Package ‘mldr.datasets’

August 29, 2016

Title   R Ultimate Multilabel Dataset Repository
Version 0.3.15
Date    2016-01-16
Description Large collection of multilabel datasets along with the functions
needed to export them to several formats, to make partitions, and to obtain
bibliographic information.

URL https://github.com/fcharte/mldr.datasets
Depends R (>= 3.0.0)
Imports utils
License LGPL (>= 3) | file LICENSE
LazyData true
RoxygenNote 5.0.1
Suggests mldr
NeedsCompilation no
Author David Charte [cre],
        Francisco Charte [aut]
Maintainer David Charte <fdavidcl@outlook.com>
Repository CRAN
Date/Publication 2016-01-16 13:12:17

R topics documented:

bibtex ................................................................. 3
birds ...................................................................... 4
bookmarks ............................................................. 4
cal500 ................................................................. 5
check_n_load.mldr .................................................. 6
corel16k001 ......................................................... 6
corel16k002 ......................................................... 7
corel16k003 ......................................................... 8
topics documented:
corel16k004 ......................................................... 8
corel16k005 ......................................................... 9
corel16k006 ......................................................... 10
corel16k007 ......................................................... 10
corel16k008 ......................................................... 11
corel16k009 ......................................................... 12
corel16k010 ......................................................... 12
corel15k ............................................................... 13
delicious ............................................................... 14
delicious ............................................................... 14
enron ................................................................. 15
eurlexdc_test ......................................................... 16
eurlexdc_tra ......................................................... 16
eurlexev_test ......................................................... 17
eurlexev_tra ......................................................... 18
eurlexsm_test ......................................................... 18
eurlexsm_tra ......................................................... 19
flags ................................................................. 20
genbase ............................................................... 20
imdb ................................................................. 21
langlog ............................................................... 22
mediamill ............................................................. 22
medical ............................................................... 23
mldrs ................................................................. 24
ng20 ................................................................. 24
nuswide_BoW ......................................................... 25
nuswide_VLAD ....................................................... 25
ohsumed ............................................................. 26
random.kfolds ....................................................... 27
rcv1sub1 .............................................................. 27
rcv1sub2 .............................................................. 28
rcv1sub3 .............................................................. 29
rcv1sub4 .............................................................. 29
rcv1sub5 .............................................................. 30
reutersk500 .......................................................... 31
scene ................................................................. 31
slashdot .............................................................. 32
stackex_chemistry .................................................... 32
stackex_chess ......................................................... 33
stackex_coffee ....................................................... 34
stackex_cooking ..................................................... 34
stackex_cs ........................................................... 35
stackex_philosophy ............................................... 36
stratified.kfolds .................................................... 36
tmc2007 ............................................................. 37
tmc2007_500 .......................................................... 38
toBibtex.mldr ......................................................... 39
write.mldr .......................................................... 39
bibtex

Dataset with BibTeX entries

Description
Multilabel dataset from the text domain.

Usage
bibtex()

Format
An mldr object with 7395 instances, 1836 attributes and 159 labels

Source
Katakis, I. and Tsoumakas, G. and Vlahavas, I., "Multilabel Text Classification for Automated Tag Suggestion", in Proc. ECML PKDD08 Discovery Challenge, Antwerp, Belgium, pp. 75-83, 2008

Examples
## Not run:
bibtex() # Check and load the dataset
tobibtex(bibtex)
bibtex$measures

## End(Not run)
**birds**  
*Dataset with sounds produced by birds and the species they belong to*

**Description**
Multilabel dataset from the sound domain.

**Usage**
birds

**Format**
An mldr object with 645 instances, 260 attributes and 19 labels

**Source**

**Examples**
```r
toBibtex(birds)  
birds$measures
```

**bookmarks**  
*Dataset with data from web bookmarks and their categories*

**Description**
Multilabel dataset from the text domain.

**Usage**
bookmarks()

**Format**
An mldr object with 87856 instances, 2150 attributes and 208 labels
**Source**

Katakis, I. and Tsoumakas, G. and Vlahavas, I., "Multilabel Text Classification for Automated Tag Suggestion", in Proc. ECML PKDD08 Discovery Challenge, Antwerp, Belgium, pp. 75-83, 2008

**Examples**

```r
## Not run:
bookmarks()  # Check and load the dataset
toBibtex(bookmarks)
bookmarks$measures

## End(Not run)
```

---

**cal500**

*Dataset with music data along with labels for emotions, instruments, genres, etc.*

**Description**

Multilabel dataset from the music domain.

**Usage**

```r
cal500
```

**Format**

An `mlrd` object with 502 instances, 68 attributes and 174 labels

**Source**


**Examples**

```r
## Not run:
toBibtex(cal500)
cal500$measures

## End(Not run)
```
**check_n_load.mldr**  
*Check if an mldr object is locally available and download it if needed*

**Description**

This function checks if the mldr object whose name is given as input is locally available, loading it in memory. If necessary, the dataset will be downloaded from the GitHub repository and saved locally.

