Package 'smallsets'

December 5, 2023

Title Visual Documentation for Data Preprocessing

Version 2.0.0

Maintainer Lydia R. Lucchesi <Lydia.Lucchesi@anu.edu.au>

Description Data practitioners regularly use the 'R' and 'Python' programming languages to prepare data for analyses. Thus, they encode important data preprocessing decisions in 'R' and 'Python' code. The 'smallsets' package subsequently decodes these decisions into a Smallset Timeline, a static, compact visualisation of data preprocessing decisions (Lucchesi et al. (2022) <doi:10.1145/3531146.3533175>). The visualisation consists of small data snapshots of different preprocessing steps. The 'smallsets' package builds this visualisation from a user's dataset and preprocessing code located in an 'R', 'R Markdown', 'Python', or 'Jupyter Notebook' file. Users simply add structured comments with snapshot instructions to the preprocessing code. One optional feature in 'smallsets' requires installation of the 'Gurobi' optimisation software and 'gurobi' 'R' package, available from <https://www.gurobi.com>. More information regarding the optional feature and 'gurobi' installation can be found in the 'smallsets' vignette.

URL https://lydialucchesi.github.io/smallsets/, https://github.com/lydialucchesi/smallsets

License GPL (>= 3)

Encoding UTF-8

LazyData TRUE

RoxygenNote 7.2.3

Depends R (>= 3.5.0)

Imports callr, colorspace, flextable, ggplot2, ggtext, knitr, patchwork, plotrix, reticulate, rmarkdown

Suggests gurobi, testthat (>= 3.0.0)

VignetteBuilder knitr

Config/testthat/edition 3

NeedsCompilation no

Author Lydia R. Lucchesi [aut, cre] (<https://orcid.org/0000-0002-1901-4301>), Petra M. Kuhnert [ths], Jenny L. Davis [ths], Lexing Xie [ths]

Repository CRAN

Date/Publication 2023-12-05 00:00:02 UTC

R topics documented:

`sets_labelling` *Sets labelling*

Description

Sets labelling parameters for the Smallset Timeline.

Usage

```
sets_labelling(labelCol = NULL, labelColDif = NULL)
```

Arguments

`labelCol`	Either "lighter" or "darker" for colour of column names and printed data, in comparison to tile colours. Default is "darker".
`labelColDif`	Value between 0-1 determining how much lighter or darker. Default is .5.

Details

Passed to `labelling` in Smallset_Timeline.

Value

Returns a list with two elements (the labelling parameters).

Examples

```
# labels and printed data are black
Smallset_Timeline(
  data = s_data,
  code = system.file("s_data_preprocess.R", package = "smallsets"),
  printedData = TRUE,
  truncateData = 4,
  labelling = sets_labelling(labelCol = "darker", labelColDif = 1))
```

```
# labels and printed data are midpoint between tile colour and white
Smallset_Timeline(
  data = s_data,
  code = system.file("s_data_preprocess.R", package = "smallsets"),
  colours = 3,
  printedData = TRUE,
  truncateData = 4,
  labelling = sets_labelling(labelCol = "lighter", labelColDif = .5))
```

`sets_sizing` *Sets sizing*

Description

Sets sizing parameters for the Smallset Timeline.

Usage

```
sets_sizing(
  captions = NULL,
  columns = NULL,
  data = NULL,
  icons = NULL,
  legend = NULL,
  resume = NULL,
  tiles = NULL
)
```

Arguments

`captions`	Positive numeric value for caption text size. Default is 3.
`columns`	Positive numeric value for column name text size. Default is 3.
`data`	Positive numeric value for printed data text size. Default is 2.
`icons`	Positive numeric value for legend icon size. Default is 1.
`legend`	Positive numeric value for legend text size. Default is 10.
`resume`	Positive numeric value for resume marker size. Default is 1.
`tiles`	Positive numeric value for Smallset tile size. Default is .2.

Details

Passed to `sizing` in Smallset_Timeline.

Value

Returns a list with seven elements (the sizing parameters).

Examples

```
# increase size of caption text
# and add more caption space, so larger caption text fits
Smallset_Timeline(
   data = s_data,
   code = system.file("s_data_preprocess.R", package = "smallsets"),
   sizing = sets_sizing(captions = 4),
   spacing = sets_spacing(captions = 4)
)
```

`sets_spacing` *Sets spacing*

Description

Sets spacing parameters for the Smallset Timeline.

