

Package: pkgmatch (via r-universe)

February 13, 2025

Title Find R Packages Matching Either Descriptions or Other R Packages

Version 0.4.3.027

Description Find R packages matching either descriptions or other R packages.

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>

Requires R (>= 4.1.0)

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

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

LinkingTo Rcpp

Depends R (>= 3.5.0)

NeedsCompilation yes

Encoding UTF-8

Language en-GB

Roxygen list(markdown = TRUE)

RoxygenNote 7.3.2

Config/testthat/edition 3

VignetteBuilder knitr

Config/pak/sysreqs make libicu-dev libxml2-dev libssl-dev

Repository <https://ropensci.r-universe.dev>

RemoteUrl <https://github.com/ropensci-review-tools/pkgmatch>

RemoteRef main

RemoteSha c0f035b0b6c4e250b5c3419affad0b13b215792

Contents

get_ollama_url	2
head.pkgmatch	3
ollama_check	4
pkgmatch_bm25	4
pkgmatch_bm25_fn_calls	5
pkgmatch_browse	6
pkgmatch_embeddings_from_pkgs	7
pkgmatch_embeddings_from_text	8
pkgmatch_load_data	9
pkgmatch_similar_fns	10
pkgmatch_similar_pkgs	11
pkgmatch_treesitter_fn_tags	13
pkgmatch_update_cache	14
pkgmatch_update_data	14
print.pkgmatch	15
set_ollama_url	16
text_is_code	16
Index	18

get_ollama_url	<i>Get the URL for local ollama API</i>
----------------	---

Description

Return the URL of the specified ollama API. Default is "127.0.0.1:11434"

Usage

```
get_ollama_url()
```

Value

The ollama API URL

See Also

set_ollama_url

Other ollama: [ollama_check\(\)](#), [set_ollama_url\(\)](#)

head.pkgmatch	<i>Head method for 'pkgmatch' objects</i>
---------------	---

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: [pkgmatch_browse\(\)](#), [pkgmatch_load_data\(\)](#), [pkgmatch_update_cache\(\)](#), [print.pkgmatch\(\)](#), [text_is_code\(\)](#)

Examples

```
## Not run:  
input <- "Download open spatial data from NASA"  
p <- pkgmatch_similar_pkgs (input)  
p # Default print method, lists 5 best matching packages  
head (p) # Shows first 5 rows of full `data.frame` object  
  
## End(Not run)
```

ollama_check	<i>Check that ollama is installed with required models, and download if not.</i>
--------------	--

Description

Note that the URL of a locally-running ollama instance is presumed by default to be "127.0.0.1:11434". Other values can be set using the [set_ollama_url](#) function.

Usage

```
ollama_check(sudo = is_docker_sudo())
```

Arguments

sudo Set to TRUE if ollama is running in docker with sudo privileges.

Value

TRUE if everything works okay, otherwise the function will error before returning.

See Also

Other ollama: [get_ollama_url\(\)](#), [set_ollama_url\(\)](#)

Examples

```
## Not run:
chk <- ollama_check ()

## End(Not run)
```

pkgmatch_bm25	<i>Calculate the "BM25" = "Best Matching 25" ranking function between text input and all R packages within specified corpus.</i>
---------------	--

Description

See https://en.wikipedia.org/wiki/Okapi_BM25.

Usage

```
pkgmatch_bm25(input, txt = NULL, idfs = NULL, corpus = "ropensci")
```

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.
ids	Optional list of Inverse Document Frequency weightings generated by the internal <code>bm25_idf</code> function. If not specified, values for the <code>rOpenSci</code> corpus will be automatically downloaded and used.
corpus	If <code>txt</code> 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", or "cran". Note that the "ropensci-fns" corpus contains entries for every single function of every <code>rOpenSci</code> package, 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.