**Usage**

```r
cHECK_n_load_mldr(mldr.name)
```

**Arguments**

- `mldr.name`  
  Name of the dataset to load

**Examples**

```r  
## Not run:  
library(mldr.datasets)  
check_n_load.mldr("bibtex")  
bibtex$measures
## End(Not run)
```

---

**corel16k001**  
*Datasets with data from the Corel image collection. There are 10 subsets in corel16k*

**Description**

Multilabel dataset from the image domain.

**Usage**

```r
corel16k001()
```

**Format**

An mldr object with 13766 instances, 500 attributes and 153 labels

**Source**

Barnard, K. and Duygulu, P. and Forsyth, D. and de Freitas, N. and Blei, D. M. and Jordan, M. I.,  
**corel16k002**

### Examples

```r
## Not run:
corel16k001()  # Check and load the dataset
toBibtex(corel16k001)
corel16k001$measures

## End(Not run)
```

---

**corel16k002**

Datasets with data from the Corel image collection. There are 10 subsets in corel16k

---

**Description**

Multilabel dataset from the image domain.

**Usage**

```r
corel16k002()
```

**Format**

An mldr object with 13761 instances, 500 attributes and 164 labels

**Source**


**Examples**

```r
## Not run:
corel16k002()  # Check and load the dataset
toBibtex(corel16k002)
corel16k002$measures

## End(Not run)
```
corel16k003

Datasets with data from the Corel image collection. There are 10 subsets in corel16k

Description

Multilabel dataset from the image domain.

Usage

corel16k003()

Format

An mldr object with 13760 instances, 500 attributes and 154 labels

Source


Examples

```r
## Not run:
corel16k003() # Check and load the dataset
toBibtex(corel16k003)
corel16k003$measures

## End(Not run)
```

corel16k004

Datasets with data from the Corel image collection. There are 10 subsets in corel16k

Description

Multilabel dataset from the image domain.

Usage

corel16k004()

Format

An mldr object with 13837 instances, 500 attributes and 162 labels
Source


Examples

```r
## Not run:
corel16k005() # Check and load the dataset
toBibTex(corel16k005)
corel16k005$measures

## End(Not run)
```

datasets with data from the Corel image collection. There are 10 sub-
sets in corel16k

Description

Multilabel dataset from the image domain.

Usage

```r
corel16k005()
```

Format

An mldr object with 13847 instances, 500 attributes and 160 labels

Source


Examples

```r
## Not run:
corel16k005() # Check and load the dataset
toBibTex(corel16k005)
corel16k005$measures

## End(Not run)
```
Datasets with data from the Corel image collection. There are 10 subsets in corel16k

Description
Multilabel dataset from the image domain.

Usage
corel16k006()

Format
An mldr object with 13859 instances, 500 attributes and 162 labels

Source

Examples
```r
## Not run:
corel16k006() # Check and load the dataset
toBibtex(corel16k006)
corel16k006$measures

## End(Not run)
```

corel16k007

Datasets with data from the Corel image collection. There are 10 subsets in corel16k

Description
Multilabel dataset from the image domain.

Usage
corel16k007()

Format
An mldr object with 13915 instances, 500 attributes and 174 labels
Source


Examples

```r
## Not run:
corel16k008() # Check and load the dataset
toBibtex(corel16k008)
corel16k008$measures

## End(Not run)
```

datasets with data from the Corel image collection. There are 10 subsets in corel16k

Description

Multilabel dataset from the image domain.

Usage

```r
corel16k008()
```

Format

An mldr object with 13864 instances, 500 attributes and 168 labels

Source


Examples

```r
## Not run:
corel16k008() # Check and load the dataset
toBibtex(corel16k008)
corel16k008$measures

## End(Not run)
```
corel16k009

Datasets with data from the Corel image collection. There are 10 subsets in corel16k

Description
Multilabel dataset from the image domain.

Usage
corel16k009()

Format
An mldr object with 13884 instances, 500 attributes and 173 labels

Source

Examples
## Not run:
corel16k009() # Check and load the dataset
toBibtex(corel16k009)
corel16k009$measures

## End(Not run)

---

corel16k010

Datasets with data from the Corel image collection. There are 10 subsets in corel16k

Description
Multilabel dataset from the image domain.

Usage
corel16k010()

Format
An mldr object with 13618 instances, 500 attributes and 144 labels
Source


Examples

```r
## Not run:
corel16k010() # Check and load the dataset
toBibtex(corel16k010)
corel16k010$measures

## End(Not run)
```

---

**corel5k**

*Dataset with data from the Corel image collection*

Description

Multilabel dataset from the image domain.

Usage

`corel5k()`

Format

An mldr object with 5000 instances, 499 attributes and 374 labels

Source


Examples

```r
## Not run:
corel5k() # Check and load the dataset
toBibtex(corel5k)
corel5k$measures

## End(Not run)
```
delicious

Dataset generated from the del.icio.us site bookmarks

Description
Multilabel dataset from the text domain.

