Package 'pkgmatch'

June 10, 2026

Title Find R Packages Matching Either Descriptions or Other R Packages

Version 0.5.4

Description Find R packages from CRAN, 'rOpenSci', or Bioconductor corpora. Packages can be matched to general text descriptions, to names of installed packages, or to local paths to entire source repositories. The package is used to list the most similar packages for each new submission to the 'rOpenSci' software peer-review program ('rOpenSci' authors, 2026; <doi:10.5281/zenodo.18885936>).

License MIT + file LICENSE

URL https://docs.ropensci.org/pkgmatch/, https://github.com/ropensci-review-tools/pkgmatch

BugReports https://github.com/ropensci-review-tools/pkgmatch/issues

Depends R (>= 4.1.0)

Imports brio, checkmate, cli, curl (>= 6.0.0), dplyr, fs, httr2, memoise, piggyback, Rcpp, rvest, tibble, tidyr, tokenizers, treesitter, treesitter.r, vctrs

Suggests gert, hms, httptest2, jsonlite, pkgbuild, rappdirs, roxygen2, testthat (>= 3.0.0), withr, knitr, rmarkdown

LinkingTo Rcpp

NeedsCompilation yes

Encoding UTF-8

Language en-GB

Config/testthat/edition 3

Config/ropensci/maintainer staff

VignetteBuilder knitr

Config/roxygen2/version 8.0.0

Author Mark Padgham [aut, cre] (ORCID: <https://orcid.org/0000-0003-2172-5265>), Davis Vaughan [ctb]

Maintainer Mark Padgham <mark.padgham@email.com>

Repository CRAN

Date/Publication 2026-06-10 07:40:02 UTC

Contents

`generate_pkgmatch_example_data`
Generate example data to use with pkgmatch

Description

This function generates a selection of test data for the "cran" corpus, to allow functions to be run offline, without having to download the large datasets otherwise required for the package to function.

Note that these data are randomly generated, and results will be generally meaningless. They are generated solely to demonstrate how the package functions, and are not intended to derive meaningful outputs.

Usage

```
generate_pkgmatch_example_data(corpus = "cran")
```

Arguments

`corpus`	One of "ropensci" or "cran", where "ropensci" generates additional data on function call frequencies.

Value

(Invisibly) The path to the temporary directory containing the package data.

See Also

Other utils: head.pkgmatch(), pkgmatch_browse(), pkgmatch_load_data(), pkgmatch_update_cache(), print.pkgmatch()

Examples

```
generate_pkgmatch_example_data ()
input <- "curl" # Name of a single installed package
pkgmatch_similar_pkgs (input, corpus = "cran")
```

---

head.pkgmatch *Head method for 'pkgmatch' objects*

Description

Head method for 'pkgmatch' objects

Usage

```
## S3 method for class 'pkgmatch'
head(x, n = 5L, ...)
```

Arguments

x	Object for which head is to be printed
n	Number of rows of full pkgmatch object to be displayed
...	Not used

Value

A (usually) smaller version of x, with all columns displayed.

See Also

Other utils: generate_pkgmatch_example_data(), pkgmatch_browse(), pkgmatch_load_data(), pkgmatch_update_cache(), print.pkgmatch()

Examples

```
corpus <- "cran"
generate_pkgmatch_example_data (corpus = corpus)
input <- "Download open spatial data from NASA"
p <- pkgmatch_similar_pkgs (input, corpus = corpus)
head (p) # Shows first 5 rows of full `data.frame` object
```

pkgmatch_bm25 *The "Best Matching 25" (BM25) ranking function.*

Description

BM25 values match single inputs to document corpora by weighting terms by their inverse frequencies, so that relatively rare words contribute more to match scores than common words. For each input, the BM25 value is the sum of relative frequencies of each term in the input multiplied by the Inverse Document Frequency (IDF) of that term in the entire corpus. See the Wikipedia page at https://en.wikipedia.org/wiki/Okapi_BM25 for further details.

Usage

```
pkgmatch_bm25(input, txt = NULL, idfs = NULL, corpus = NULL, minchar = 3L)
```

Arguments

input	A single character string to match against the second parameter of all input documents.
txt	An optional list of input documents. If not specified, data will be loaded as specified by the corpus parameter.
idfs	Optional list of Inverse Document Frequency weightings generated by the internal bm25_idf function. If not specified, values for the rOpenSci corpus will be automatically downloaded and used.
corpus	If txt is not specified, data for nominated corpus will be downloaded to local cache directory, and BM25 values calculated against those. Must be one of "ropensci", "ropensci-fns", "cran", or "bioc" (for BioConductor). Note that the "ropensci-fns" and "bioc_fns" corpora contain entries for every single function of every rOpenSci and BioConductor package, respectively, and the resulting BM25 values can be used to determine the best-matching function. The other two corpora are package-based, and the results can be used to find the best-matching package.
minchar	Minimal number of characters; tokens with less than this number are discarded.