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

```
## Not run:
input <- "Download open spatial data from NASA"
bm25 <- pkgmatch_bm25 (input)
# Or pre-load document-frequency weightings:
ids <- pkgmatch_load_data ("ids", fns = FALSE)
bm25 <- pkgmatch_bm25 (input, ids = ids)

## End(Not run)
```

`pkgmatch_bm25_fn_calls`

Calculate a "BM25" index from function-call frequencies between a local R package and all packages in specified corpus.

Description

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

Usage

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

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

```
## Not run:  
u <- "https://cran.r-project.org/src/contrib/odbc_1.5.0.tar.gz"  
path <- file.path(tempdir(), basename(u))  
download.file(u, destfile = path)  
bm25 <- pkgmatch_bm25_fn_calls(path)  
  
## End(Not run)
```

pkgmatch_browse	<i>Open web pages for pkgmatch results</i>
-----------------	--

Description

Open web pages for pkgmatch results

Usage

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

Arguments

p	A pkgmatch object returned from either pkgmatch_similar_pkgs or pkgmatch_similar_fns .
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: [head.pkgmatch\(\)](#), [pkgmatch_load_data\(\)](#), [pkgmatch_update_cache\(\)](#), [print.pkgmatch\(\)](#), [text_is_code\(\)](#)

Examples

```
## Not run:
input <- "genomics and transcriptomics sequence data"
p <- pkgmatch_similar_pkgs (input)
pkgmatch_browse (p) # Open main package pages on rOpenSci
p <- pkgmatch_similar_pkgs (input, corpus = "cran")
pkgmatch_browse (p) # Open main package pages on CRAN
p <- pkgmatch_similar_fns (input)
pkgmatch_browse (p) # Open pages for best-matching rOpenSci functions

## End(Not run)
```

pkgmatch_embeddings_from_pkgs

Return raw language model ('LM') embeddings from package text and function definitions.

Description

The embeddings are currently retrieved from a local 'ollama' server running Jina AI embeddings.

Usage

```
pkgmatch_embeddings_from_pkgs(packages = NULL, functions_only = FALSE)
```

Arguments

packages A vector of local paths to directories containing R packages.

functions_only If TRUE, calculate embeddings for function descriptions only. This is intended to generate a separate set of embeddings which can then be used to match plain-text queries of functions, rather than entire packages.

Value

If !functions_only, a list of two matrices of embeddings: one for the text descriptions of the specified packages, including individual descriptions of all package functions, and one for the entire code base. For functions_only, a single matrix of embeddings for all function descriptions.

See Also

Other embeddings: [pkgmatch_embeddings_from_text\(\)](#)

Examples

```
## Not run:
packages <- c("cli", "fs")
emb_fns <- pkgmatch_embeddings_from_pkgs (packages, functions_only = TRUE)
colnames (emb_fns) # All functions of the two packages
emb_pkg <- pkgmatch_embeddings_from_pkgs (packages, functions_only = FALSE)
names (emb_pkg) # text_with_fns, text_wo_fns, code
colnames (emb_pkg$text_with_fns) # cli, fs

## End(Not run)
```

pkgmatch_embeddings_from_text

Return raw language model ('LM') embeddings from a vector of text strings.

Description

The embeddings are currently retrieved from a local 'ollama' server running Jina AI embeddings.

Usage

```
pkgmatch_embeddings_from_text(input = NULL)
```

Arguments

input A vector of one or more text strings for which embeddings are to be extracted.

Value

A matrix of embeddings, one column for each input item, and a fixed number of rows defined by the embedding length of the language models.

See Also

Other embeddings: [pkgmatch_embeddings_from_pkgs\(\)](#)

Examples

```
## Not run:
input <- "Download open spatial data from NASA"
emb <- pkgmatch_embeddings_from_text (input = input)

## End(Not run)
```

pkgmatch_load_data	<i>Load embeddings generated by the pkgmatch_embeddings_from_pkgs function, either for all rOpenSci packages or, if fns = TRUE, all individual functions within those packages.</i>
--------------------	---

Description

Load embeddings generated by the [pkgmatch_embeddings_from_pkgs](#) function, either for all rOpenSci packages or, if fns = TRUE, all individual functions within those packages.