Usage
delicious()

Format
An mldr object with 16105 instances, 500 attributes and 983 labels

Source
Tsoumakas, G. and Katakis, I. and Vlahavas, I., "Effective and Efficient Multilabel Classification in Domains with Large Number of Labels", in Proc. ECML/PKDD Workshop on Mining Multidimensional Data, Antwerp, Belgium, MMD08, pp. 30–44, 2008

Examples
```r
## Not run:
delicious()  # Check and load the dataset
tobibtex(delicious)
delicious$measures

## End(Not run)
```

emotions

Dataset with features extracted from music tracks and the emotions they produce

Description
Multilabel dataset from the music domain.

Usage
emotions

Format
An mldr object with 593 instances, 72 attributes and 6 labels
# enron

## Source


## Examples

```r
## Not run:
toBibtex(emotions)
emotions$measures

## End(Not run)
```

<table>
<thead>
<tr>
<th>enron</th>
<th>Dataset with email messages and the folders where the users stored them</th>
</tr>
</thead>
</table>

## Description

Multilabel dataset from the text domain.

## Usage

```r
enron()
```

## Format

An mldr object with 1702 instances, 1001 attributes and 53 labels

## Source

Klimt, B. and Yang, Y., "The Enron Corpus: A New Dataset for Email Classification Research", in Proc. ECML04, Pisa, Italy, pp. 217-226, 2004

## Examples

```r
## Not run:
enron() # Check and load the dataset
toBibtex(enron)
enron$measures

## End(Not run)
```
eurlexdc_test

List with 10 folds of the test data from the EUR-Lex directory codes dataset

Description
Multilabel dataset from the text domain.

Usage
eurlexdc_test()

Format
An mldr object with 1935 instances, 5000 attributes and 412 labels

Source
Mencia, E. L. and Furnkranz, J., "Efficient pairwise multilabel classification for large-scale problems in the legal domain", Machine Learning and Knowledge Discovery in Databases, pp. 50–65, 2008

Examples
## Not run:
eurlexdc_test()  # Check and load the dataset
toBibtex(eurlexdc_test[[1]])
eurlexdc_test[[1]]$measures
## End(Not run)

eurlexdc_tra

List with 10 folds of the train data from the EUR-Lex directory codes dataset

Description
Multilabel dataset from the text domain.

Usage
eurlexdc_tra()

Format
An mldr object with 17413 instances, 5000 attributes and 412 labels
Source

Mencia, E. L. and Furnkranz, J., "Efficient pairwise multilabel classification for large-scale problems in the legal domain", Machine Learning and Knowledge Discovery in Databases, pp. 50–65, 2008

Examples

## Not run:
```r
eurlexdc_tral( # Check and load the dataset
tobibtex(eurlexdc_test[[1]]))
eurlexdc_test[[1]]$measures
```

## End(Not run)

----

| eurlexev_test | List with 10 folds of the test data from the EUR-Lex EUROVOC descriptors dataset |

Description

Multilabel dataset from the text domain.

Usage

```r
eurlexev_test()
```

Format

An mldr object with 1935 instances, 5000 attributes and 3993 labels

Source

Mencia, E. L. and Furnkranz, J., "Efficient pairwise multilabel classification for large-scale problems in the legal domain", Machine Learning and Knowledge Discovery in Databases, pp. 50–65, 2008

Examples

## Not run:
```r
eurlexev_test() # Check and load the dataset
tobibtex(eurlexev_test[[1]])
eurlexev_test[[1]]$measures
```

## End(Not run)
eurlexev_tra

List with 10 folds of the train data from the EUR-Lex EUROVOC descriptors dataset

Description
Multilabel dataset from the text domain.

Usage
eurlexev_tra()

Format
An mldr object with 17413 instances, 5000 attributes and 3993 labels

Source
Mencia, E. L. and Furnkranz, J., "Efficient pairwise multilabel classification for large-scale problems in the legal domain", Machine Learning and Knowledge Discovery in Databases, pp. 50–65, 2008

Examples
```r
## Not run:
eurlexev_tra()  # Check and load the dataset
toBibtex(eurlexev_tra[[1]])
eurlexev_tra[[1]]$measures
```

```r
## End(Not run)
```

eurlexsm_test

List with 10 folds of the test data from the EUR-Lex subject matters dataset

Description
Multilabel dataset from the text domain.

Usage
eurlexsm_test()

Format
An mldr object with 1935 instances, 5000 attributes and 201 labels
Source

Mencia, E. L. and Furnkranz, J., "Efficient pairwise multilabel classification for large-scale problems in the legal domain", Machine Learning and Knowledge Discovery in Databases, pp. 50–65, 2008

Examples

```r
## Not run:
eurlexsm_test() # Check and load the dataset
toBibtex(eurlexsm_test[[1]])
eurlexsm_test[[1]]$measures

## End(Not run)
```

---

**eurlexsm_tra** List with 10 folds of the train data from the EUR-Lex subject matters dataset

Description

Multilabel dataset from the text domain.

