Package 'goldi'

June 28, 2017

Type Package

Title Gene Ontology Label Discernment and Identification

Version 1.0.1

Date 2017-06-27

Encoding UTF-8

Maintainer Christopher B. Cole <chris.c.1221@gmail.com>

Description A tool for identifying multiple word key terms in free text with application to Gene Ontology labels.

LinkingTo Rcpp, RcppArmadillo

Depends R (>= 2.15.0)

Imports dplyr, Rcpp, tm, SnowballC, magrittr, futile.logger

Suggests testthat, covr, rmarkdown, knitr, pdftools, RISmed

License MIT + file LICENSE

BugReports https://github.com/Chris1221/goldi/issues

URL https://github.com/Chris1221/goldi

LazyData TRUE

RoxygenNote 6.0.1

VignetteBuilder knitr

NeedsCompilation yes

Author Christopher B. Cole [aut, cre, cph],
Sejal Patel [ctb],
Jo Knight [ctb]

Repository CRAN

Date/Publication 2017-06-28 15:06:20 UTC

R topics documented:

`enrichment` *Calculate enriched terms in a target set*

Description

Given a target set of articles under question, we wish to compare the frequencies of term occurance to a control set of articles. We set a threshold above which to investigate terms (setting this threshold higher reduces spurious associated but decreases power to identify true associations) and calculate the enrichment and the P value of association (see `hgt` for more details).

Usage

```
enrichment(target, control, threshold, correction = "fdr")
```

Arguments

`target`	Set of `goldi` output for a target set of articles.
`control`	Set of `goldi` output for a control set of articles.
`threshold`	Only investigate associations which have been founder greater than this number of times.
`correction`	Correction to impliment on association P values. Users may choose any value which `p.adjust` may accept.

Details

This function mimics a truncated version of the output of GOrilla by identifying and quantifying the enrichment of terms in a target set of articles. Given N articles and B associations of the given term, the enrichment of b terms in the n articles in the target set is given by $\frac{\frac{b}{n}}{\frac{B}{N}}$.

Value

A formated data.frame with three columns: terms, enrichments, and P values.

References

Eden, E., Navon, R., Steinfeld, I., Lipson, D., & Yakhini, Z. (2009). GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. BMC Bioinformatics, 10(1), 1–7. http://doi.org/10.1186/1471-2105-10-48

`goldi` *Identify terms present in document.*

Description

This function takes as input a document which the user wishes to mine, a list of terms which they wish to identify, and an acceptance function for deciding on associations. This is the main function of the package; all others are helper functions, exported for your convenience. For full instructions on this function's usage, please see the documentation at github.com/Chris1221/goldi, or read the associated publication. We recommend it as background regardless.

Usage

```
goldi(doc,
  terms = "You must put your terms here if not using a precomputed TDM.",
  lims = c(1, 2, 3, 3, 4, 5, 6, 6, 7, 8, 8), output, syn = FALSE,
  syn.list = NULL, object = FALSE, log = NULL, reader = "local",
  term_tdm = NULL, log.level = "warn")
```

Arguments

`doc`	Either a file path to a document which will be read in, or a string already read into R. See `"reader"` for more details. Depending on the `"reader"` selected, there are four options for document input.
`terms`	Either a character vector of terms, with each element being a separate term, or a file path to a newline seperated text document which may be parsed into terms.
`lims`	Number of identical (or synonymous) words which must be present in a sentence in order for it to be accepted as a match for the term. "interactive" is default and allows you to interavtively build your own list, but a list or vector of n elements can be supplied where n is the largest term you wish to search for.
`output`	path to output file
`syn`	If you would like to use synonyms, set `"syn = TRUE"` with `"syn.list"` left as default to launch the interactive generator (`"goldi::make.syn()"`), or give a list if synonyms are already formatted.
`syn.list`	LIST of synonyms to be used. First element of each list item is the word that will counted if any of the other elements of that list item are present.
`object`	Return as an R object?
`log`	If specified, the path to the log you wish to keep.
`reader`	Option for how to read in the text files. See details.
`term_tdm`	If using a precompiled TDM.
`log.level`	Logging level. See `?flog.threshold` for details.