Usage

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

Arguments

what	One of: <ul style="list-style-type: none">• "embeddings" to load pre-generated embeddings;• "idfs" to load pre-generated Inverse Document Frequency weightings;• "functions" to load pre-generated frequency tables for text descriptions of function calls; or• "calls" to load pre-generated frequency tables for actual function calls.
corpus	If embeddings or idfs parameters are not specified, they are automatically downloaded for the corpus specified by this parameter. Must be one of "ropensci" or "cran". The function will then return the most similar package from the specified corpus.
fns	If FALSE (default), load embeddings for all rOpenSci packages; otherwise load (considerably larger dataset of) embeddings for all individual functions.
raw	Only has effect of what = "calls", in which case default of FALSE loads single Inverse Document Frequency table to entire corpus; otherwise if TRUE, loads raw function call counts for each package in corpus.

Value

The loaded data.frame.

See Also

Other utils: [head.pkgmatch\(\)](#), [pkgmatch_browse\(\)](#), [pkgmatch_update_cache\(\)](#), [print.pkgmatch\(\)](#), [text_is_code\(\)](#)

Examples

```
## Not run:
embeddings <- pkgmatch_load_data ("embeddings")
embeddings_fns <- pkgmatch_load_data ("embeddings", fns = TRUE)
idfs <- pkgmatch_load_data ("idfs")
idfs_fns <- pkgmatch_load_data ("idfs", fns = TRUE)

## End(Not run)
```

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

Description

Function matching is only available for functions from the corpus of rOpenSci packages.

Usage

```
pkgmatch_similar_fns(input, embeddings = NULL, n = 5L, browse = FALSE)
```

Arguments

input	A text string.
embeddings	Large Language Model embeddings for all rOpenSci packages, generated from pkgmatch_embeddings_from_pkgs . If not provided, pre-generated embeddings 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 character vector of function names in the form "::".

See Also

Other main: [pkgmatch_similar_pkgs\(\)](#)

Examples

```
## Not run:
input <- "Process raster satellite images"
p <- pkgmatch_similar_fns (input)
p # Default print method, lists 5 best matching packages
head (p) # Shows first 5 rows of full `data.frame` object

## End(Not run)
```

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, or a path to a local R package, and returns information on R packages which best match that input. Matches are found from within a specified "corpus", currently all packages from either [rOpenSci's package suite](#), or from [CRAN](#).

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. This information is in the form of a `data.frame`, with one column for the package name, and one or more additional columns of integer ranks for each package. 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()`.

Ranks are obtained from scores derived from:

- Cosine similarities between Language Model (LM) embeddings for the input, and corresponding embeddings for the specified corpus.
- **"Best Match 25" (BM25)** scores based on document token frequencies.

Ranks for text matches are generally obtained from packages both including and excluding function descriptions as part of the package text. This results in up to four scores for each input. These scores are then combined to a final ranking using the [Reciprocal Rank Fusion \(RRF\) algorithm](#). The additional parameter of `lm_proportion` determines the extent to which the final ranking weights the LM versus BM25 components.

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. This means the effect of changing `lm_proportion` can easily be examined by simply repeating calls to this function.

