Package 'gutenbergr'

May 29, 2025

Title Download and Process Public Domain Works from Project Gutenberg

Version 0.3.0

Description Download and process public domain works in the Project
Gutenberg collection <https://www.gutenberg.org/>. Includes metadata
for all Project Gutenberg works, so that they can be searched and
retrieved.

License GPL-2

URL <https://docs.ropensci.org/gutenbergr/>,
<https://github.com/ropensci/gutenbergr>

BugReports <https://github.com/ropensci/gutenbergr/issues>

Depends R (>= 4.1)

Imports cli, dplyr, glue, purrr, readMDTable, readr, rlang, stringr,
tibble, urltools

Suggests curl, knitr, rmarkdown, testthat (>= 3.0.0), tidyr, tidytext,
withr

VignetteBuilder knitr

Encoding UTF-8

Language en-US

LazyData TRUE

LazyDataCompression xz

RoxygenNote 7.3.2

Config/testthat/edition 3

NeedsCompilation no

Author Jon Harmon [aut, cre] (ORCID: <https://orcid.org/0000-0003-4781-4346>),
Myfanwy Johnston [aut],
Jordan Bradford [aut] (ORCID: <https://orcid.org/0009-0000-8570-3474>),
David Robinson [aut, cph]

Maintainer Jon Harmon <jonthegeek@gmail.com>

Repository CRAN

Date/Publication 2025-05-29 19:40:20 UTC

Contents

---

gutenberg_authors   *Metadata about Project Gutenberg authors*

Description

Data frame with metadata about each author of a Project Gutenberg work. Although the Project Gutenberg raw data also includes metadata on contributors, editors, illustrators, etc., this dataset contains only people who have been the single author of at least one work.

Usage

```
gutenberg_authors
```

Format

A tbl_df (see tibble or dplyr) with one row for each author, with the columns

gutenberg_author_id Unique identifier for the author that can be used to join with the [gutenberg_metadata](#) dataset

author The agent_name field from the original metadata

alias Alias

birthdate Year of birth

deathdate Year of death

wikipedia Link to Wikipedia article on the author. If there are multiple, they are "|"-delimited

aliases Character vector of aliases. If there are multiple, they are "/"-delimited

Details

To find the date on which this metadata was last updated, run attr(gutenberg_authors, "date_updated").

See Also

[gutenberg_metadata](#), [gutenberg_subjects](#)

Examples

```
# date last updated
attr(gutenberg_authors, "date_updated")
```

---

gutenberg_download   *Download one or more works using a Project Gutenberg ID*

Description

Download one or more works by their Project Gutenberg IDs into a data frame with one row per line per work. This can be used to download a single work of interest or multiple at a time. You can look up the Gutenberg IDs of a work using gutenberg_works() or the gutenberg_metadata dataset.

Usage

```
gutenberg_download(
  gutenberg_id,
  mirror = NULL,
  strip = TRUE,
  meta_fields = character(),
  verbose = TRUE
)
```

Arguments

gutenberg_id	A vector of Project Gutenberg IDs, or a data frame containing a gutenberg_id column, such as from the results of gutenberg_works().
mirror	A mirror URL to retrieve the books from. By default uses the mirror from gutenberg_get_mirror().
strip	Whether to strip suspected headers and footers using gutenberg_strip().
meta_fields	Additional fields describing each book, such as title and author, to add from gutenberg_metadata.
verbose	Whether to show messages about the Project Gutenberg mirror that was chosen

Value

A two column tbl_df (see tibble::tibble()) with one row for each line of the text or texts, with columns

gutenberg_id Integer column with the Project Gutenberg ID of each text

text A character vector of lines of text

Examples

```
# download The Count of Monte Cristo
gutenberg_download(1184)

# download two books: Wuthering Heights and Jane Eyre
books <- gutenberg_download(c(768, 1260), meta_fields = "title")
books
dplyr::count(books, title)

# download all books from Jane Austen
austen <- gutenberg_works(author == "Austen, Jane") |>
  gutenberg_download(meta_fields = "title")
austen
dplyr::count(austen, title)
```

---

gutenberg_get_all_mirrors

Get all mirror data from Project Gutenberg

Description

Get all mirror data from `https://www.gutenberg.org/MIRRORS.ALL`. This only includes mirrors reported to Project Gutenberg and verified to be relatively stable. For more information on mirroring and getting your own mirror listed, see `https://www.gutenberg.org/help/mirroring.html`.

