 to current day.

License GPL (>= 2)

Encoding UTF-8

NeedsCompilation no

Author Marco Bastos [aut, cre],
Cornelius Puschmann [aut],
Mark Johnman [ctb]

Repository CRAN

Date/Publication 2018-06-30 23:16:17 UTC

R topics documented:

get_guardian .................................................. 2
get_json ...................................................... 4
GuardianR ..................................................... 4
parse_json_to_df .......................................... 6

Index 7
get_guardian

Search Guardian API for news articles that match the criteria

Description

The function get_guardian takes four variables (keyword(s), starting date, end date, and API-key) and returns a data frame with 27 columns, with the last column including the full text of the article.

Search criteria accepts single or multiple keywords concatenated with "+" (e.g. "keyword1+keyword2"). It also accepts Boolean queries with and/or/not between words to refine searches. For exact phrases and matches, please encapsulate the keywords in %22 (e.g. "%22Death+of+Margaret+Thatcher%22").

From version 0.5 onwards, the function get_guardian returns the full text of articles and requires a Guardian API-key. Guardian API-key can be obtained by registering at <http://open-platform.theguardian.com/access/>.

Usage

get_guardian(keywords, section, format, fromNdate, toNdate, apiNkey)

Arguments

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>keywords</td>
<td>Keyword to search Guardian API. Example: &quot;Thatcher&quot;. For multiple keywords use &quot;Margaret+Hilda+Thatcher&quot;.</td>
</tr>
<tr>
<td>section</td>
<td>Specifies news sections to narrow the query (articles from other news desks will be filtered out). It doesn not accept blank spaces, so use &quot;world&quot; if querying for articles in the section &quot;world news&quot; (and so on). It defaults to NULL in which case the API request is performed across all sections. For reference, these are some of the most common news sections: entertainment, sport, world, opinion, economy, national, artanddesign, politics, environment, lifestyle, jobs, education, technology, science, fashion, culture, society</td>
</tr>
</tbody>
</table>
get_guardian

- tourism
- health
- cars
- uk-news
- business
- football
- stage
- money
- commentisfree

format
Specifies JSON or XML format. It defaults to JSON.

fromNdate
Start date of search.

toNdate
End date of search.

apiNkey
A Guardian API-key is necessary to retrieve the full text of news articles. A Guardian API-key can be obtained by registering at <http://open-platform.theguardian.com/access/>.

Value
Returns a data frame with 27 variables. Rows refer to news articles and columns to the news item’s data.

Author(s)
Marco Bastos & Cornelius Puschmann

References
Information about the Content API is available on Guardian Item Reference Guide (http://open-platform.theguardian.com/documentation/).

Examples
```r
## Not run:
results <- get_guardian("islamic+state", 
section="world", 
fromNdate="2014-09-16", 
toNdate="2014-09-16", 
apiNkey="212d23d3-c7b2-4273-8f1b-289a0803ca4b")
## End(Not run)
```
get_json

Request json at Guardian API

Description

The function get_json makes a call at Guardian API to retrieve articles that match the criteria. The function is called by get_guardian and should not be used standalone.

Usage

get_json(keywords, section, format, from.date, to.date, api.key)

Arguments

keywords Keyword to search Guardian API. Exemple: "Thatcher". For multiple keywords use "Margaret+Hilda+Thatcher".
section Specifies news sections to narrow the query (articles from other news desks will be filtered out). It doesn not accept blank spaces, so use "word" if querying for articles in the "World news" section (and so on). It defaults to NULL in which case API search is performed across all sections.
format Specifies format JSON or XML formar. It defaults to JSON.
from.date Start date of search.
to.date End date of search.
api.key Guardian API-key.

Author(s)

Marco Bastos & Cornelius Puschmann

GuardianR

The Guardian API Wrapper

Description

The package provides an interface to the Open Platform’s Content API of the Guardian Media Group. It retrieves content from news outlets The Observer, The Guardian, and guardian.co.uk from 1999 to current day. As of version 0.5, GuardianR connects to the Guardian Content API v2, which is much faster, but requires an API key and implements the limit of 12 Calls per second or 5,000 calls per day.
The results of keyword search include news articles that match the search terms and were published after 1999. Function `get_guardian` accepts date intervals with multiple keywords (format: "keyword1+keyword2") and returns a data frame. Rows refer to news articles and columns to the news item’s data. The resulting data frame includes the following 27 columns:

- **id**: News article ID
- **sectionId**: API news section ID to the news article
- **sectionName**: Guardian news section ID to the news article
- **webPublicationDate**: News article online publication date
- **webTitle**: Online headline
- **webUrl**: News article URL
- **apiUrl**: News article API URL
- **newspaperPageNumber**: Page where the news article was featured in the printed newspaper
- **trailText**: Introductions to the news article
- **headline**: Print headline
- **showInRelatedContent**: Related content (logical)
- **lastModified**: Date the news article was last modified (POSIXct)
- **hasStoryPackage**: (logical)
- **score**: Score provided by the API relative to the query and importance of the result
- **standfirst**: Short summary to the news article
- **shortUrl**: Short URL to the news article
- **wordcount**: News article’s word count
- **commentable**: Comments included (logical)
- **allowUgc**: Allows User Generated Content (logical)
- **isPremoderated**: Pre-moderated news articles (logical)
- **byline**: News article author
- **publication**: Source of the published news article (Oberserver, The Guardian, or guardian.co.uk)
- **newspaperEditionDate**: Publication date of print news article (POSIXct)
- **shouldHideAdverts**: Hide ads (logical)
- **liveBloggingNow**: live blogging (logical)
- **commentCloseDate**: Date comments closed (POSIXct)
body News article full text
At the time of this writing, Microsoft Windows does not have integration between the OS default CA certificates and OpenSS. Because of that, the function get_guardian will verify the operation system and when necessary download the SSL certificate "cacert.perm" to access https links.
To see how to cite the package, check citation("GuardianR").

Author(s)
Marco Bastos & Cornelius Puschmann
Maintainer: Marco Bastos <marco@toledobastos.com>

References
More information on Content API is available on Guardian Item Reference Guide (http://www.guardian.co.uk/open-platform/content-api-item-reference-guide) and the sign up for an API key page (https://open-platform.theguardian.com/access).

---

parse_json_to_df

**Parses json data to data frame**

**Description**
The function `parse_json_to_df` parses the json data to a data frame in R. The function is called by `get_guardian` and should not be used standalone.

**Usage**

```r
parse_json_to_df(api.responses)
```

**Arguments**

- `api.responses` JSON response from Guardian API

**Author(s)**
Marco Bastos & Cornelius Puschmann
Index

*Topic IO
  get_guardian, 2
  get_json, 4
  parse_json_to_df, 6

*Topic utilities
  get_guardian, 2
  get_json, 4
  parse_json_to_df, 6

get_guardian, 2
get_json, 4
GuardianR, 4
GuardianR-package (GuardianR), 4

parse_json_to_df, 6