Usage

```r
eurlexsm_tra()
```

Format

An mldr object with 17413 instances, 5000 attributes and 201 labels

Source

Mencia, E. L. and Furnkranz, J., "Efficient pairwise multilabel classification for large-scale problems in the legal domain", Machine Learning and Knowledge Discovery in Databases, pp. 50–65, 2008

Examples

```r
## Not run:
eurlexsm_test() # Check and load the dataset
toBibtex(eurlexsm_test[[1]])
eurlexsm_test[[1]]$measures

## End(Not run)
```
flags

Dataset with features corresponding to world flags

Description
Multilabel dataset from the image domain.

Usage
flags

Format
An mldr object with 194 instances, 19 attributes and 7 labels

Source

Examples
```r
## Not run:
toBibtex(flags)
flags$m$measures
## End(Not run)
```

---

genbase

Dataset with genes data and their functional expression

Description
Multilabel dataset from the biology domain.

Usage
genbase

Format
An mldr object with 662 instances, 1186 attributes and 27 labels
**Source**

Diplaris, S. and Tsoumakas, G. and Mitkas, P. and Vlahavas, I., "Protein Classification with Multiple Algorithms", in Proc. 10th Panhellenic Conference on Informatics, Volos, Greece, PCI05, pp. 448–456, 2005

**Examples**

```r
## Not run:
toBibtex(genbase)
genbase$measures

## End(Not run)
```

---

**imdb**

*Dataset generated from the IMDB film database*

**Description**

Multilabel dataset from the text domain.

**Usage**

```r
imdb()
```

**Format**

An mldr object with 120919 instances, 1001 attributes and 28 labels

**Source**


**Examples**

```r
## Not run:
imdb() # Check and load the dataset
toBibtex(imdb)
imdb$measures

## End(Not run)
```
langlog  

*Dataset with data from the Language forum discussion*

**Description**
Multilabel dataset from the text domain.

**Usage**
langlog

**Format**
An mldr object with 1460 instances, 1004 attributes and 75 labels

**Source**
Read, Jesse, "Scalable multi-label classification", University of Waikato, 2010

**Examples**
```r
## Not run:
tobibtex(langlog)
langlog$measures
## End(Not run)
```

mediamill  

*Dataset with features extracted from video sequences and semantic concepts assigned as labels*

**Description**
Multilabel dataset from the video domain.

**Usage**
mediamill()

**Format**
An mldr object with 43907 instances, 120 attributes and 101 labels

**Source**
**medical**

**Examples**

```r
## Not run:
mediamill()  # Check and load the dataset
toBibtex(mediamill)
mediamill$measures

## End(Not run)
```

---

**medical**  
*Dataset generated from medical reports*

---

**Description**

Multilabel dataset from the text domain.

**Usage**

```r
medical
```

**Format**

An mldr object with 978 instances, 1449 attributes and 45 labels

**Source**


**Examples**

```r
## Not run:
toBibtex(medical)
medical$measures

## End(Not run)
```
mldrs

*Obtain and show a list of additional datasets available to download*

**Description**

The function downloads from GitHub the most up to date list of additional datasets. Those datasets are not included into the package, but can be downloaded and saved locally.

**Usage**

mldrs()

**Examples**

```r
## Not run:
library(mldr.datasets)
mldrs()

## End(Not run)
```

ng20

*Dataset with news messages and the news groups they belong to*

**Description**

Multilabel dataset from the text domain. The original name of the dataset is 20ng

**Usage**

ng20

**Format**

An mldr object with 19300 instances, 1006 attributes and 20 labels

**Source**

Ken Lang, "Newsweeder: Learning to filter netnews", in Proc. 12th International Conference on Machine Learning, pp. 331-339, 1995

**Examples**

```r
## Not run:
toBibtex(ng20)
ng20$measures

## End(Not run)
```
**nuswide_Bow**

*Dataset obtained from the NUS-WIDE database with BoW representation*

<table>
<thead>
<tr>
<th>nuswide_Bow</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dataset obtained from the NUS-WIDE database with BoW representation</td>
</tr>
</tbody>
</table>

**Description**

Multilabel dataset from the image domain.

**Usage**

```
nuswide_Bow()
```

**Format**

An mldr object with 269648 instances, 501 attributes and 81 labels

**Source**

Chua, Tat-Seng and Tang, Jinhui and Hong, Richang and Li, Haojie and Luo, Zhiping and Zheng, Yantao, "NUS-WIDE: a real-world web image database from National University of Singapore", in Proc. of the ACM international conference on image and video retrieval, pp. 48, 2009

**Examples**

```r
## Not run:
nuswide_Bow()  # Check and load the dataset
toBibtex(nuswide_Bow)
nuswide_Bow$measures

## End(Not run)
```

---

**nuswide_VLAD**

*Dataset obtained from the NUS-WIDE database with cVLAD+ representation*

<table>
<thead>
<tr>
<th>nuswide_VLAD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dataset obtained from the NUS-WIDE database with cVLAD+ representation</td>
</tr>
</tbody>
</table>

**Description**

Multilabel dataset from the image domain.