Value

A data.frame of package names and 'BM25' measures against text from whole packages both with and without function descriptions.

See Also

Other bm25: pkgmatch_bm25_fn_calls()

Examples

```
# The following function simulates remote data in temporary directory, to
# enable package usage without downloading. Do not run for normal usage.
generate_pkgmatch_example_data ()

input <- "curl" # Name of a single installed package
pkgmatch_bm25 (input, corpus = "cran")
# Or pre-load document-frequency weightings and pass those:
idfs <- pkgmatch_load_data ("idfs", corpus = "cran", fns = FALSE)
# Those have token frequencies for both "full" text, and for descriptions
# only "desc_only":
pkgmatch_bm25 (input, corpus = "cran", idfs = idfs$full)
pkgmatch_bm25 (input, corpus = "cran", idfs = idfs$descs_only)
```

---

pkgmatch_bm25_fn_calls

The "Best Matching 25" (BM25) ranking function for function calls

---

Description

See ?pkgmatch_bm25 for details of BM25 ranks. This function calculates "BM25" ranks from function-call frequencies between a local R package and all packages in specified corpus. Values are thus higher for packages with similar patterns of function calls, weighted by inverse frequencies, so functions called infrequently across the entire corpus contribute more than common functions.

Note that the results of this function are entirely different from pkgmatch_bm25 with `corpus = "ropensci-fns"` or `corpus = "bioc-fns"`. The latter return BM25 values from text descriptions of all functions in all rOpenSci or BioConductor packages, whereas this function returns BM25 values based on frequencies of function calls within packages.

Usage

```
pkgmatch_bm25_fn_calls(path, corpus = NULL)
```

Arguments

path	Local path to source code of an R package.
corpus	One of "ropensci" or "cran"

Value

A data.frame of two columns:

- "package" Naming the package from the specified corpus;
- bm25 The "BM25" index value for the nominated packages, where high values indicate greater overlap in term frequencies.

See Also

Other bm25: pkgmatch_bm25()

Examples

```
corpus <- "ropensci"
flist <- generate_pkgmatch_example_data (corpus = corpus)

pkgmatch_bm25_fn_calls (path = "cli", corpus = corpus)
```

---

pkgmatch_browse *Open web pages for* pkgmatch *results*

Description

Open web pages for pkgmatch results

Usage

```
pkgmatch_browse(p, n = NULL)
```

Arguments

p	A pkgmatch object returned from pkgmatch_similar_pkgs.
n	Number of top-matching entries which should be opened. Defaults to the value passed to the main functions.

Value

(Invisibly) A named vector of integers, with 0 for all pages able to be successfully opened, and 1 otherwise.

See Also

Other utils: generate_pkgmatch_example_data(), head.pkgmatch(), pkgmatch_load_data(), pkgmatch_update_cache(), print.pkgmatch()

Examples

```
input <- "genomics and transcriptomics sequence data"

p <- pkgmatch_similar_pkgs (input, corpus = "ropensci")

## Not run:
pkgmatch_browse (p) # Open main package pages on rOpenSci

## End(Not run)
```

```
p <- pkgmatch_similar_pkgs (input, corpus = "cran")

## Not run:
pkgmatch_browse (p) # Open main package pages on CRAN

## End(Not run)
```

`pkgmatch_load_data` *Load 'pkgmatch' data for specified corpus.*

Description

Load pre-computed data for a specified corpus. Data types are:

- "idfs" for Inverse Document Frequency weightings;
- "functions" for frequency tables for text descriptions of function calls; or
- "calls" for frequency tables for actual function calls.

This function is called within the main pkgmatch_similar_pkgs function to load required data there, and should not generally need to be explicitly called.

Usage

```
pkgmatch_load_data(what = "idfs", corpus = "ropensci", fns = FALSE)
```

Arguments

`what`	One of the three data types described above: "idfs", "functions", or "calls".
`corpus`	Must be specified as one of "ropensci", "cran", or "bioc" (for BioConductor). If `idfs` parameter is not specified, data will be automatically downloaded for the corpus specified by this parameter. The function will then return the most similar package from the specified corpus. Note that calculations will `corpus = "cran"` will generally take longer, because the corpus is much larger.
`fns`	If `FALSE` (default), load data for all packages; otherwise load (considerably larger dataset of) data for all individual functions.

Value

The loaded data.

