Package ‘PRISMA’

October 12, 2022

Type Package
Title Protocol Inspection and State Machine Analysis
Version 0.2-7
Date 2018-05-26
Depends R (>= 2.10), Matrix, gplots, methods, ggplot2
Suggests tm (>= 0.6)
Author Tammo Krueger, Nicole Kraemer
Maintainer Tammo Krueger <tammokrueger@googlemail.com>
Description Loads and processes huge text
corpora processed with the sally toolbox (<http://www.mlsec.org/sally/>).
sally acts as a very fast preprocessor which splits the text files into
tokens or n-grams. These output files can then be read with the PRISMA
package which applies testing-based token selection and has some
replicate-aware, highly tuned non-negative matrix factorization and
principal component analysis implementation which allows the processing of
very big data sets even on desktop machines.
License GPL (>= 2.0)
NeedsCompilation no
Repository CRAN
Date/Publication 2018-05-26 22:01:47 UTC

R topics documented:

PRISMA-package .................................................... 2
asap ................................................................. 3
corpusToPrisma .................................................... 4
estimateDimension ............................................... 5
getDuplicateData ................................................ 6
getMatrixFactorizationLabels .................................. 6
loadPrismaData .................................................... 7
plot.prisma .......................................................... 8
plot.prismaDimension ........................................... 9
Description

Loads and processes huge text corpora processed with the sally toolbox (<http://www.mlsec.org/sally/>). sally acts as a very fast preprocessor which splits the text files into tokens or n-grams. These output files can then be read with the PRISMA package which applies testing-based token selection and has some replicate-aware, highly tuned non-negative matrix factorization and principal component analysis implementation which allows the processing of very big data sets even on desktop machines.

Details

Package: PRISMA
Type: Package
Title: Protocol Inspection and State Machine Analysis
Version: 0.2-7
Date: 2018-05-26
Depends: Matrix, gplots, methods, ggplot2
Suggests: tm (>= 0.6)
Author: Tammo Krueger, Nicole Kraemer
Maintainer: Tammo Krueger <tammokrueger@googlemail.com>
Description: Loads and processes huge text corpora processed with the sally toolbox (<http://www.mlsec.org/sally/>). sally
License: GPL (>=2.0)

Index of help topics:

<table>
<thead>
<tr>
<th>PRISMA-package</th>
<th>Protocol Inspection and State Machine Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>asap</td>
<td>The ASAP Data Set</td>
</tr>
<tr>
<td>corpusToPrisma</td>
<td>Convert tm copus to PRISMA</td>
</tr>
<tr>
<td>estimateDimension</td>
<td>Estimate Inner Dimension</td>
</tr>
<tr>
<td>getDuplicateData</td>
<td>Restores Data with Duplicates</td>
</tr>
<tr>
<td>getMatrixFactorizationLabels</td>
<td>Convert Coordinates of Matrix Factorization to Labels</td>
</tr>
<tr>
<td>loadPrismaData</td>
<td>Load PRISMA Data Files</td>
</tr>
<tr>
<td>plot.prisma</td>
<td>Generics For PRISMA Objects</td>
</tr>
</tbody>
</table>
plot.prismaDimension  Generics For PRISMA Objects
plot.prismaMF     Generics For PRISMA Objects
prismaDuplicatePCA Matrix Factorization Based on Replicate-Aware
                     PCA
prismaHclust      Matrix Factorization Based on Hierarchical
                     Clustering
prismaNMF         Matrix Factorization Based on Replicate-Aware
                     NMF
thesis            The Thesis Data Set

Further information is available in the following vignettes:

PRISMA    Quick introduction (source)

Author(s)

Tammo Krueger, Nicole Kraemer
Maintainer: Tammo Krueger <tammokrueger@googlemail.com>

References

Honeypots 5th ACM Workshop on Artificial Intelligence and Security (AISEC 2012), accepted
Krueger, T., Kraemer, N., Rieck, K. (2011) ASAP: Automatic Semantics-Aware Analysis of Net-
work Payloads Privacy and Security Issues in Data Mining and Machine Learning - International
ECML/PKDD Workshop. Lecture Notes in Computer Science 6549, Springer. 50 - 63

Examples

# please see the vingette for examples

______________
asap The ASAP Data Set
______________

Description

Toy data set to show the capabilities of the PRISMA package.

Usage

asap

Format

A prisma object.
**Author(s)**

Tammo Krueger <tammokrueger@googlemail.com>

**References**


---

**corpusToPrisma**  
*Convert tm corpus to PRISMA*

**Description**

Converts a tm corpus object to a PRISMA object.

**Usage**

```r
corpusToPrisma(corpus, alpha = 0.05, skipFeatureCorrelation = FALSE)
```

**Arguments**

- `corpus`: a tm corpus
- `alpha`: significance level for the feature tests. If NULL, all features are kept.
- `skipFeatureCorrelation`: should the grouping of features based on correlation analysis be skipped.

**Value**

- `prismaData`: data object representing the tokenized documents as features x samples matrix.