Usage

```
sets_spacing(
  captions = NULL,
  degree = NULL,
  header = NULL,
  right = NULL,
  rows = NULL,
  columns = NULL
)
```

Arguments

`captions`	Positive numeric value for amount of caption space. When alignment is horizontal, refers to the space below snapshots. When alignment is vertical, refers to the space to the right of the snapshots. Default is 2.5.
`degree`	Integer between 0-90 (degrees) to rotate column names. Default is 0.
`header`	Positive numeric value for amount of column name space. Default is 1.
`right`	Positive numeric value (>=.5) for amount of space to the right of each snapshot.
`rows`	Integer for number of Smallset Timeline rows (applicable when the alignment is horizontal). Default is 1.
`columns`	Integer for number of Smallset Timeline columns (applicable when the alignment is vertical). Default is 1.

Details

Passed to `spacing` in Smallset_Timeline.

Value

Returns a list with six elements (the spacing parameters).

Examples

```
# increase space for captions and rotate column names
Smallset_Timeline(
   data = s_data,
   code = system.file("s_data_preprocess.R", package = "smallsets"),
   spacing = sets_spacing(captions = 5, degree = 45)
)
```

Smallset_Timeline *Smallset Timeline*

Description

Builds a Smallset Timeline to visualise data preprocessing decisions.

Usage

```
Smallset_Timeline(
  data,
  code,
  rowCount = 5,
  rowSelect = NULL,
  rowReturn = FALSE,
  rowIDs = NULL,
  ignoreCols = NULL,
  colours = 1,
  printedData = FALSE,
  truncateData = NULL,
  ghostData = TRUE,
  missingDataTints = FALSE,
  align = "horizontal",
  font = "sans",
  sizing = sets_sizing(),
  spacing = sets_spacing(),
  labelling = sets_labelling(),
  altText = FALSE
)
```

Arguments

data	Dataset that is being preprocessed.

code	R, R Markdown, Python, or Jupyter Notebook data preprocessing script. Include the filename extension (e.g., "my_code.R", "my_code.Rmd", "my_code.py", or "my_code.ipynb"). If the script is not in the working directory, include the full file path.
rowCount	Integer between 5-15 for number of Smallset rows.
rowSelect	NULL, 1, or 2. If NULL, Smallset rows are randomly sampled. If 1, Smallset rows are selected using the coverage optimisation model. If 2, Smallset rows are selected using the coverage + variety optimisation model, which has a long run time for large datasets. Options 1 and 2 use the Gurobi solver (v9.1.2) and require a Gurobi license. Please visit https://www.gurobi.com to obtain a license (free academic licenses are available).
rowReturn	A logical. TRUE prints, to the console, the row numbers of the rows selected for the Smallset.
rowIDs	If R preprocessing code, a character vector of row names for rows to include in the Smallset. If Python preprocessing code, a numeric vector of indices for rows to include in the Smallset.
ignoreCols	Character vector of column names indicating which to exclude from the Smallset Timeline.
colours	Either 1, 2, or 3 for one of the built-in colour schemes (all are colourblind-friendly and 2 works on a grey scale, i.e., it's printer-friendly) or a list with four hex colour codes for added, deleted, edited, and unchanged (e.g., list(added = "#5BA2A6", deleted = "#DDC492", edited = "#FFC500", unchanged = "#E6E3DF")).
printedData	A logical. TRUE prints data values in the Smallset snapshots.
truncateData	Integer for the number of characters in each printed data value. Results in characters plus an ellipsis.
ghostData	A logical. TRUE includes empty tiles where data have been removed.
missingDataTints	A logical. TRUE plots a lighter colour value for a missing data value.
align	Either "horizontal" or "vertical". For horizontal, snapshots are plotted left to right. For vertical, snapshots are plotted top to bottom.
font	Any font installed in R.
sizing	sets_sizing for size specifications.
spacing	sets_spacing for space specifications.
labelling	sets_labelling for label specifications.
altText	A logical. TRUE generates alternative text (alt text) for the Smallset Timeline and prints it to the console.

Details

Prior to running this command, structured comments with snapshot instructions must be added to the preprocessing script passed to code. See section titled "Structured comments" in `vignette("smallsets")` or in the online user guide.

Value

Returns a Smallset Timeline object, which is a plot consisting of ‘ggplot‘ objects assembled with ‘patchwork‘.

Examples

```
set.seed(145)

Smallset_Timeline(
  data = s_data,
  code = system.file("s_data_preprocess.R", package = "smallsets")
)
```

---

s_data *Synthetic dataset*

Description

A synthetic dataset generated for illustrative purposes.

Usage

```
s_data
```

Format

A data frame with 100 rows and 8 columns:

C1 discrete

C2 binary

C3 discrete

C4 discrete

C5 continuous

C6 continuous

C7 continuous

C8 continuous

Examples

```
str(s_data)
```

Index