**Usage**

```
nuswide_VLAD()
```

**Format**

An mldr object with 269648 instances, 129 attributes and 81 labels
Source

Chua, Tat-Seng and Tang, Jinhui and Hong, Richang and Li, Haojie and Luo, Zhiping and Zheng, Yantao, "NUS-WIDE: a real-world web image database from National University of Singapore", in Proc. of the ACM international conference on image and video retrieval, pp. 48, 2009

Examples

## Not run:
nuswide_VLAD() # Check and load the dataset
toBibtex(nuswide_VLAD)
nuswide_VLAD$measures

## End(Not run)

__ohsumed__

Dataset generated from a subset of the Medline database

Description

Multilabel dataset from the text domain.

Usage

ohsumed()

Format

An mldr object with 13929 instances, 1002 attributes and 23 labels

Source

Joachims, Thorsten, "Text Categorization with Suport Vector Machines: Learning with Many Relevant Features", in Proc. 10th European Conference on Machine Learning, pp. 137-142, 1998

Examples

## Not run:

ohsumed() # Check and load the dataset
toBibtex(ohsumed)
ohsumed$measures

## End(Not run)
**random.kfolds**

**Partition an mldr object into k folds**

**Description**

This method randomly partitions the given dataset into k folds, providing training and test partitions for each fold.

**Usage**

```r
random.kfolds(mld, k = 5, seed = 10)
```

**Arguments**

- `mld`: The mldr object to be partitioned
- `k`: The number of folds to be generated. By default is 5
- `seed`: The seed to initialize the random number generator. By default is 10. Change it if you want to obtain partitions containing different samples, for instance to use a 2x5 cv strategy

**Value**

An mldr.folds object. This is a list containing k elements, one for each fold. Each element is made up of two mldr objects, called train and test

**Examples**

```r
## Not run:
library(mldr.datasets)
library(mldr)
folds.emotions <- random.kfolds(emotions)
summary(folds.emotions[[1]]$train)
summary(folds.emotions[[1]]$test)
## End(Not run)
```

---

**rcv1sub1**

*Dataset from the Reuters corpus (subset 1)*

**Description**

Multilabel dataset from the text domain.

**Usage**

```r
rcv1sub1()
```
Format

An mldr object with 6000 instances, 47236 attributes and 101 labels

Source


Examples

```r
## Not run:
rcv1sub2() # Check and load the dataset
toBibtex(rcv1sub2)
rcv1sub2$measures
## End(Not run)
```

---

rcv1sub2

*Dataset from the Reuters corpus (subset 2)*

Description

Multilabel dataset from the text domain.

Usage

`rcv1sub2()`

Format

An mldr object with 6000 instances, 47236 attributes and 101 labels

Source


Examples

```r
## Not run:
rcv1sub2() # Check and load the dataset
toBibtex(rcv1sub2)
rcv1sub2$measures
## End(Not run)
```
Dataset from the Reuters corpus (subset 3)

Description
Multilabel dataset from the text domain.

Usage
rcv1sub3()

Format
An mldr object with 6000 instances, 47236 attributes and 101 labels

Source

Examples

```r
## Not run:
rcv1sub3() # Check and load the dataset
toBibtex(rcv1sub3)
rcv1sub3$measures

## End(Not run)
```

Dataset from the Reuters corpus (subset 4)

Description
Multilabel dataset from the text domain.

Usage
rcv1sub4()

Format
An mldr object with 6000 instances, 47229 attributes and 101 labels
Source


Examples

```r
## Not run:
rcv1sub5() # Check and load the dataset
tobilbex(rcv1sub5)
rcv1sub5$measures

## End(Not run)
```

---

rcv1sub5      Dataset from the Reuters corpus (subset 5)

Description

Multilabel dataset from the text domain.

Usage

```r
crv1sub5()
```

Format

An mldr object with 6000 instances, 47235 attributes and 101 labels

Source


Examples

```r
## Not run:
rcv1sub5() # Check and load the dataset
tobilbex(rcv1sub5)
rcv1sub5$measures

## End(Not run)
```
**reutersk500**

*Dataset from the Reuters Corpus with the 500 most relevant features selected*

**Description**

Multilabel dataset from the text domain.

**Usage**

`reutersk500()`

**Format**

An mldr object with 6000 instances, 500 attributes and 103 labels

**Source**

Read, Jesse, "Scalable multi-label classification", University of Waikato, 2010

**Examples**

```r
## Not run:
reutersk500()  # Check and load the dataset
tobibtex(reutersk500)
reutersk500$measures

## End(Not run)
```

**scene**

*Dataset from images with different natural scenes*

**Description**

Multilabel dataset from the image domain.

**Usage**

`scene()`

**Format**

An mldr object with 2407 instances, 294 attributes and 6 labels

**Source**

### slashdot

Dataset generated from slashdot.org site entries

**Description**

Multilabel dataset from the text domain.