Usage

```
pkgmatch_similar_pkgs(  
  input,  
  corpus = "ropensci",  
  embeddings = NULL,  
  idfs = NULL,  
  input_is_code = text_is_code(input),  
  lm_proportion = 0.5,  
  n = 5L,  
  browse = FALSE  
)
```

Arguments

`input` Either a path to local source code of an R package, or a text string.

corpus	If embeddings or idfs parameters are not specified, they are automatically downloaded for the corpus specified by this parameter. Must be one of "ropensci" or "cran". The function will then return the most similar package from the specified corpus.
embeddings	Large Language Model embeddings for all rOpenSci packages, generated from pkgmatch_embeddings_from_pkgs . If not provided, pre-generated embeddings will be downloaded and stored in a local cache directory.
idfs	Inverse Document Frequency tables for all rOpenSci packages, generated from pkgmatch_bm25 . If not provided, pre-generated IDF tables will be downloaded and stored in a local cache directory.
input_is_code	A binary flag indicating whether input is code or plain text. Ignored if input is path to a local package; otherwise can be used to force appropriate interpretation if input type.
lm_proportion	A value between 0 and 1 to control the relative contributions of results from Language Models ("LMs") versus results from traditional token-frequency models. Final rankings are generated by combining these two kinds of results, so that <code>lm_proportion = 0</code> will return results from token frequency analyses only, while <code>lm_proportion = 1</code> will return results from LMs only.
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 one or more columns of package ranks in terms of text similarity and, if `input` is a local path to an entire R package, of similarity in code structure. As described above, the default print method prints package names only. To see full result, use `as.data.frame()`.

Note

The first time this function is run without passing either `embeddings` or `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

`input_is_code`

Other main: [pkgmatch_similar_fns\(\)](#)

Examples

```
## Not run:
input <- "Download open spatial data from NASA"
p <- pkgmatch_similar_pkgs (input)
p # Default print method, lists 5 best matching packages
head (p) # Shows first 5 rows of full `data.frame` object
# This second call will be much faster than first call:
```

```
p2 <- pkgmatch_similar_pkgs (input, lm_proportion = 0.25)

## End(Not run)
```

pkgmatch_treesitter_fn_tags

Use "treesitter" to tag all function calls made within local package, and to associate those calls with package namespaces. This is used as input to the [pkgmatch_bm25_fn_calls](#) function.

Description

Use "treesitter" to tag all function calls made within local package, and to associate those calls with package namespaces. This is used as input to the [pkgmatch_bm25_fn_calls](#) function.

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

```
## Not run:
u <- "https://cran.r-project.org/src/contrib/odbc_1.5.0.tar.gz"
path <- file.path(tempdir(), basename(u))
download.file(u, destfile = path)
tags <- pkgmatch_treesitter_fn_tags(path)

## End(Not run)
```

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

Description

Locally-cached data used by this package are generally updated every 30 days, or according to `options("pkgmatch_update_frequency")` if that is set. This function forces all locally-cached data to be updated, regardless of update frequencies. Remote data are provided on the latest release of GitHub repository at <https://github.com/ropensci-review-tools/pkgmatch/releases>.

Usage

```
pkgmatch_update_cache()
```

See Also

Other utils: [head.pkgmatch\(\)](#), [pkgmatch_browse\(\)](#), [pkgmatch_load_data\(\)](#), [print.pkgmatch\(\)](#), [text_is_code\(\)](#)

Examples

```
## Not run:  
pkgmatch_update_cache ()  
  
## End(Not run)
```

pkgmatch_update_data *Update pkgmatch' data for both CRAN and rOpenSci packages on GitHub release*

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!

Usage

```
pkgmatch_update_data(upload = TRUE)
```

Arguments

upload If TRUE, upload updated results to GitHub release.

Details

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.

Value

Local path to directory containing updated results.

