Smallset features

Lydia Lucchesi

28/04/2021

Smallset comments

There are four types of smallset comments. These are added to the data preprocessing script to inform the smallset package where to start tracking the data preprocessing, take snapshots of the data, and end tracking.

- Start comment
  - `# start smallset`
  - This comment tells the package to start tracking the code. There may be code included in the script prior to the preprocessing of interest. This tells smallset to ignore that code. When the tracking starts, a snapshot of the data is automatically taken.
- Snap comment
  - `# snap` + name of data object (`# snap mydata`)
  - This comment tells the package to take a snapshot of the data. The snapshot is taken after the line of R code below the snap comment. After `# snap`, the data object name must be included. The data object name can change throughout the preprocessing.
- Resume comment
  - `# resume smallset`
  - This comment is optional. Sometimes, a data practioner will stop preprocessing and start analysing the data only to realise that additional preprocessing must be done. This comment tells the package that the preprocessing was stopped and then restarted, and a visual marker will be included in the timeline to represent this aspect of the process.
- End comment
  - `# end smallset`
  - This comment tells the package to end tracking. There may be code included in the script after the preprocessing part. This tells smallset to ignore that code. When the tracking ends, a snapshot of the data is automatically taken.

Example

```
other code
other code

# start smallset
# snap mydata
preprocessing command
# snap mydata
preprocessing command
preprocessing command
preprocessing command
# snap mydata2
preprocessing command
# end smallset
```

```
other code
other code
```

First function: `prepare_smallset`

```
prepare_smallset(
  data,
  code,
  dir = getwd(),
  rowCount = 6,
  rowNums = NULL,
  runBig = FALSE,
  ignoreCols = NULL,
  captionTemplateName = "captionTemplate",
  captionTemplateDir = getwd(),
  captionTemplateAuthor = NULL
)
```

This function selects a smallset, takes snapshots, and identifies the changes between snapshots. The output can be passed to the second function, `create_timeline`, to create a smallset timeline.

Arguments

- data
  - The data set being preprocessed.
- code
  - The data preprocessing code.
- dir
  - The file path to the data preprocessing code.
- rowCount
  - Integer greater than or equal to five. Number of rows to include in the smallset. Default is six rows.
- rowNums
  - Numeric vector of row numbers. Indicates particular rows from the data set to include in the smallset. It must be less than or equal to the rowCount.
- runBig
  - TRUE or FALSE. FALSE means preprocessing code will be run on smallset. TRUE means preprocessing code will be run on the data set passed to the data argument, and the smallset will be extracted from that output at each snap point.
- ignoreCols
  - Character vector of column names. Indicates which columns from the data set should not be included in the smallset. Columns in this vector should usually not be referenced in the data preprocessing code.
- captionTemplateName
  - File name for the caption template.
- captionTemplateDir
  - File path for the caption template.
- captionTemplateAuthor
  - Name of author for the caption template.

Caption template

Second function: `create_timeline`

```
create_timeline(
  snapshotList,
  constant = list("#cecfd6", 0.8),
  changed = list("#0f3d1c", 0.8),
  added = list("#a35222", 0.8),
  deleted = list("#3e4d63", 0.8),
  colScheme = NULL,
  abstract = TRUE,
  ghostData = TRUE,
  highlightNA = FALSE,
  sizing = list(columns = 2, tiles = 1, captions = 8, data = 2.5, legendText = 7,
    legendIcons = 1, title = 10, subtitle = 8, footnote = 7, resume = 0.25),
  truncateData = FALSE,
  accentCols = "darker",
  accentColsDif = 0.5,
  otherTextCol = 1,
  timelineRows = NULL,
  timelineFont = "sans",
  captionSpace = 1,
  captionTemplateName = NULL,
  captionTemplateDir = NULL
)
```

This function creates a smallset timeline using output from prepare_smallset. Timelines have many customisation options. They are detailed below.