Usage

```
gutenberg_get_all_mirrors()
```

Value

A tbl_df of Project Gutenberg mirrors and related data

continent Continent where the mirror is located

nation Nation where the mirror is located

location Location of the mirror

provider Provider of the mirror

url URL of the mirror

note Special notes

Examples

```
gutenberg_get_all_mirrors()
```

---

gutenberg_get_mirror   *Get the recommended mirror for Gutenberg files*

Description

Get the recommended mirror for Gutenberg files and set the global `gutenberg_mirror` options.

Usage

```
gutenberg_get_mirror(verbose = TRUE)
```

Arguments

verbose   Whether to show messages about the Project Gutenberg mirror that was chosen

Value

A character vector with the url for the chosen mirror.

Examples

```
gutenberg_get_mirror()
```

---

gutenberg_languages   *Metadata about Project Gutenberg languages*

Description

Data frame with metadata about the languages of each Project Gutenberg work.

Usage

```
gutenberg_languages
```

Format

A tbl_df (see tibble or dplyr) with one row for each author, with the columns

gutenberg_id Unique identifier for the work that can be used to join with the [gutenberg_metadata](#) dataset

language Language ISO 639 code. Two letter code if one exists, otherwise three letter.

total_languages Number of languages for this work.

Details

To find the date on which this metadata was last updated, run attr(gutenberg_languages, "date_updated").

See Also

gutenberg_metadata, gutenberg_subjects

Examples

```
# date last updated
attr(gutenberg_languages, "date_updated")
```

---

gutenberg_metadata   *Gutenberg metadata about each work*

Description

Selected fields of metadata about each of the Project Gutenberg works.

Usage

```
gutenberg_metadata
```

Format

A `tibble::tibble()` with one row for each work in Project Gutenberg and the following columns:

gutenberg_id Numeric ID, used to retrieve works from Project Gutenberg

title Title

author Author, if a single one given. Given as last name first (e.g. "Doyle, Arthur Conan")

author_id Project Gutenberg author ID

language Language ISO 639 code, separated by / if multiple. Two letter code if one exists, otherwise three letter. See https://en.wikipedia.org/wiki/List_of_ISO_639-2_codes

gutenberg_bookshelf Which collection or collections this is found in, separated by / if multiple

rights Generally one of three options: "Public domain in the USA." (the most common by far), "Copyrighted. Read the copyright notice inside this book for details.", or "None"

has_text Whether there is a file containing digits followed by .txt in Project Gutenberg for this record (as opposed to, for example, audiobooks). If not, cannot be retrieved with gutenberg_download()

Details

To find the date on which this metadata was last updated, run attr(gutenberg_metadata, "date_updated").

See Also

gutenberg_works, gutenberg_authors, gutenberg_subjects

Examples

```
library(dplyr)
library(stringr)

gutenberg_metadata

gutenberg_metadata |>
  count(author, sort = TRUE)

# look for Shakespeare, excluding collections (containing "Works") and
# translations
shakespeare_metadata <- gutenberg_metadata |>
  filter(
    author == "Shakespeare, William",
    language == "en",
    !str_detect(title, "Works"),
    has_text,
    !str_detect(rights, "Copyright")
  ) |>
  distinct(title)

# note that the gutenberg_works() function filters for English
# non-copyrighted works and does de-duplication by default:

shakespeare_metadata2 <- gutenberg_works(
  author == "Shakespeare, William",
  !str_detect(title, "Works")
)

# date last updated
attr(gutenberg_metadata, "date_updated")
```

| gutenberg_strip | *Strip header and footer content from a Project Gutenberg book* |

Description

Strip header and footer content from a Project Gutenberg book. This is based on some formatting guesses so it may not be perfect. It will also not strip tables of contents, prologues, or other text that appears at the start of a book.

Usage

```
gutenberg_strip(text)
```

Arguments

text   A character vector with lines of a book.

Value

A character vector with Project Gutenberg headers and footers removed

Examples

```
book <- gutenberg_works(title == "Pride and Prejudice") |>
  gutenberg_download(strip = FALSE)

head(book$text, 10)
tail(book$text, 10)

text_stripped <- gutenberg_strip(book$text)

head(text_stripped, 10)
tail(text_stripped, 10)
```

---

gutenberg_subjects   *Gutenberg metadata about the subject of each work*

Description

Gutenberg metadata about the subject of each work, particularly Library of Congress Classifications
(lcc) and Library of Congress Subject Headings (lcsh).