**Author(s)**

Tammo Krueger <tammokrueger@googlemail.com>

**Examples**

```r
if (require("tm") && packageVersion("tm") >= '0.6') {
  data(thesis)
  thesis
  thesis = corpusToPrisma(thesis, NULL, TRUE)
  thesis
}
```
Description

Matrix factorization methods compress the original data matrix \( A \in \mathbb{R}^{f,N} \) with \( f \) features and \( N \) samples into two parts, namely \( A = BC \) with \( B \in \mathbb{R}^{f,k} \), \( C \in \mathbb{R}^{k,N} \). The function estimateDimension estimates \( k \) based on a noise model estimated from a scrambled version of the original data matrix.

Usage

```r
estimateDimension(prismaData, alpha = 0.05, nScrambleSamples = NULL)
```

Arguments

- `prismaData` A prismaData object loaded via loadPrismaData
- `alpha` Error probability for confidence intervals
- `nScrambleSamples` The number of scrambled samples that should be used to estimate the noise model. NULL means to use the complete data set.

Value

- `estDim` prismaDimension object that can be printed and plotted.

Author(s)

Tammo Krueger <tammokrueger@googlemail.com>

References


Examples

```r
# please see the vingette for examples
```
getDuplicateData

Restores Data with Duplicates

Description

The `loadPrismaData` function triggers a feature selection and data combination methods which subsequently remove duplicate entries for efficient representation of the data. The `getDuplicateData` rebuilds the data matrix with explicit representation of all duplicate entries.

Usage

```
getDuplicateData(prismaData)
```

Arguments

`prismaData` prisma data loaded via `loadPrismaData`

Value

`dataWithDuplicates` Data matrix containing explicit copies of all duplicates.

Author(s)

Tammo Krueger <tammokrueger@googlemail.com>

Examples

```
data(asap)
dataWithDuplicates = getDuplicateData(asap)
```

getMatrixFactorizationLabels

Convert Coordinates of Matrix Factorization to Labels

Description

Given a matrix factorization object \( A = BC \), this function returns for each document the index of the inner dimension which has the maximal coordinate. Thus, it converts the fuzzy clustering found in the columns of the \( C \) matrix into a hard clustering by returning the position with the maximal coordinate value.

Usage

```
getMatrixFactorizationLabels(prismaMF)
```
**loadPrismaData**

**Arguments**
- `prismaMF` a matrix factorization object.

**Value**
- `labels` vector containing the label assignment for each document.

**Author(s)**
Tammo Krueger <tammokrueger@googlemail.com>

**See Also**
- `prismaNMF`

---

**loadPrismaData**  
*Load PRISMA Data Files*

**Description**
Loads files generated by the sally tool (see [http://www.mlsec.org/sally/](http://www.mlsec.org/sally/)) and represents the data as binary token/ngrams x documents matrix. After loading, statistical tests are applied to find features which are not volatile nor constant. Co-occurring features are grouped to further compactify the data. See `system.file("extdata","sallyPreprocessing.py", package="PRISMA")` for a Python script which generates the corresponding .fsally file from a .sally file which reduce the loading time via `loadPrismaData` considerably.

**Usage**
```r
loadPrismaData(path, maxLines = -1, fastSally = TRUE,
                alpha = 0.05, skipFeatureCorrelation=FALSE)
```

**Arguments**
- `path` path of the data file without the .sally extension. loadPrisma loads path.sally or path.fsally depending on the fastSally switch.
- `maxLines` maximal number of lines to read from the data file. -1 means to read all lines.
- `fastSally` should the fsally file be used, which drastically decreases loading time.
- `alpha` significance level for the feature tests. If NULL, all features are kept.
- `skipFeatureCorrelation` should the grouping of features based on correlation analysis be skipped.

**Value**
- `prismaData` data object representing the tokenized documents as features x samples matrix.
plot.prisma

Author(s)
Tammo Krueger <tammokrueger@googlemail.com>

References
See http://www.mlsec.org/sally/ for the sally utility.

Examples
# please see the vinette for examles
# please see system.file("extdata","asap.tar.gz", package="PRISMA") for
# an example sally output

data(asap)
print(asap)
plot(asap)

plot.prisma
Generics For PRISMA Objects

Description
Print and plot generic for the PRISMA objects.

Usage
## S3 method for class 'prisma'
print(x, ...)
## S3 method for class 'prisma'
plot(x, ...)

Arguments
x PRISMA data loaded via loadPrismaData
... not used

Author(s)
Tammo Krueger <tammokrueger@googlemail.com>

See Also
estimateDimension, prismaHclust, prismaDuplicatePCA, prismaNMF

Examples
data(asap)
print(asap)
plot(asap)
plot.prismaDimension

Generics For PRISMA Objects

Description

Print and plot generic for the PRISMA dimension objects.