Arguments

- snapshotList
  - List output from `prepare_smallset`.
- constant
  - Hex colour code. Colour represents data that have not changed since previous snapshot. Can pass in a list with a colour and transparency value (0 to 1) for that colour.
- changed
  - Hex colour code. Colour represents data that have changed since previous snapshot. Can pass in a list with a colour and transparency value (0 to 1) for that colour.
- added
  - Hex colour code. Colour represents data that have been added since previous snapshot. Can pass in a list with a colour and transparency value (0 to 1) for that colour.
- deleted
  - Hex colour code. Colour represents data that will be deleted prior to next snapshot. Can pass in a list with a colour and transparency value (0 to 1) for that colour.
- colScheme
  - NULL, colour scheme name, or vector. If NULL, uses four colour arguments above. If colour scheme name, uses built-in scheme with colours pre-assigned to the four preprocessing states (constant, changed, added, deleted). If vector, it must be a vector of length five, with the first element being the colour scheme name followed by the four preprocessing states in the order that they should be assigned to scheme colours (e.g., c("colScheme1", "changed", "constant", "deleted", "added")).
- abstract

  - TRUE or FALSE. FALSE prints data values in tables.
- ghostData
  - TRUE or FALSE. TRUE includes blank tiles where data have been removed.
- highlightNA
  - TRUE or FALSE. TRUE plots a lighter colour value to signal data value is missing.
- sizing
  - List of size specifications. Can specify sizes for column names, table tiles, caption text, printed data, legend text, legend icons, timeline title, timeline subtitle, timeline footnote, and preprocessing resume indicator.
- truncateData
  - TRUE or FALSE. FALSE if data do not need to be truncated to fit within table tiles. Otherwise, an integer specifying width of data value (width includes "...").
- accentCols
  - Either "darker" or "lighter" for stamp colour. Can enter a list corresponding to specific actions.
- accentColsDif
  - Value between 0 and 1. Corresponds to how much lighter or darker accent colour will be. Can pass a list with different accent values for different colours.
- otherTextCol
  - Value between 0 and 1. Default is 1, which is when column names are black. 0 means columns will be the constant colour.
- timelineRows
  - Integer greater than or equal to one. Number of rows to divide the smallset timeline into.
- timelineFont
  - Choose one of sans, serif, or mono.
- captionSpace
  - Value greater than or equal to .5. Higher values create more caption space. Default is 1.
- captionTemplateName
  - Name of caption template. Can be included so template is not overwritten when running `prepare_smallset`.
- captionTemplateDir
  - Name of caption template directory. Can be included so template is not overwritten when running `prepare_smallset`.

One timeline designed several different ways

Set up

```
library(smallset)
```

```
## Registered S3 method overwritten by 'gdata':
##   method         from
##   reorder.factor gplots
```

```
set.seed(6)
bb <- read.csv("~/Desktop/bbData.csv")

columnNames <- c(
  "Day",
  "Tmin",
  "Tmax",
  "Rain",
  "Temp9",
  "Hum9",
  "Cloud9",
```

```
  "Dir9",
  "Speed9"
)

colnames(bb) <- columnNames

bb$Dir9 <- ifelse(bb$Dir9 == "", NA, bb$Dir9)
bb$Speed9 <- ifelse(bb$Speed9 == "", NA, bb$Speed9)

mylist <- prepare_smallset(
  data = bb,
  code = "~/Desktop/batemansProcess.R",
  rowCount = 6,
  rowNums = c(1, 28),
  runBig = TRUE,
  ignoreCols = NULL,
  captionTemplateName = "mycaptions"
)
```

