

Package ‘proxyC’

April 21, 2019

Type Package

Title Computes Proximity in Large Sparse Matrices

Version 0.1.3

Description Computes proximity between rows or columns of large matrices efficiently in C++. Functions are optimised for large sparse matrices using the Armadillo and Intel TBB libraries. Among several built-in similarity/distance measures, computation of correlation, cosine similarity and Euclidean distance is particularly fast.

Encoding UTF-8

LazyData true

LinkingTo Rcpp, RcppParallel, RcppArmadillo (>= 0.7.600.1.0)

BugReports <https://github.com/koheiw/proxyC/issues>

SystemRequirements C++11

License GPL-3

Depends R (>= 3.1.0), methods

Imports Matrix (>= 1.2), Rcpp (>= 0.12.12), RcppParallel

Suggests testthat, proxy

RoxygenNote 6.1.1

NeedsCompilation yes

Author Kohei Watanabe [cre, aut, cph]
(<<https://orcid.org/0000-0001-6519-5265>>),
Robrecht Cannoodt [aut] (<<https://orcid.org/0000-0003-3641-729X>>)

Maintainer Kohei Watanabe <watanabe.kohei@gmail.com>

Repository CRAN

Date/Publication 2019-04-21 21:50:03 UTC

R topics documented:

simil	2
Index	3

simil	<i>Compute similarity/distance between rows or columns of large matrices</i>
-------	--

Description

Fast similarity/distance computation function for large sparse matrices. You can floor small similarity value to save computation time and storage space by an arbitrary threshold (`min_simil`) or rank (`rank`). Please increase the number of threads for better performance using [setThreadOptions](#).

Usage

```
simil(x, y = NULL, margin = 1, method = c("cosine", "correlation",
    "jaccard", "ejaccard", "dice", "edice", "hamman", "simple matching",
    "faith"), min_simil = NULL, rank = NULL, digits = 14)

dist(x, y = NULL, margin = 1, method = c("euclidean", "chisquared",
    "hamming", "kullback", "manhattan", "maximum", "canberra", "minkowski"),
    p = 2, digits = 14)
```

Arguments

<code>x</code>	a matrix or Matrix object
<code>y</code>	if a matrix or Matrix object is provided, proximity between documents or features in <code>x</code> and <code>y</code> is computed.
<code>margin</code>	integer indicating margin of similarity/distance computation. 1 indicates rows or 2 indicates columns.
<code>method</code>	method to compute similarity or distance
<code>min_simil</code>	the minimum similarity value to be recorded.
<code>rank</code>	an integer value specifying top-n most similarity values to be recorded.
<code>digits</code>	determines rounding of small values towards zero. Use primarily to correct rounding errors in C++. See zapsmall .
<code>p</code>	weight for minkowski distance

See Also

[zapsmall](#)

Examples

```
mt <- Matrix::rsparsematrix(100, 100, 0.01)
simil(mt, method = "cosine")[1:5, 1:5]
mt <- Matrix::rsparsematrix(100, 100, 0.01)
dist(mt, method = "euclidean")[1:5, 1:5]
```

Index

`dist (simil)`, [2](#)

`Matrix`, [2](#)

`matrix`, [2](#)

`setThreadOptions`, [2](#)

`simil`, [2](#)

`zapsmall`, [2](#)