

Package ‘gower’

February 23, 2017

Maintainer Mark van der Loo <mark.vanderloo@gmail.com>

License GPL-3

Title Gower's Distance

LazyData no

Type Package

LazyLoad yes

Description Compute Gower's distance (or similarity) coefficient between records. Compute the top-n matches between records. Core algorithms are executed in parallel on systems supporting OpenMP.

Version 0.1.2

URL <https://github.com/markvanderloo/gower>

BugReports <https://github.com/markvanderloo/gower/issues>

Date 2017-02-23

Suggests testthat, knitr, rmarkdown

RoxygenNote 6.0.1

VignetteBuilder knitr

NeedsCompilation yes

Author Mark van der Loo [aut, cre]

Repository CRAN

Date/Publication 2017-02-23 23:35:17

R topics documented:

gower-package	2
gower_dist	2
gower_topn	3
Index	5

gower-package	<i>Gower's distance/similarity measure.</i>
---------------	---

Description

A C-based implementation of Gower's distance.

gower_dist	<i>Gower's distance</i>
------------	-------------------------

Description

Compute Gower's distance, pairwise between records in two data sets *x* and *y*. Records from the smallest data set are recycled over.

Usage

```
gower_dist(x, y, pair_x = NULL, pair_y = NULL, eps = 1e-08,
           nthread = getOption("gd_num_thread"))
```

Arguments

<i>x</i>	[data.frame]
<i>y</i>	[data.frame]
<i>pair_x</i>	[numeric character] (optional) Columns in <i>x</i> used for comparison. See Details below.
<i>pair_y</i>	[numeric character] (optional) Columns in <i>y</i> used for comparison. See Details below.
<i>eps</i>	[numeric] (optional) Computed numbers (variable ranges) smaller than <i>eps</i> are treated as zero.
<i>nthread</i>	Number of threads to use for parallelization. By default, for a dual-core machine, 2 threads are used. For any other machine <i>n</i> -1 cores are used so your machine doesn't freeze during a big computation. The maximum nr of threads are determined from <code>omp::get_max_threads</code> .

Value

A numeric vector of length `max(nrow(x),nrow(y))`.

Details

There are three ways to specify which columns of *x* should be compared with what columns of *y*. The first option is do give no specification. In that case columns with matching names will be used. The second option is to use only the `pair_y` argument, specifying for each column in *x* in order, which column in *y* must be used to pair it with (use 0 to skip a column in *x*). The third option is to explicitly specify the columns to be matched using `pair_x` and `pair_y`.

Note

Gower (1971) originally defined a similarity measure (s , say) with values ranging from 0 (completely dissimilar) to 1 (completely similar). The distance returned here equals $1 - s$.

References

Gower, John C. "A general coefficient of similarity and some of its properties." *Biometrics* (1971): 857-871.

See Also

[gower_topn](#)

gower_topn	<i>Find the top-n matches</i>
------------	-------------------------------

Description

Find the top-n matches in y for each record in x.

Usage

```
gower_topn(x, y, pair_x = NULL, pair_y = NULL, n = 5, eps = 1e-08,
           nthread = getOption("gd_num_thread"))
```

Arguments

x	[data.frame]
y	[data.frame]
pair_x	[numeric character] (optional) Columns in x used for comparison. See Details below.
pair_y	[numeric character] (optional) Columns in y used for comparison. See Details below.
n	The top-n indices and distances to return.
eps	[numeric] (optional) Computed numbers (variable ranges) smaller than eps are treated as zero.
nthread	Number of threads to use for parallelization. By default, for a dual-core machine, 2 threads are used. For any other machine n-1 cores are used so your machine doesn't freeze during a big computation. The maximum nr of threads are determined from omp: :get_max_threads.

Value

A list with two array elements: index and distance. Both have size $n \times \text{nrow}(x)$. Each i th column corresponds to the top-n best matches of x with rows in y.

See Also[gower_dist](#)**Examples**

```
# find the top 4 best matches in the iris data set with itself.  
x <- iris[1:3,]  
lookup <- iris[1:10,]  
gower_topn(x=x,y=lookup,n=4)
```

Index

`gower-package`, 2

`gower_dist`, 2, 4

`gower_topn`, 3, 3