Package 'rscielo'

August 22, 2019

Type Package

Title A Scraper for Scientific Journals Hosted on Scielo

Version 1.0.0

Description Scrapes data from scientific articles hosted on the Scientific Electronic Library Online Platform <http://www.scielo.br/>. The data information includes author's names, articles' metadata and contents, among others. The package also provides additional functions to easily summarize the scraped data.

License GPL-3

Depends R (>= 3.1)

Imports graphics, magrittr, stats, xml2, httr (>= 0.5), rvest, stringr, tibble, purrr, dplyr

LazyData TRUE

URL https://github.com/meirelesff/rscielo

BugReports https://github.com/meirelesff/rscielo/issues

RoxygenNote 6.1.1

NeedsCompilation no

Author Fernando Meireles [aut, cre] (<https://orcid.org/0000-0002-7027-2058>), Denisson Silva [aut] (<https://orcid.org/0000-0003-2771-8146>), Rogerio Barbosa [aut] (<https://orcid.org/0000-0002-7027-2058>)

Maintainer Fernando Meireles <fernando.meireles@iesp.uerj.br>

Repository CRAN

Date/Publication 2019-08-22 12:10:02 UTC

R topics documented:

`rscielo-package` *rscielo: A Scraper for Scientific Journals Hosted on Scielo*

Description

rscielo provides functions to easily scrape bibliometric information from scientific journals and articles hosted on the Scientific Electronic Library Online Platform (Scielo.br). The retrieved data includes a journal's details and citation counts; article's contents, footnotes, bibliographic references; and several other common information used in bibliometric studies. The package also offers functions to quickly summarize the scrapped data.

Details

To learn more about rscielo, check the package documentation.

Author(s)

Fernando Meireles [author], Denisson Silva [author], Rogerio Barbosa [author]

See Also

Useful links:

- https://github.com/meirelesff/rscielo
- Report bugs at https://github.com/meirelesff/rscielo/issues

---

`get_article` *Scrape text from a single article hosted on Scielo*

Description

`get_article()` scrapes the full text from an article hosted on Scielo. In bilingual journals, the text retrieved is in the journal's main language used for publication (most of the time, it is English).

Usage

```
get_article(x, output_text = TRUE)
```

Arguments

x	a character vector with the link or id of the article hosted on Scielo to be scrapped.
output_text	a logical indicating whether get_article() should return a character vector or a tibble (defaults to TRUE).

Value

When the argument output_text is TRUE, the function returns a character vector with the requested article's content. When output_text is FALSE, the function returns a tibble with the following variables:

- text: article's full text (character).
- doi: article's Digital Object Identifier (DOI, (character)).

Note

Sometimes, the Scielo website is offline for maintaince, in which cases this function will not work (i.e., users will get HTML status different from the usual 200 OK).

Examples

```
article <- get_article(x = "http://www.scielo.br/scielo.php?
script=sci_arttext&pid=S1981-38212016000200201&lng=en&nrm=iso&tlng=en")
```

---

get_article_footnotes *Scrape footnotes from a single article hosted on Scielo*

Description

get_article_footnotes() scrapes all the footnotes iin an article hosted on Scielo.

Usage

```
get_article_footnotes(x)
```

Arguments

x	a character vector with the link or id of the article hosted on Scielo to be scrapped.

Value

The function returns a tibble with the following variables:

- footnote: article's footnotes (character).
- doi: article's Digital Object Identifier (character).

Note

Sometimes, the Scielo website is offline for maintaince, in which cases this function will not work (i.e., users will get HTML status different from the usual 200 OK).

Examples

```
df <- get_article_fnotes(x = "http://www.scielo.br/scielo.php?
script=sci_arttext&pid=S1981-38212016000200201&lng=en&nrm=iso&tlng=en")
```

---

get_article_id *Get the ID of a scientific article hosted on Scielo*

Description

`get_article_id()` extracts the ID of an article's URL

Usage

```
get_article_id(url)
```

Arguments

`url`	a character vector with the URL of an article hosted on Scielo.

Value

The function returns a `character` vector with the article ID.

Examples

```
id <- get_article_id(url = "http://www.scielo.br/scielo.php?script=sci_arttext&
pid=S1981-38212016000200201&lng=en&nrm=iso&tlng=en")
```

`get_article_meta` *Scrape meta-data from a single article hosted on Scielo*

Description

`get_article_meta()` scrapes meta-data information from an article hosted on Scielo.

Usage

```
get_article_meta(x)
```

Arguments

`x`	a character vector with the link or id of the article hosted on Scielo to be scrapped.

Details

This functions scrapes several meta-data information, such as author's names, article title, year of publication, journal issue and number of pages.

Value

The function returns a `tibble` with the following variables:

- author: Author name.
- first_author_surname: First author surname.
- institution: Author's institution.
- inst_adress: Author's institution address.
- country: Author's country.
- title: Article title.
- year: Year of publication.
- journal: Journal name.
- volume: Volume.
- number: Number.
- first_page: Article's first page.
- last_page: Article's last page
- abstratc: Article's abstract.
- keywords: Article's keywords.
- article_id:
- doi: DOI.
- n_authors: Number of authors.
- n_pages: Number of pages.
- n_refs: Number of references.

