Package 'ctrialsgov'

July 6, 2025

Type Package

Title Query Data from U.S. National Library of Medicine's Clinical Trials Database

Version 0.2.7

Maintainer Taylor Arnold <tarnold2@richmond.edu>

Description Tools to create and query database from the U.S. National Library of Medicine's Clinical Trials database <https://clinicaltrials.gov/>. Functions provide access a variety of techniques for searching the data using range queries, categorical filtering, and by searching for full-text keywords.

License MIT + file LICENSE

Encoding UTF-8

Depends R (>= 3.5.0)

Imports dplyr, utils, lubridate, purrr, rlang, stringi, tibble, DBI, methods, Matrix

Suggests covr, knitr, rmarkdown, testthat (>= 3.0.0), usethis

VignetteBuilder knitr

LazyData true

LazyDataCompression bzip2

Config/testthat/edition 3

RoxygenNote 7.3.2

NeedsCompilation no

Author Taylor Arnold [aut, cre] (ORCID: <https://orcid.org/0000-0003-0576-0669>), Auston Wei [aut], Michael J. Kane [aut] (ORCID: <https://orcid.org/0000-0003-1899-6662>)

Repository CRAN

Date/Publication 2025-07-06 20:30:02 UTC

Contents

`cancer_studies` *Sample of Industry Cancer Trials from 2021*

Description

Cancer clinical trials based on a query where: ‘study_type‘ is "Interventional"; ‘sponsor_type‘ is "Industry"; ‘date_range‘ is trials from 2021-01-01 or newer; The ‘description‘ includes the keyword "cancer"; ‘phase‘ is reported (not NA); ‘primary_purpose‘ is "Treatment"; ‘minimum_enrollment‘ is 100.

`ctgov_create_data` *Initialize the connection*

Description

This function must be run prior to other functions in the package. It creates a parsed and cached version of the clinical trials dataset in memory in R. This makes other function calls relatively efficient. other

Usage

```
ctgov_create_data(con, verbose = TRUE)
```

Arguments

`con`	an DBI connection object to the database
`verbose`	logical flag; should progress messages be printed?; defaults to `TRUE`

Value

does not return any value; used only for side effects

Author(s)

Taylor B. Arnold, <taylor.arnold@acm.org>

`ctgov_kwic` *Keywords in Context*

Description

Takes a keyword and vector of text and returns instances where the keyword is found within the text.

Usage

```
ctgov_kwic(
  term,
  text,
  names = NULL,
  n = Inf,
  ignore_case = TRUE,
  use_color = FALSE,
  width = 20L,
  output = c("cat", "character", "data.frame")
)
```

Arguments

`term`	search term as a string
`text`	vector of text to search
`names`	optional vector of names corresponding to the text
`n`	number of results to return; default is Inf
`ignore_case`	should search ignore case? default is TRUE
`use_color`	printed results include ASCII color escape sequences; these are set to `FALSE` because they only work correctly when returned in the terminal
`width`	how many characters to show as context
`output`	what kind of output to provide; default prints the results using `cat`

Value

either nothing, character vector, or data frame depending on the the requested return type

ctgov_load_cache *Download and/or load cached data*

Description

This function downloads a saved version of the full clinical trials dataset from the package's development repository on GitHub (~150MB) and loads it into R for querying. The data will be cached so that it can be re-loaded without downloading. We try to update the cache frequently so this is a convenient way of grabbing the data if you do not need the most up-to-date version of the database.

Usage

```
ctgov_load_cache(force_download = FALSE)
```

Arguments

force_download	logical flag; should the cache be re-downloaded if it already exists? defaults to FALSE

Value

does not return any value; used only for side effects

Author(s)

Taylor B. Arnold, <taylor.arnold@acm.org>

ctgov_load_sample *Load sample dataset*

Description

This function loads a sample dataset for testing and prototyping purposes. after running, all of the functions in the package can then be used with this sample data. It consists of a 2.5 from ClinicalTrials.gov at the time of the package creation.

Usage

```
ctgov_load_sample()
```

Value

does not return any value; used only for side effects

Author(s)

Taylor B. Arnold, <taylor.arnold@acm.org>

ctgov_query *Query the ClinicalTrials.gov dataset*

Description

This function selects a subset of the clinical trials data by using a a variety of different search parameters. These include free text search keywords, range queries for the continuous variables, and exact matches for categorical fields. The function `ctgov_query_terms` shows the categorical levels for the latter. The function will either take the entire dataset loaded into the package environment or a previously queried input.

Usage

```
ctgov_query(
  data = NULL,
  description_kw = NULL,
  sponsor_kw = NULL,
  brief_title_kw = NULL,
  official_title_kw = NULL,
  criteria_kw = NULL,
  intervention_kw = NULL,
  intervention_desc_kw = NULL,
  outcome_kw = NULL,
  outcome_desc_kw = NULL,
  conditions_kw = NULL,
  population_kw = NULL,
  date_range = NULL,
  enrollment_range = NULL,
  minimum_age_range = NULL,
  maximum_age_range = NULL,
  study_type = NULL,
  allocation = NULL,
  intervention_model = NULL,
  observational_model = NULL,
  primary_purpose = NULL,
  time_perspective = NULL,
  masking_description = NULL,
  sampling_method = NULL,
  phase = NULL,
  gender = NULL,
  sponsor_type = NULL,
  ignore_case = TRUE,
  match_all = FALSE
)
```

Arguments

`data`	a dataset to search over; set to `NULL` to use the full dataset that is currently loaded

`description_kw` character vector of keywords to search in the intervention description field. Set to `NULL` to avoid searching this field.