Usage

```
gutenberg_subjects
```

Format

A tbl_df (see tibble or dplyr) with one row for each pairing of work and subject, with columns:

gutenberg_id ID describing a work that can be joined with gutenberg_metadata

subject_type Either "lcc" (Library of Congress Classification) or "lcsh" (Library of Congress Subject Headings)

subject Subject

Details

Find more information about Library of Congress Categories here: https://www.loc.gov/catdir/cpso/lcco/, and about Library of Congress Subject Headings here: https://id.loc.gov/authorities/subjects.html.

To find the date on which this metadata was last updated, run `attr(gutenberg_subjects, "date_updated")`.

See Also

[gutenberg_metadata,](#) [gutenberg_authors](#)

Examples

```
library(dplyr)
library(stringr)

gutenberg_subjects |>
  filter(subject_type == "lcsh") |>
  count(subject, sort = TRUE)

sherlock_holmes_subjects <- gutenberg_subjects |>
  filter(str_detect(subject, "Holmes, Sherlock"))

sherlock_holmes_subjects

sherlock_holmes_metadata <- gutenberg_works() |>
  filter(author == "Doyle, Arthur Conan") |>
  semi_join(sherlock_holmes_subjects, by = "gutenberg_id")

sherlock_holmes_metadata


holmes_books <- gutenberg_download(sherlock_holmes_metadata$gutenberg_id)

holmes_books


# date last updated
attr(gutenberg_subjects, "date_updated")
```

---

gutenberg_works   *Get a filtered table of Gutenberg work metadata*

Description

Get a table of Gutenberg work metadata that has been filtered by some common (settable) defaults, along with the option to add additional filters. This function is for convenience when working with common conditions when pulling a set of books to analyze. For more detailed filtering of the entire Project Gutenberg metadata, use the [gutenberg_metadata](#) and related datasets.

Usage

```
gutenberg_works(
  ...,
  languages = "en",
```

```
  only_text = TRUE,
  rights = c("Public domain in the USA.", "None"),
  distinct = TRUE,
  all_languages = FALSE,
  only_languages = TRUE
)
```

Arguments

`...`	Additional filters, given as expressions using the variables in the gutenberg_metadata dataset (e.g. author == "Austen, Jane")
`languages`	Vector of languages to include
`only_text`	Whether the works must have Gutenberg text attached. Works without text (e.g. audiobooks) cannot be downloaded with gutenberg_download
`rights`	Values to allow in the `rights` field. By default allows public domain in the US or "None", while excluding works under copyright. NULL allows any value of Rights
`distinct`	Whether to return only one distinct combination of each title and gutenberg_author_id. If multiple occur (that fulfill the other conditions), it uses the one with the lowest ID
`all_languages`	Whether, if multiple languages are given, all of them need to be present in a work. For example, if `c("en", "fr")` are given, whether only en/fr as opposed to English or French works should be returned
`only_languages`	Whether to exclude works that have other languages besides the ones provided. For example, whether to include en/fr when English works are requested

Details

By default, returns

- English-language works
- That are in text format in Gutenberg (as opposed to audio)
- Whose text is not under copyright
- At most one distinct field for each title/author pair

Value

A tbl_df (see the tibble or dplyr packages) with one row for each work, in the same format as gutenberg_metadata.

Examples

```
library(dplyr)

gutenberg_works()
```

```
# filter conditions
gutenberg_works(author == "Shakespeare, William")

# language specifications

gutenberg_works(languages = "es") |>
  count(language, sort = TRUE)

gutenberg_works(languages = c("en", "es")) |>
  count(language, sort = TRUE)

gutenberg_works(languages = c("en", "es"), all_languages = TRUE) |>
  count(language, sort = TRUE)

gutenberg_works(languages = c("en", "es"), only_languages = FALSE) |>
  count(language, sort = TRUE)
```

---

sample_books   *Sample Book Downloads*

Description

A tibble of book text for two sample books, generated using gutenberg_download().

Usage

```
sample_books
```

Format

A tbl_df (from tibble::tibble()) with one row for each line of text from each book, with columns:

gutenberg_id Unique identifier for the work that can be used to join with the gutenberg_metadata dataset.

text A character vector of lines of text.

title The title of this work.

author The author of this work.

Details

This code was used to download the books: gutenberg_download(c(109, 105), meta_fields = c("title", "author"))

Index