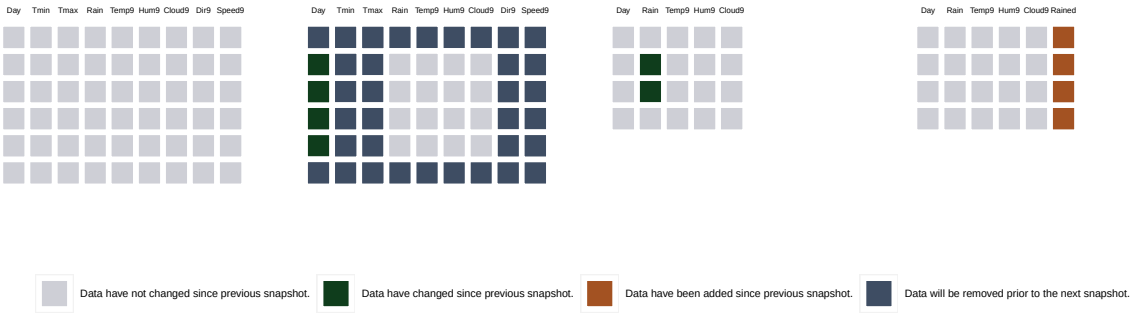
```
## [1] "Summary: 4 snapshots taken"
## [1] "First snapshot:"
##           Day Tmin Tmax Rain Temp9 Hum9 Cloud9 Dir9 Speed9
## 1   2021-01-1   NA   NA   NA    NA   NA     NA   NA     NA
## 11 2021-01-11 13.5 29.6  0.0  23.4   72      2   NA     NA
## 14 2021-01-14 20.6 32.2  0.2  27.4   69      0    5      2
## 15 2021-01-15 14.6 25.5  0.0  21.4   75      0    4      2
## 22 2021-01-22 14.2 38.1  0.0  26.5   68      0   NA     NA
## 28 2021-01-28   NA 19.0   NA    NA   NA     NA   NA     NA
## [1] "Last snapshot:"
##    Day Rain Temp9 Hum9 Cloud9 Rained
## 11  11  0.0  23.4   72      2      0
## 14  14  0.8  27.4   69      0      1
## 15  15  0.2  21.4   75      0      1
## 22  22  0.0  26.5   68      0      0
## [1] "Edits, additions, and deletions identified and mycaptions.Rmd file created at /Users/luc093/Desk
```

Example 1

```
check <- create_timeline(
  snapshotList = mylist,
  constant = list("#cecfd6", 1),
  changed = list("#0f3d1c", 1),
  added = list("#a35222", 1),
  deleted = list("#3e4d63", 1),
  abstract = TRUE,
  ghostData = FALSE,
  highlightNA = FALSE,
  sizing =
    list(
      "columns" = 1,
      "tiles" = 1,
      "legendText" = 5,
      "legendIcons" = 1
```

```
    ),
  accentCols = "darker",
  accentColsDif = .9,
  otherTextCol = 1,
  timelineRows = 1,
  timelineFont = "sans"
)

check
```

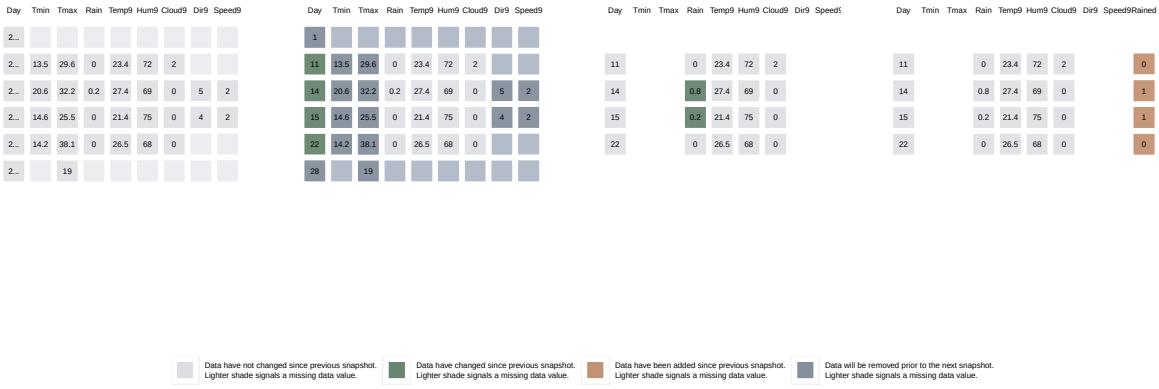


Example 2

```
check <- create_timeline(
  snapshotList = mylist,
  constant = list("#cecfd6", .6),
  changed = list("#0f3d1c", .6),
  added = list("#a35222", .6),
  deleted = list("#3e4d63", .6),
  abstract = FALSE,
  ghostData = TRUE,
  highlightNA = TRUE,
  sizing =
    list(
      "columns" = 1,
      "tiles" = 1,
      "legendText" = 4,
      "legendIcons" = .75,
      "captions" = 1.3,
      "title" = 8,
      "subtitle" = 6,
      "footnote" = 4,
      "data" = 1

    ),
  truncateData = 4,
  accentCols = "darker",
  accentColsDif = 1,
  otherTextCol = 1,
  timelineRows = 1,
  timelineFont = "sans",
  captionSpace = 4
)
```