See Also

Other utils: `generate_pkgmatch_example_data()`, `head.pkgmatch()`, `pkgmatch_browse()`, `pkgmatch_update_cache()`, `print.pkgmatch()`

Examples

```
corpus <- "cran"
generate_pkgmatch_example_data (corpus = corpus)
idfs <- pkgmatch_load_data ("idfs", corpus = corpus)
idfs_fns <- pkgmatch_load_data ("idfs", fns = TRUE, corpus = corpus)
```

---

pkgmatch_similar_fns *Identify R functions best matching a given input string*

Description

Function matching is only available for functions from the corpora of rOpenSci or Bioconductor packages, and not for CRAN packages.

Usage

```
pkgmatch_similar_fns(input, corpus = "ropensci", n = 5L, browse = FALSE)
```

Arguments

input	A text string.
corpus	One of "ropensci" or "bioc" (for BioConductor). It is not possible to match functions again CRAN packages.
n	When the result of this function is printed to screen, the top n packages will be displayed.
browse	If TRUE, automatically open webpages of the top n matches in local browser.

Value

A modified data.frame object of class "pkgmatch". The data.frame has 3 columns:

1. "pkg_fn" with the name of the function in the form "package::function";
2. "simil" with a similarity score between 0 and 1; and
3. "rank" as an integer index, with the highest rank of 1 as the first row.

The return object has a default print method which prints the names only of the first 5 best matching functions; see ?print.pkgmatch for details.

See Also

Other main: pkgmatch_similar_pkgs()

Examples

```
corpus <- "ropensci"
generate_pkgmatch_example_data (corpus = corpus)
input <- "Process raster satellite images"
p <- pkgmatch_similar_fns (input, corpus = corpus)
p # Default print method, lists 5 best matching functions
head (p) # Shows first 5 rows of full `data.frame` object
```

`pkgmatch_similar_pkgs` *Find R packages matching an input of either text or another package*

Description

This function accepts as `input` either a text description, the name of a locally-installed package, or a path to a local directory containing an R package. It ranks all R packages within the specified corpus in terms of how well they match that input. The "corpus" argument can specify either rOpenSci's package suite, CRAN, or Bioconductor.

Ranks are obtained from scores derived from "Best Match 25" (BM25) scores based on document token frequencies.

Ranks are generally obtained by matching both for full package text from the specified corpus, including all long-form documentation, and by matching package descriptions only. The function returns a single rank derived by combining individual ranks using the Reciprocal Rank Fusion (RRF) algorithm.

Finally, all components of this function are locally cached for each call (by the **memoise** package), so additional calls to this function with the same `input` and `corpus` should be much faster than initial calls.

Usage

```
pkgmatch_similar_pkgs(
  input,
  corpus = NULL,
  idfs = NULL,
  n = 5L,
  browse = FALSE
)
```

Arguments

`input`	Either a text string, a path to local source code of an R package, or the name of any installed R package.
`corpus`	Must be specified as one of "ropensci", "cran", or "bioc" (for BioConductor). If `idfs` parameter is not specified, data will be automatically downloaded for the corpus specified by this parameter. The function will then return the most similar package from the specified corpus. Note that calculations will `corpus = "cran"` will generally take longer, because the corpus is much larger.

idfs	Inverse Document Frequency tables for a suite of packages, generated from pkgmatch_bm25. If not provided, pre-generated IDF tables will be downloaded and stored in a local cache directory.
n	When the result of this function is printed to screen, the top n packages will be displayed.
browse	If TRUE, automatically open webpages of the top n matches in local browser.

Value

A data.frame with a "package" column naming packages, and a column of package ranks, with 1 being most similar. For the CRAN corpus, a column of package versions is also included.

The returned object has a default print method which prints the best 5 matches directly to the screen, yet returns information on all packages within the specified corpus. There is also a head method to print the first few entries of these full data (default n = 5). To see all data, use as.data.frame().

Note

The first time this function is run without passing idfs, required values will be automatically downloaded and stored in a locally persistent cache directory. Especially for the "cran" corpus, this downloading may take quite some time.

See Also

Other main: pkgmatch_similar_fns()

Examples

```
# The following function simulates remote data in temporary directory, to
# enable package usage without downloading. Do not run for normal usage.
generate_pkgmatch_example_data ()

input <- "curl" # Name of a single installed package
p <- pkgmatch_similar_pkgs (input, corpus = "cran")
p # Default print method, lists 5 best matching packages
head (p) # Shows first 5 rows of full `data.frame` object
```

---

pkgmatch_treesitter_fn_tags

Identify all function calls made within a package.

Description

This function uses "treesitter" (https://github.com/tree-sitter/tree-sitter) to tag all function calls made within a local package, and to associate those calls with package namespaces.

This is used as input to the pkgmatch_bm25_fn_calls function, to enable function calls within a local package to be inversely weighted by frequencies within all packages within a corpus. The results of applying this function to the full corpora used in this package are contained within the data listed on https://github.com/ropensci-review-tools/pkgmatch/releases/tag/v0.5.2, as "fn-calls-ropensci.Rds" and "fn-calls-cran.Rds".