Note

Sometimes, the Scielo website is offline for maintaince, in which cases this function will not work (i.e., users will get HTML status different from the usual 200 OK).

See Also

`get_journal`

Examples

```
article_meta <- get_article_meta(x = "http://www.scielo.br/scielo.php?
script=sci_arttext&pid=S1981-38212016000200201&lng=en&nrm=iso&tlng=en")
```

---

get_article_references

Scrape bibliographic references from a single article hosted on Scielo

---

Description

`get_article_references()` scrapes a list of bibliographic references cited by an article hosted on Scielo.

Usage

```
get_article_references(x)
```

Arguments

x	a character vector with the link or the id of the article hosted on Scielo to be scrapped.

Value

The function a `tibble` with the following variables:

- references: an article's bibliographic reference (`character`).
- doi: article's Digital Object Identifier (DOI).

Note

Sometimes, the Scielo website is offline for maintaince, in which cases this function will not work (i.e., users will get HTML status different from the usual 200 OK).

Examples

```
refs <- get_article_references(x = "http://www.scielo.br/scielo.php?
script=sci_arttext&pid=S1981-38212016000200201&lng=en&nrm=iso&tlng=en")
```

`get_journal` *Scrape meta-data from articles published by a journal hosted on Scielo*

Description

`get_journal()` scrapes meta-data information from articles of a journal hosted on Scielo. In bilingual journals, articles' titles, abstracts and other relevant information are retrieved in the journal's main language used for publication (most of the time, it is English). The function can extract information from all articles ever published by the journal or only the ones in its latest issue.

Usage

```
get_journal(journal_id, last_issue = TRUE)
```

Arguments

`journal_id`	a character vector with the ID of the journal hosted on Scielo (the `get_ournal_id` function can be used to find the journal ID from its URL).
`last_issue`	a logical vector, if `FALSE` scrapes all issues of the journal, if `TRUE` (default) only scrapes its last issue.

Details

This functions scrapes several meta-data information, such as author's names, articles' titles, year of publication, edition and number of pages, that can be summarized with specific `summary` method.

Value

The function returns a `tibble` with the following variables:

- author: Author name.
- first_author_surname: First author surname.
- institution: Author's institution.
- inst_adress: Author's institution address.
- country: Author's country.
- title: Article title.
- year: Year of publication.
- journal: Journal name.
- volume: Volume.

- number: Number.
- first_page: Article's first page.
- last_page: Article's last page
- abstratc: Article's abstract.
- keywords: Article's keywords.
- article_id:
- doi: DOI.
- n_authors: Number of authors.
- n_pages: Number of pages.
- n_refs: Number of references.

Note

Sometimes, the Scielo website is offline for maintaince, in which cases this function will not work (i.e., users will get HTML status different from the usual 200 OK).

See Also

get_article_meta

Examples

```
df <- get_journal(journal_id = "1981-3821")
summary(df)
```

---

get_journal_id *Get the ID of a journal hosted on Scielo*

Description

`get_journal_id()` extracts the numerical ID (pid) from a journal's URL.

Usage

```
get_journal_id(url)
```

Arguments

`url`	a character vector with the URL of a journal hosted on Scielo.

Value

The function returns a `character` vector with the journal ID.

Examples

```
id <- get_journal_id(url = "http://www.scielo.br/scielo.php?
script=sci_serial&pid=1981-3821&lng=en&nrm=iso")
```

`get_journal_info` *Scrape the description of a journal hosted on Scielo*

Description

`get_journal_info()` scrapes the description (publisher, issn, and mission) information of a journal hosted on Scielo.

Usage

```
get_journal_info(journal_id)
```

Arguments

`journal_id`	a character vector with the ID of the journal hosted on Scielo (the `get_journal_id` function can be used to find a journal's ID from its URL).

Value

The function returns a `tibble` with the journal's description.

Note

Sometimes, the Scielo website is offline for maintaince, in which cases this function will not work (i.e., users will get HTML status different from the usual 200 OK).

Examples

```
journal_info <- get_journal_info(journal_id = "1981-3821")
```

get_journal_list *Scrape a list with all the journals hosted on Scielo*

Description

get_journal_list() scrapes the title, the numerical ID (pid) and the URL of all journals hosted on Scielo.

Usage

```
get_journal_list()
```

Value

The function returns a tibble with each journal's title, ID, and URL

Examples

```
journal_list <- get_journal_list()
```

get_journal_metrics *Scrape publication and citation counts of a journal hosted on Scielo*

Description

get_journal_metrics() scrapes publication and citation counts of a jornal hosted on Scielo.

Usage

```
get_journal_metrics(journal_id)
```

Arguments

journal_id	a character vector with the ID of the journal hosted on Scielo (the get_journal_id function can be used to find a journal's ID from its URL).

Value

The function returns a tibble with the following variables:

- year: Year.
- n_issues: Number of issues in that year.
- n_articles: Number of articles in that year.
- granted_citations: Granted citations by the journal in that year.
- received_citations: Received citations by the journal in that year.
- avg_art_per_issues: Average number of articles published by the journal in that year.

Note

Sometimes, the Scielo website is offline for maintaince, in which cases this function will not work (i.e., users will get HTML status different from the usual 200 OK).

Examples

```
df <- get_journal_metrics(journal_id = "1981-3821")
```

Index