**Usage**

`slashdot`

**Format**

An mldr object with 3782 instances, 1079 attributes and 22 labels

**Source**


**Examples**

```r
## Not run:
toBibtex(slashdot)
slashdot$measures
## End(Not run)
```

### stackex_chemistry

Dataset from the Stack Exchange’s chemistry forum

**Description**

Multilabel dataset from the text domain.

**Usage**

`stackex_chemistry()`
Stack Exchange Chess

**Format**

An mldr object with 6961 instances, 540 attributes and 175 labels

**Source**

Charte, Francisco and Rivera, Antonio J. and del Jesus, Maria J. and Herrera, Francisco, "QUINTA: A question tagging assistant to improve the answering ratio in electronic forums", in EUROCON 2015 - International Conference on Computer as a Tool (EUROCON), IEEE, pp. 1-6, 2015

**Examples**

```r
## Not run:
stackex_chess()  # Check and load the dataset
toBibtex(stackex_chess)
stackex_chess$measures

## End(Not run)
```

---

**Description**

Dataset from the Stack Exchange’s chess forum

**Usage**

`stackex_chess`

**Format**

An mldr object with 1675 instances, 585 attributes and 227 labels

**Source**

Charte, Francisco and Rivera, Antonio J. and del Jesus, Maria J. and Herrera, Francisco, "QUINTA: A question tagging assistant to improve the answering ratio in electronic forums", in EUROCON 2015 - International Conference on Computer as a Tool (EUROCON), IEEE, pp. 1-6, 2015

**Examples**

```r
## Not run:
toBibtex(stackex_chess)
stackex_chess$measures

## End(Not run)
```
Dataset from the Stack Exchange’s coffee forum

Description
Multilabel dataset from the text domain.

Usage
stackex_coffee()

Format
An mldr object with 225 instances, 1763 attributes and 123 labels

Source
Charte, Francisco and Rivera, Antonio J. and del Jesus, Maria J. and Herrera, Francisco, "QUINTA: A question tagging assistant to improve the answering ratio in electronic forums", in EUROCON 2015 - International Conference on Computer as a Tool (EUROCON), IEEE, pp. 1-6, 2015

Examples
## Not run:
stackex_coffee()
toBibtex(stackex_coffee)
stackex_coffee$measures
## End(Not run)

Dataset from the Stack Exchange’s cooking forum

Description
Multilabel dataset from the text domain.

Usage
stackex_cooking()

Format
An mldr object with 10491 instances, 577 attributes and 400 labels
**Source**

Charte, Francisco and Rivera, Antonio J. and del Jesus, Maria J. and Herrera, Francisco, "QUINTA: A question tagging assistant to improve the answering ratio in electronic forums", in EUROCON 2015 - International Conference on Computer as a Tool (EUROCON), IEEE, pp. 1-6, 2015

**Examples**

```
## Not run:
stackex_cooking() # Check and load the dataset
toBibtex(stackex_cooking)
stackex_cooking$measures

## End(Not run)
```

---

**stackex_cs**

*Dataset from the Stack Exchange's computer science forum*

---

**Description**

Multilabel dataset from the text domain.

**Usage**

```
stackex_cs()
```

**Format**

An mldr object with 9270 instances, 635 attributes and 274 labels

**Source**

Charte, Francisco and Rivera, Antonio J. and del Jesus, Maria J. and Herrera, Francisco, "QUINTA: A question tagging assistant to improve the answering ratio in electronic forums", in EUROCON 2015 - International Conference on Computer as a Tool (EUROCON), IEEE, pp. 1-6, 2015

**Examples**

```
## Not run:
stackex_cs() # Check and load the dataset
toBibtex(stackex_cs)
stackex_cs$measures

## End(Not run)
```
stackex_philosophy Dataset from the Stack Exchange’s philosophy forum

Description
Multilabel dataset from the text domain.

Usage
stackex_philosophy()

Format
An mldr object with 3971 instances, 842 attributes and 233 labels

Source
Charte, Francisco and Rivera, Antonio J. and del Jesus, Maria J. and Herrera, Francisco, "QUINTA: A question tagging assistant to improve the answering ratio in electronic forums", in EUROCON 2015 - International Conference on Computer as a Tool (EUROCON), IEEE, pp. 1-6, 2015

Examples
```r
## Not run:
stackex_philosophy()  # Check and load the dataset
toBibtex(stackex_philosophy)
stackex_philosophy$measures

## End(Not run)
```

stratified.kfolds Partition an mldr object into k folds

Description
This method partitions the given dataset into k folds using a stratified strategy, providing training and test partitions for each fold.