Value

A data frame of terms and their context within the document.

Author(s)

Christopher B. Cole <chris.c.1221@gmail.com>

References

See ArXiv prepubliation.

Examples

```
## Not run:

# Give the free form text
doc <- "In this sentence we will talk about ribosomal chaperone activity."

# Load in the included term document matrix for the terms
data("TDM.go.df")

# Pipe output and log to /dev/null
output = "/dev/null"
log = "/dev/null"

# Run the function
goldi(doc = doc,
      term_tdm = TDM.go.df,
      output = output,
      log = log,
      object = TRUE)


## End(Not run)
```

---

hgt *Calculate the hypergeometric tail*

Description

Given N articles, B of which are annotated to a given term, the chance that .b of these articles are annotated in a test set of size .n is equal to the hypergeometric tail function.

Usage

```
hgt(.b, N, B, .n)
```

Arguments

.b	Number of annotations in target group
N	Total number of articles
B	Total number of annotations
.n	Number of articles in target group

Details

P value is computed as in referenced article (GOrilla). Briefly, the P value is the sum from .b to the minimum of .n and B of .n choose i plus N-.n choose B - i all divided by N choose B.

Value

P value of the hypergeometric distribution.

References

Eden, E., Navon, R., Steinfeld, I., Lipson, D., & Yakhini, Z. (2009). GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. BMC Bioinformatics, 10(1), 1–7. http://doi.org/10.1186/1471-2105-10-48

---

make.lim *Construct Constraint Limits for Fuzzy Identification*

Description

Construct Constraint Limits for Fuzzy Identification

Usage

```
make.lim(int = TRUE, list = NULL, length = 10)
```

Arguments

int	interactive (TRUE) or specified list (FALSE)
list	preconstructed list if i is FALSE
length	number of parameters you wish to enter. Defaults to 5.

make.syn *Construct synonym matrix for internal or external use*

Description

Construct synonym matrix for internal or external use

Usage

```
make.syn(return = FALSE)
```

Arguments

return	Return value (T) or stout (F). Defaults to T.

match *Match terms*

Description

Match terms in C++

Usage

```
match(term_vector, pdf_tdm, term_tdm, thresholds, pdf_index, terms, sentences)
```

Arguments

term_vector	Index vector of where each of the terms is in the pdf_tdm. i.e. the ith element of term_vector is j. Therefor, term i is at column j in the pdf_tdm.
pdf_tdm	Term document matrix of words in the PDF
term_tdm	Term document matrix of words in the terms and pdf sentences.
thresholds	Acceptance thresholds
pdf_index	Index of terms in PDF
terms	List of terms used, this is the vector of column names of term_tdm.
sentences	Vector of sentences read in from the PDf.

Value

List of matched terms.

MF_terms *Molecular Function terms A dataset of molecular function terms.*

Description

Molecular Function terms A dataset of molecular function terms.

Usage

```
MF_terms
```

Format

An object of class `data.table` (inherits from `data.frame`) with 10834 rows and 1 columns.

replaceExpressions *Internal expression replacement with grep*

Description

Internal expression replacement with grep

Usage

```
replaceExpressions(x)
```

Arguments

x	pattern

replaceExpressions.character *Internal expression replacement with grep*

Description

Internal expression replacement with grep

Usage

```
## S3 method for class 'character'
replaceExpressions(x)
```

Arguments

x	pattern

replaceExpressions.PlainTextDocument *Internal expression replacement with grep*

Description

Internal expression replacement with grep

Usage

```
## S3 method for class 'PlainTextDocument'
replaceExpressions(x)
```

Arguments

x	pattern

TDM.go.df *Term document matrix of Gene Ontology Molecular Function terms*

Description

This is a provided data set for speeding up computations and also for use in unit testing. It has been compiled through the scripts available in `raw-data`

Usage

```
TDM.go.df
```

Format

An object of class `data.frame` with 4987 rows and 10824 columns.

Index