Examples

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

## End(Not run)
```

print.pkgmatch	<i>Print method for 'pkgmatch' objects</i>
----------------	--

Description

Print method for 'pkgmatch' objects

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: [head.pkgmatch\(\)](#), [pkgmatch_browse\(\)](#), [pkgmatch_load_data\(\)](#), [pkgmatch_update_cache\(\)](#), [text_is_code\(\)](#)

Examples

```
## Not run:
input <- "Download open spatial data from NASA"
p <- pkgmatch_similar_pkgs (input)
p # Default print method, lists 5 best matching packages
head (p) # Shows first 5 rows of full `data.frame` object

## End(Not run)
```

set_ollama_url	<i>Set the URL for local ollama API</i>
----------------	---

Description

Set the URL for local ollama API

Usage

```
set_ollama_url(ollama_url)
```

Arguments

ollama_url The desired ollama API URL

Value

The ollama API URL

See Also

[get_ollama_url\(\)](#)

Other ollama: [get_ollama_url\(\)](#), [ollama_check\(\)](#)

text_is_code	<i>Estimate whether input text string is code or English prose text.</i>
--------------	--

Description

This is only approximate, and there are even software packages which can give false negatives and be identified as prose (like rOpenSci's "geonames" package), and prose which may be wrongly identified as code.

Usage

```
text_is_code(txt)
```


Arguments

txt Single input text string

Value

Logical value indicating whether or not txt was identified as code.

See Also

Other utils: [head.pkgmatch\(\)](#), [pkgmatch_browse\(\)](#), [pkgmatch_load_data\(\)](#), [pkgmatch_update_cache\(\)](#), [print.pkgmatch\(\)](#)

Examples

```
txt <- "Some text without any code"
text_is_code (txt)
txt <- "this_is_code <- function (x) { x }"
text_is_code (txt)
```

Index

- * **bm25**
 - pkgmatch_bm25, [4](#)
 - pkgmatch_bm25_fn_calls, [5](#)
 - * **data**
 - pkgmatch_update_data, [14](#)
 - * **embeddings**
 - pkgmatch_embeddings_from_pkgs, [7](#)
 - pkgmatch_embeddings_from_text, [8](#)
 - * **main**
 - pkgmatch_similar_fns, [10](#)
 - pkgmatch_similar_pkgs, [11](#)
 - * **ollama**
 - get_ollama_url, [2](#)
 - ollama_check, [4](#)
 - set_ollama_url, [16](#)
 - * **treесitter**
 - pkgmatch_treesitter_fn_tags, [13](#)
 - * **utils**
 - head.pkgmatch, [3](#)
 - pkgmatch_browse, [6](#)
 - pkgmatch_load_data, [9](#)
 - pkgmatch_update_cache, [14](#)
 - print.pkgmatch, [15](#)
 - text_is_code, [16](#)
- [get_ollama_url](#), [2](#), [4](#), [16](#)
[get_ollama_url\(\)](#), [16](#)
- [head.pkgmatch](#), [3](#), [7](#), [9](#), [14](#), [15](#), [17](#)
- [ollama_check](#), [2](#), [4](#), [16](#)
- [pkgmatch_bm25](#), [4](#), [5](#), [6](#), [12](#)
[pkgmatch_bm25_fn_calls](#), [5](#), [5](#), [13](#)
[pkgmatch_browse](#), [3](#), [6](#), [9](#), [14](#), [15](#), [17](#)
[pkgmatch_embeddings_from_pkgs](#), [7](#), [8–10](#),
[12](#)
[pkgmatch_embeddings_from_text](#), [8](#), [8](#)
[pkgmatch_load_data](#), [3](#), [7](#), [9](#), [14](#), [15](#), [17](#)
[pkgmatch_similar_fns](#), [6](#), [10](#), [12](#)
[pkgmatch_similar_pkgs](#), [6](#), [10](#), [11](#)
[pkgmatch_treesitter_fn_tags](#), [13](#)
[pkgmatch_update_cache](#), [3](#), [7](#), [9](#), [14](#), [15](#), [17](#)
[pkgmatch_update_data](#), [14](#)
[print.pkgmatch](#), [3](#), [7](#), [9](#), [14](#), [15](#), [17](#)
[set_ollama_url](#), [2](#), [4](#), [16](#)
[text_is_code](#), [3](#), [7](#), [9](#), [14](#), [15](#), [16](#)