Package ‘proxyC’

December 31, 2018

Type Package

Title Computes Proximity in Large Sparse Matrices

Version 0.1.0

Description Computes proximity between rows or columns of large matrices efficiently in C++. Functions are optimized 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

Collate 'RcppExports.R' 'proxy.R'

RoxygenNote 6.1.1

NeedsCompilation yes

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

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

Repository CRAN

Date/Publication 2018-12-31 22:10:02 UTC

R topics documented:

simil — *Compute similiarty/distance between raws or columns of large matrices*

Description

Fast similarity/distance computation function for large sparse matrices. You can floor small similairty value to to save computation time and storage space by an arbitrary threashold (`min_simil`) or rank (`rank`). Please increase the numbner of threads for better perfromance 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)

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

Arguments

`x` a matrix or Matrix object

`y` if a matrix or Matrix object is provided, proximity between documents or features in `x` and `y` is computed.

`margin` integer indicating margin of similiarty/distance computation. 1 indicates rows or 2 indicates columns.

`method` method to compute similarity or distance

`min_simil` the minimum similiarty value to be recoded.

`rank` an integer value specifying top-n most similiarty values to be recorded.

`p` weight for minkowski distance

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