Usage

```
pkgmatch_treesitter_fn_tags(path)
```

Arguments

path	Path to local package, or .tar.gz file of package source.

Value

A data.frame of all function calls made within the package, with the following columns:

- 'fn' Name of the package function within which call is made, including namespace identifiers of "::" for exported functions and ":::" for non-exported functions.
- name Name of function being called, including namespace.
- start Byte number within file corresponding to start of definition
- end Byte number within file corresponding to end of definition
- file Name of file in which fn call is defined.

Examples

```
# Get function calls made within locally-installed packages:
fn_tags <- pkgmatch_treesitter_fn_tags ("curl") # Name of installed package
fn_tags <- pkgmatch_treesitter_fn_tags ("cli") # Name of installed package

# Or get calls from full source code:
u <- "https://cran.r-project.org/src/contrib/Archive/odbc/odbc_1.5.0.tar.gz"
path <- file.path (tempdir (), basename (u))

download.file (u, destfile = path, quiet = TRUE)
fn_tags <- pkgmatch_treesitter_fn_tags (path)
```

---

pkgmatch_update_cache *Update all locally-cached* pkgmatch *data to latest versions.*

Description

This function forces all locally-cached data to be updated with latest version of remote data provided on the latest release of GitHub repository at https://github.com/ropensci-review-tools/pkgmatch/releases.

Caching strategies are described in the "*Data Caching and Updating*" vignette, accessible either locally via vignette("data-caching-and-updating", package = "pkgmatch"), or online at https://docs.ropensci.org/pkgmatch/articles/B_data-caching-and-updating.html. In short, locally-cached data used by this package are updated by default every 30 days (with the vignette describing how to modify this default behaviour). This function forces all locally-cached data to be updated, regardless of update frequencies.

Usage

```
pkgmatch_update_cache()
```

Value

(Invisibly) A list of full local paths to all files which were updated.

See Also

Other utils: generate_pkgmatch_example_data(), head.pkgmatch(), pkgmatch_browse(), pkgmatch_load_data(), print.pkgmatch()

Examples

```
## Not run:
pkgmatch_update_cache ()

## End(Not run)
```

---

pkgmatch_update_data *Update pkgmatch corpus data on GitHub*

Description

This function is intended for internal rOpenSci use only. Usage by any unauthorized users will error and have no effect unless run with `upload = FALSE`, in which case updated data will be created in the sub-directory "pkgmatch-results" of R's current temporary directory. This updating may take a very long time!

The function does not update the BioConductor data. Because those are fixed to specific BioConductor releases, they are only updated manually with the internal `pkgmatch_generate_bioc()` function.

Note that this function is categorically different from pkgmatch_update_cache. This function updates the internal data used by the `pkgmatch` package, and should only ever be run by package maintainers. The pkgmatch_update_cache downloads the latest versions of these data to a local cache for use in this package.

Usage

```
pkgmatch_update_data(
  upload = TRUE,
  local_cran_mirror = NULL,
  local_ropensci_mirror = NULL
)
```

Arguments

upload	If TRUE, upload updated results to GitHub release.
local_cran_mirror	Optional path to a local directory with full CRAN mirror. If specified, data will use packages from this local source for updating. Default behaviour if not specified is to download new packages into tempdir, and delete once data have been updated.
local_ropensci_mirror	Optional path to a local directory with full rOpenSci mirror. If specified, data will use repositories from this local source for updating. Default behaviour if not specified is to clone new repositories into tempdir, and delete once data have been updated.

Value

Local path to directory containing updated results.

Examples

```
## Not run:
pkgmatch_update_data (upload = FALSEE)

## End(Not run)
```

---

print.pkgmatch *Print method for 'pkgmatch' objects*

Description

The main pkgmatch function, pkgmatch_similar_pkgs, returns data.frame objects of class "pkgmatch". This class exists primarily to enable this print method, which summarises by default the top 5 matching packages or functions. Objects can be converted to standard data.frames with as.data.frame().

Usage

```
## S3 method for class 'pkgmatch'
print(x, ...)
```

Arguments

x	Object to be printed
...	Additional parameters passed to default 'print' method.

Value

The result of printing x, in form of either a single character vector, or a named list of character vectors.

See Also

Other utils: generate_pkgmatch_example_data(), head.pkgmatch(), pkgmatch_browse(), pkgmatch_load_data(), pkgmatch_update_cache()

Examples

```
corpus <- "cran"
generate_pkgmatch_example_data (corpus = corpus)
input <- "Download open spatial data from NASA"
p <- pkgmatch_similar_pkgs (input, corpus = corpus)
head (p) # Shows first 5 rows of full `data.frame` object
p # Default print method, lists 5 best matching packages
```

Index