Usage
stratified.kfolds(mld, k = 5, seed = 10)
Arguments

mld The mldr object to be partitioned
k The number of folds to be generated. By default is 5
seed The seed to initialize the random number generator. By default is 10. Change it if you want to obtain partitions containing different samples, for instance to use a 2x5 fcv strategy

Value

An mldr.folds object. This is a list containing k elements, one for each fold. Each element is made up of two mldr objects, called train and test

Examples

## Not run:
library(mldr.datasets)
library(mldr)
folds.emotions <- stratified.kfolds(emotions)
summary(folds.emotions[[1]]$train)
summary(folds.emotions[[1]]$test)

## End(Not run)

tmc2007 Dataset from airplanes failures reports

Description

Multilabel dataset from the text domain.

Usage

tmc2007()

Format

An mldr object with 28596 instances, 49060 attributes and 22 labels

Source

Srivastava, A. N. and Zane-Ulman, B., "Discovering recurring anomalies in text reports regarding complex space systems", Aerospace Conference, pp. 3853-3862, 2005
Examples

```r
## Not run:
tmc2007()  # Check and load the dataset
toBibtex(tmc2007)
tmc2007$measures

## End(Not run)
```

---

tmc2007_500  

*Dataset from airplanes failures reports (500 most relevant features extracted)*

Description

Multilabel dataset from the text domain.

Usage

```r
tmc2007_500()
```

Format

An mldr object with 28596 instances, 500 attributes and 22 labels

Source

Srivastava, A. N. and Zane-Ulman, B., "Discovering recurring anomalies in text reports regarding complex space systems", Aerospace Conference, pp. 3853-3862, 2005

Examples

```r
## Not run:
tmc2007_500()  # Check and load the dataset
toBibtex(tmc2007_500)
tmc2007_500$measures

## End(Not run)
```
**toBibtex.mldr**

BibTeX entry associated to an mldr object

**Description**

Gets the content of the bibtex member of the mldr object and returns it

**Usage**

```r
## S3 method for class 'mldr'

## Not run:
library(mldr.datasets)
cat(toBibtex(emotions))
```

**Arguments**

- `object` The mldr object whose BibTeX entry is needed
- `...` Additional parameters from the generic toBibtex function not used by toBib-

**Value**

A string with the BibTeX entry

**Examples**

```r
## Not run:
library(mldr.datasets)
cat(toBibtex(emotions))
```

**write.mldr**

Export an mldr object or set of mldr objects to different file formats

**Description**

Writes one or more files in the specified formats with the content of the mldr or mldr.folds given as parameter

**Usage**

```r
write.mldr(mld, format = c("MULAN", "MEKA"), sparse = FALSE,
          basename = ifelse(!is.null(mld$name) & nchar(mld$name) > 0,
           regmatches(mld$name, regexpr("(\w)+", mld$name)), "unnamed_mldr"))
```
Arguments

- **mld**: The mldr/mldr.folds object to be exported.
- **format**: A vector of strings stating the desired file formats. It can contain the values 'MULAN', 'MEKA', 'KEEL', 'CSV' and 'LIBSVM'.
- **sparse**: Boolean value indicating if sparse representation has to be used for ARFF-based file formats.
- **basename**: Base name for the files. 'unnamed_mldr' is used by default.

Examples

```r
## Not run:
library(mldr.datasets)
write.mldr(emotions, format = c('CSV', 'KEEL'))

## End(Not run)
```

---

**yahoo_arts**

Dataset generated from the Yahoo! web site index (arts category)

Description

Multilabel dataset from the text domain.

Usage

```r
yahoo_arts()
```

Format

An mldr object with 7484 instances, 23146 attributes and 26 labels.

Source


Examples

```r
## Not run:
yahoo_arts()  # Check and load the dataset
toBibtex(yahoo_arts)
yahoo_arts$measures

## End(Not run)
```
**yahoo_business**

Dataset generated from the Yahoo! web site index (business category)

Description

Multilabel dataset from the text domain.

Usage

```r
yahoo_business()
```

Format

An `mlr` object with 11214 instances, 21924 attributes and 30 labels

Source


Examples

```r
## Not run:
yahoo_business() # Check and load the dataset
toBibTeX(yahoo_business)
yahoo_business$measures
## End(Not run)
```

**yahoo_computers**

Dataset generated from the Yahoo! web site index (computers category)

Description

Multilabel dataset from the text domain.

Usage

```r
yahoo_computers()
```

Format

An `mlr` object with 12444 instances, 34096 attributes and 33 labels
Source


Examples

```r
## Not run:
yahoo_computers() # Check and load the dataset
toBibtex(yahoo_computers)
yahoo_computers$measures

## End(Not run)
```

---

**yahoo_education**

Dataset generated from the Yahoo! web site index (arts education)

Description

Multilabel dataset from the text domain.

Usage

`yahoo_education()`

Format

An mldr object with 12030 instances, 27534 attributes and 33 labels

Source


Examples

```r
## Not run:
yahoo_education() # Check and load the dataset
toBibtex(yahoo_education)
yahoo_education$measures

## End(Not run)
```
Yahoo Entertainment

Dataset generated from the Yahoo! web site index (Arts Entertainment)

Description
Multilabel dataset from the text domain.