`sponsor_kw` character vector of keywords to search in the sponsor (the company that submitted the study). Set to `NULL` to avoid searching this field.

`brief_title_kw` character vector of keywords to search in the brief title field. Set to `NULL` to avoid searching this field.

`official_title_kw` character vector of keywords to search in the official title field. Set to `NULL` to avoid searching this field.

`criteria_kw` character vector of keywords to search in the criteria field. Set to `NULL` to avoid searching this field.

`intervention_kw` character vector of keywords to search in the intervention names field. Set to `NULL` to avoid searching this field.

`intervention_desc_kw` character vector of keywords to search in the intervention description field. Set to `NULL` to avoid searching this field.

`outcome_kw` character vector of keywords to search in the outcome measures field. Set to `NULL` to avoid searching this field.

`outcome_desc_kw` character vector of keywords to search in the outcome description field. Set to `NULL` to avoid searching this field.

`conditions_kw` character vector of keywords to search in the conditions field. Set to `NULL` to avoid searching this field.

`population_kw` character vector of keywords to search in the population field. Set to `NULL` to avoid searching this field.

`date_range` string of length two formatted as "YYYY-MM-DD" describing the earliest and latest data to include in the results. Use a missing value for either value search all dates. Set to `NULL` to avoid searching this field.

`enrollment_range` numeric of length two describing the smallest and largest enrollment sizes to include in the results. Use a missing value for either value to avoid filtering. Set to `NULL` to avoid searching this field.

`minimum_age_range` numeric of length two describing the smallest and largest minmum age (in years) to include in the results. Use a missing value for either value to avoid filtering. Set to `NULL` to avoid searching this field.

`maximum_age_range` numeric of length two describing the smallest and largest maximum age (in years) to include in the results. Use a missing value for either value to avoid filtering. Set to `NULL` to avoid searching this field.

`study_type` character vector of study types to include in the output. Set to `NULL` to avoid searching this field.

`allocation` character vector of allocations to include in the output. Set to `NULL` to avoid searching this field.

intervention_model
: character vector of interventions to include in the output. Set to NULL to avoid searching this field.

observational_model
: character vector of observations to include in the output. Set to NULL to avoid searching this field.

primary_purpose
: character vector of primary purposes to include in the output. Set to NULL to avoid searching this field.

time_perspective
: character vector of time perspectives to include in the output. Set to NULL to avoid searching this field.

masking_description
: character vector of maskings to include in the output. Set to NULL to avoid searching this field.

sampling_method
: character vector of sampling methods to include in the output. Set to NULL to avoid searching this field.

phase
: character vector of phases to include in the output. Set to NULL to avoid searching this field.

gender
: character vector of genders to include in the output. Set to NULL to avoid searching this field.

sponsor_type
: character vector of sponsor types to include in the output. Set to NULL to avoid searching this field.

ignore_case
: logical. Should the search ignore capitalization. The default is TRUE.

match_all
: logical. Should the results required matching all the keywords? The default is FALSE.

Value

a tibble object queried from the loaded database

Author(s)

Taylor B. Arnold, <taylor.arnold@acm.org>

---

ctgov_query_terms *Query the ClinicalTrials.gov dataset*

Description

Returns a list showing the available category levels for querying the data with the ctgov_query function.

Usage

```
ctgov_query_terms()
```

Value

a named list of allowed categorical values for the query

---

ctgov_schema *Get and Set the Default Schema*

Description

This function sets the schema in which tables in which the CT Trials tables reside.

Get the current schema eiter of the following.

ctgov_schema() ctgov_get_schema()

Set the current schema with the following.

ctgov_schema(<SCHEMA NAME>) ctgov_set_schema(<SCHEMA NAME>)

A return of "" from the get functions indicates a schema is not specified.

Usage

```
ctgov_schema(schema = NULL)
```

Arguments

schema	the name of the schema. (Default is NULL - None)

Value

no return value; used for side effects

---

ctgov_text_similarity *Similarity Matrix*

Description

Takes one or more vectors of text and returns a similarity matrix.

Usage

```
ctgov_text_similarity(
  ...,
  max_terms = 10000,
  tolower = TRUE,
  min_df = 0,
  max_df = 1
)
```

Arguments

...	one or more vectors of text to search; must all be the same length
max_terms	maximum number of terms to consider for keywords
tolower	should keywords respect the case of the raw terms
min_df	minimum proportion of documents that a term should be present in to be included in the keywords
max_df	maximum proportion of documents that a term should be present in to be included in the keywords

Value

a distance matrix

---

ctgov_tfidf *TF-IDF Keywords*

Description

Takes one or more vectors of text and returns a vector of keywords.

Usage

```
ctgov_tfidf(
  ...,
  max_terms = 10000,
  tolower = TRUE,
  nterms = 5L,
  min_df = 0,
  max_df = 1
)
```

Arguments

...	one or more vectors of text to search; must all be the same length
max_terms	maximum number of terms to consider for keywords
tolower	should keywords respect the case of the raw terms
nterms	number of keyord terms to include
min_df	minimum proportion of documents that a term should be present in to be included in the keywords
max_df	maximum proportion of documents that a term should be present in to be included in the keywords

Value

a character vector of detected keywords

---

has_term *Does a Term Appear in a Vector of Strings?*

Description

Does a Term Appear in a Vector of Strings?

Usage

```
has_term(s, pattern, ignore_case = TRUE)
```

Arguments

s	the vector of strings.
pattern	the pattern to search for.
ignore_case	should the case be ignored? Default TRUE

Value

a single logical value

---

tbl_join_sample *Sample Clinical Trials Dataset*

Description

Data frame containing a 2.5 percent random sample of clinical trials.

Index