Package: gower (via r-universe)

September 20, 2024

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

License GPL-3

Title Gower's Distance

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 1.0.1

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

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

Suggests tinytest (>= 0.9.3),

RoxygenNote 7.2.1

NeedsCompilation yes

Author Mark van der Loo [aut, cre], David Turner [ctb]

Repository CRAN

Date/Publication 2022-12-22 09:40:02 UTC

Contents

`gower-package` *Gower’s distance/similarity measure.*

Description

A C-based implementation of Gower’s distance.

`gower_dist` *Gower’s distance*

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,
  weights = NULL,
  ignore_case = FALSE,
  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.

`eps` `[numeric] (optional)` Computed numbers (variable ranges) smaller than `eps` are treated as zero.

`weights` `[numeric] (optional)` A vector of weights of length `ncol(x)` that defines the weight applied to each component of the gower distance.

`ignore_case` `[logical]` Toggle ignore case when neither `pair_x` nor `pair_y` are user-defined.

`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 using `omp_get_max_threads` at C level.

Value

A `numeric` vector of length `max(nrow(x),nrow(y))`. When there are no columns to compare, a message is printed and both `numeric(0)` is returned invisibly.

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 `pairs_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` *Find the top-n matches*

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,
  weights = NULL,
  ignore_case = FALSE,
  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.

weights [numeric] (optional) A vector of weights of length ncol(x) that defines the weight applied to each component of the gower distance.

ignore_case [logical] Toggle ignore case when neither pair_x nor pair_y are user-defined.

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 using omp_get_max_threads at C level.

Value

A list with two array elements: index and distance. Both have size n X nrow(x). Each ith column corresponds to the top-n best matches of x with rows in y. When there are no columns to compare, a message is printed and both distance and index will be empty matrices; the list is then returned invisibly.

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