Usage
yahoo_entertainment()

Format
An mldr object with 12730 instances, 32001 attributes and 21 labels

Source

Examples
```r
## Not run:
yahoo_entertainment()  # Check and load the dataset
toBibtex(yahoo_entertainment)
yahoo_entertainment$measures

## End(Not run)
```

Yahoo Health

Dataset generated from the Yahoo! web site index (Health Category)

Description
Multilabel dataset from the text domain.

Usage
yahoo_health()

Format
An mldr object with 8205 instances, 30605 attributes and 32 labels
Source


Examples

```r
## Not run:
yahoo_health()  # Check and load the dataset
toBibtex(yahoo_health)
yahoo_health$measures

## End(Not run)
```

**yahoo_recreation**

Dataset generated from the Yahoo! web site index (recreation category)

Description

Multilabel dataset from the text domain.

Usage

```r
yahoo_recreation()
```

Format

An mldr object with 12828 instances, 30324 attributes and 22 labels

Source


Examples

```r
## Not run:
yahoo_recreation()  # Check and load the dataset
toBibtex(yahoo_recreation)
yahoo_recreation$measures

## End(Not run)
```
Yahoo Reference Dataset

Description
Multilabel dataset from the text domain.

Usage
yahoo_reference()

Format
An mldr object with 8027 instances, 39679 attributes and 33 labels

Source

Examples
```r
## Not run:
yahoo_reference() # Check and load the dataset
toBibtex(yahoo_reference)
yahoo_reference$measures

## End(Not run)
```

Yahoo Science Dataset

Description
Multilabel dataset from the text domain.

Usage
yahoo_science()

Format
An mldr object with 6428 instances, 37187 attributes and 40 labels
Yahoo Social

Source


Examples

## Not run:
yahoo_social() # Check and load the dataset
toBibtex(yahoo_social)
yahoo_social$measures

## End(Not run)

---

yahoo_social  

Dataset generated from the Yahoo! web site index (social category)

Description

Multilabel dataset from the text domain.

Usage

yahoo_social()

Format

An mldr object with 12111 instances, 52350 attributes and 39 labels

Source


Examples

## Not run:
yahoo_social() # Check and load the dataset
toBibtex(yahoo_social)
yahoo_social$measures

## End(Not run)
yahoo_society

Dataset generated from the Yahoo! web site index (society category)

Description
Multilabel dataset from the text domain.

Usage

```
yahoo_society()
```

Format
An mldr object with 14512 instances, 31802 attributes and 27 labels

Source

Examples

```
## Not run:
yahoo_society()  # Check and load the dataset
toBibtex(yahoo_society)
yahoo_society$measures

## End(Not run)
```

yeast

Dataset with protein profiles and their categories

Description
Multilabel dataset from the biology domain.

Usage

```
yeast()
```

Format
An mldr object with 2417 instances, 103 attributes and 14 labels
Source

Examples

```r
## Not run:
yeast()  # Check and load the dataset
toBibtex(yeast)
yeast$measures

## End(Not run)
```
Index

*Topic datasets
- birds, 4
- cal500, 5
- emotions, 14
- flags, 20
- genbase, 20
- langlog, 22
- medical, 23
- ng20, 24
- slashdot, 32
- stackex_chess, 33

- bibtex, 3
- birds, 4
- bookmarks, 4
- cal500, 5
- check_n_load.mldr, 6
- corel16k001, 6
- corel16k002, 7
- corel16k003, 8
- corel16k004, 8
- corel16k005, 9
- corel16k006, 10
- corel16k007, 10
- corel16k008, 11
- corel16k009, 12
- corel16k010, 12
- corel5k, 13

- delicious, 14
- emotions, 14
- enron, 15
- eurlexdc_test, 16
- eurlexdc_tra, 16
- eurlexev_test, 17
- eurlexev_tra, 18
- eurlexsm_test, 18
- eurlexsm_tra, 19

- flags, 20
- genbase, 20
- imdb, 21
- langlog, 22
- mediamill, 22
- medical, 23
- mldrs, 24
- ng20, 24
- nuswide_BoW, 25
- nuswide_VLAD, 25
- ohsumed, 26
- random.kfolds, 27
- rcv1sub1, 27
- rcv1sub2, 28
- rcv1sub3, 29
- rcv1sub4, 29
- rcv1sub5, 30
- reutersk500, 31
- scene, 31
- slashdot, 32
- stackex_chemistry, 32
- stackex_chess, 33
- stackex_coffee, 34
- stackex_cooking, 34
- stackex_cs, 35
- stackex_philosophy, 36
- stratified.kfolds, 36
- tmc2007, 37
- tmc2007_500, 38
- toBibtex.mldr, 39

- write.mldr, 39
yahoo_arts, 40
yahoo_business, 41
yahoo_computers, 41
yahoo_education, 42
yahoo_entertainment, 43
yahoo_health, 43
yahoo_recreation, 44
yahoo_reference, 45
yahoo_science, 45
yahoo_social, 46
yahoo_society, 47
yeast, 47