Usage

## S3 method for class 'prismaDimension'
print(x, ...)
## S3 method for class 'prismaDimension'
plot(x, ...)

Arguments

x \hspace{1em} \text{PRISMA dimension object generated via} \texttt{estimateDimension}

... \hspace{1em} \text{not used}

Author(s)

Tammo Krueger <tammokrueger@googlemail.com>

See Also

\texttt{estimateDimension, prismaHclust, prismaDuplicatePCA, prismaNMF}

Examples

# please see the vingette for examles

plot.prismaMF

Generics For PRISMA Objects

Description

Print and plot generic for the PRISMA matrix factorization objects.

Usage

## S3 method for class 'prismaMF'
plot(x, nLines = NULL, baseIndex = NULL, sampleIndex = NULL,
minValue = NULL, noRowClustering = FALSE, noColClustering = FALSE, type
= c("base", "coordinates"), ...)

# please see the vingette for examles
prismaDuplicatePCA

Arguments

x  
PRISMA matrix factorization object

nLines  
number of lines that should be plotted

baseIndex  
which bases should be plotted

sampleIndex  
which samples should be plotted

minValue  
cut-off value, i.e., every value smaller than minValue won’t be shown

noRowClustering  
don’t cluster the rows

noColClustering  
don’t cluster the columns

type  
show the base (type = "base", i.e. the $B$ matrix) or show the coordinate (type = "coordinates", i.e. the $C$ matrix).

...  
not used

Author(s)

Tammo Krueger <tammokrueger@googlemail.com>

See Also

estimateDimension, prismaHclust, prismaDuplicatePCA, prismaNMF

Examples

# please see the vignette for examples
**Author(s)**
Tammo Krueger <tammokrueger@googlemail.com>

**Examples**

# please see the vingette for examles

---

**prismaHclust**  
Matrix Factorization Based on Hierarchical Clustering

**Description**

A matrix factorization $A = BC$ based on the results of hclust is constructed, which holds the mean feature values for each cluster in the matrix $B$ and the indication of the cluster in the matrix $C$ for each data point (i.e. each data point is represented by its assigned cluster center).

**Usage**

prismaHclust(prismaData, ncomp, method = "single")

**Arguments**

- **prismaData**: PRISMA data for which a clustering should be calculated.
- **ncomp**: the number of components that should be extracted.
- **method**: the method used for clustering.

**Value**

prismaHclust  
Matrix factorization object containing $B$ and $C$ resulting from the hierarchical clustering of the data.

**Author(s)**
Tammo Krueger <tammokrueger@googlemail.com>

**See Also**

hclust

**Examples**

# please see the vingette for examles
prismaNMF

Matrix Factorization Based on Replicate-Aware NMF

Description

Matrix factorization $A = BC$ with strictly positiv matrices $B, C$ which minimize the reconstruction error $\|A - BC\|$. This replicate-aware version of the non-negative matrix factorization (NMF) is based on the alternating least squares approach and exploits the replicate information to speed up the calculation.

Usage

prismaNMF(prismaData, ncomp, time = 60, pca.init = TRUE, doNorm = TRUE, oldResult = NULL)

Arguments

<table>
<thead>
<tr>
<th>Argument</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>prismaData</td>
<td>PRISMA data for which a NMF should be calculated.</td>
</tr>
<tr>
<td>ncomp</td>
<td>either an integer or prismaDimension object specifying the inner dimension</td>
</tr>
<tr>
<td></td>
<td>of the matrix factorization.</td>
</tr>
<tr>
<td>time</td>
<td>seconds after which the calculation should end.</td>
</tr>
<tr>
<td>pca.init</td>
<td>should the $B$ matrix be initialized by a PCA.</td>
</tr>
<tr>
<td>doNorm</td>
<td>should the $B$ matrix normalized (i.e. all columns have the Euclidean length</td>
</tr>
<tr>
<td></td>
<td>of 1).</td>
</tr>
<tr>
<td>oldResult</td>
<td>re-use results of a previous run, i.e. $B$ and $C$ are pre-initialized with</td>
</tr>
<tr>
<td></td>
<td>the values of this previous matrix factorization object.</td>
</tr>
</tbody>
</table>

Value

prismaNMF Matrix factorization object containing the $B$ and $C$ matrix.

Author(s)

Tammo Krueger <tammokrueger@googlemail.com>

References


Examples

# please see the vignette for examples
**Description**

The 15 sections of a thesis (see references) as a tm-corpus.

**Usage**

thesis

**Format**

A tm-corpus.

**Author(s)**

Tammo Krueger <tammokrueger@googlemail.com>

**References**

Index

* datasets
  asap, 3
  thesis, 13

* package
  PRISMA-package, 2

asap, 3

corpusToPrisma, 4

estimateDimension, 5, 8–10

getDuplicateData, 6, 6
getMatrixFactorizationLabels, 6

hclust, 11

loadPrismaData, 6, 7, 7, 8

plot.prisma, 8
plot.prismaDimension, 9
plot.prismaMF, 9
print.prisma (plot.prisma), 8
print.prismaDimension
  (plot.prismaDimension), 9
PRISMA (PRISMA-package), 2
PRISMA-package, 2
prismaDuplicatePCA, 8–10, 10
prismaHclust, 8–10, 11
prismaNMF, 7–10, 12

thesis, 13