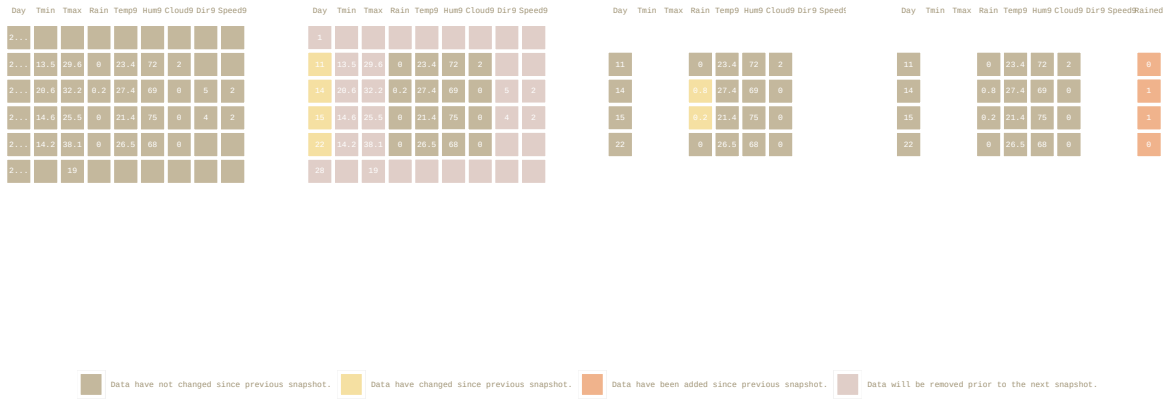
```
check
```



Example 3

```
check <- create_timeline(
  snapshotList = mylist,
  constant = list("#C4B89D", 1),
  changed = list("#F5E0A2", 1),
  added = list("#F0B38C", 1),
  deleted = list("#E0CEC8", 1),
  abstract = FALSE,
  ghostData = TRUE,
  highlightNA = FALSE,
  sizing =
    list(
      "columns" = 1,
      "tiles" = .6,
      "legendText" = 4,
      "legendIcons" = .75,
      "captions" = 1.2,
      "title" = 8,
      "subtitle" = 6,
      "footnote" = 4,
      "data" = 1
    ),
  truncateData = 4,
  accentCols = "lighter",
  accentColsDif = 1,
  otherTextCol = .3,
  timelineRows = 1,
  timelineFont = "mono",
  captionSpace = 4.5
)

check
```


Day Tmin Tmax Rain Temp9 Hum9 Cloud9 Dir9 Speed9
2... 13.5 29.6 0 23.4 72 2
2... 20.6 32.2 0.2 27.4 69 0 5 2
2... 14.6 25.5 0 21.4 75 0 4 2
2... 14.2 38.1 0 26.5 68 0
2... 19
Day Tmin Tmax Rain Temp9 Hum9 Cloud9 Dir9 Speed9
1
11 13.5 29.6 0 23.4 72 2
14 20.6 32.2 0.2 27.4 69 0 5 2
15 14.6 25.5 0 21.4 75 0 4 2
22 14.2 38.1 0 26.5 68 0
28 19
Day Tmin Tmax Rain Temp9 Hum9 Cloud9 Dir9 Speed9
11 0 23.4 72 2
14 0.8 27.4 69 0
15 0.2 21.4 75 0
22 0 26.5 68 0
Day Tmin Tmax Rain Temp9 Hum9 Cloud9 Dir9 Speed9 Rained
11 0 23.4 72 2 0
14 0.8 27.4 69 0 1
15 0.2 21.4 75 0 1
22 0 26.5 68 0 0
Data have not changed since previous snapshot.
Data have changed since previous snapshot.
Data have been added since previous snapshot.
Data will be removed prior to the next snapshot.