

Processing Observation Data from Learning Experiments

Lewis Marsh and Luis P. Prieto

08 June, 2018

This R Markdown report aims to summarise the analyses done on the observation data of several sessions of innovative learning activities, from which multimodal data has been gathered

Observation data

Preprocessing

The preprocessing is taken care by the R script in `src/processObservationData.R`. It consists of two main functions, `processObservationData` (to process the data of one session) and `processAllObservationData` (for multiple sessions).

The first of these two functions,

```
data <- processObservationData(raw_data,
                              date=as.POSIXct(strptime("10-01-2018", "%d-%m-%Y")),
                              project="Isle",
                              activitycol=F,
                              observercol=F,
                              namecols=c("timestamp","group",
                                           "StudentA","StudentB","StudentC","StudentD"))
```

takes in a dataframe with the raw observation data, together with some other parameters about the structure of the raw dataset and expected columns in the processed dataset, and returns a clean data frame with the observations for each individual.

In the original data sheets, each entry lists the observation for a whole group of students (between 3 and 6 usually) and the observations are listed as a piece of (predefined) text, such as “Talking with their group peers to solve the task” or “Totally disengaged”.

In the output this is broken up into individual entries for each student, for which a student variable is introduced. Furthermore, each possible observation is turned into a boolean variable for each of these entries, which makes it easier to conduct further statistical analysis of the dataset. Also, a column with additional notes from the original dataframe is dropped in the output, as it is difficult to make use of in a statistical test, and a global ID is created for each student, using the date, group, student number in the group and the project name. This will be of use when merging several datasets with the second function.

As students undertook several tasks in some of the experiments/projects, the input variable “activitycol” indicates whether the function has to look for an extra activity column in the input data. This will be taken over one-to-one in the output or will be filled in with “Standard” if there was only one activity. Similarly, in some projects there were more than one observer placed onto each group to enable to test the reliability of the observers and the data they produce. The variable “observercol” tells the function whether there is a separate observer column to take over from the input data. Otherwise “1-A” is filled in for each observer.

The second function, `processAllObservationData` takes in a vector of URLs in which the GoogleDocs containing the experimental data are stored in. It reads the data from the online sheets and places them in a dataframe. By examining the names of the columns in the dataframe, it detects whether there are activity or observation columns. By inspecting the first time stamp, it determines the date of the experiment. It then

cleans the data using the first function and adds it to the end of a larger dataframe which eventually gets returned after the function has iterated through all URLs.

Example of Output

To exhibit what the functions do, there is a brief example. We are given a data frame taken from one of the GoogleDocs:

```
raw_data[1,1:2]
```

```
##           Timestamp What group are you observing?
## 1 11/10/2017 10:18:07                Group 1
```

```
raw_data[1,3]
```

```
## [1] "Talking with their group peers to solve the task"
```

These are only two snippets of the data, as anything else would not fit onto this page. We see that the first two columns give a timestamp and the group. We furthermore see the observation for student A. There are 3 other students in the group with individual columns as well as a column for further comments.

By applying the `processObservationData` function, we then see how the student variable is created and the observations are turned from text into boolean values. Furthermore, the activity and observer column have been created which makes it easier to merge and compare this data with other datasets.

```
processed_data <- processAllObservationData()
summary(processed_data)
```

```
##      timestamp           group           student      disengaged
## Length:4319      Length:4319      Student B:920      Min.      :0.00000
## Class :character      Class :character      Student C:903      1st Qu.:0.00000
## Mode  :character      Mode  :character      Student D:897      Median :0.00000
##                                           Student A:853      Mean   :0.08868
##                                           Student E:624      3rd Qu.:0.00000
##                                           ann      : 62      Max.    :1.00000
##                                           (Other)  : 60
##      looking           talking           technology      resources
## Min.      :0.0000      Min.      :0.0000      Min.      :0.0000      Min.      :0.0000
## 1st Qu.:0.0000      1st Qu.:0.0000      1st Qu.:0.0000      1st Qu.:0.0000
## Median :0.0000      Median :1.0000      Median :0.0000      Median :0.0000
## Mean   :0.2135      Mean   :0.6085      Mean   :0.4971      Mean   :0.3265
## 3rd Qu.:0.0000      3rd Qu.:1.0000      3rd Qu.:1.0000      3rd Qu.:1.0000
## Max.    :1.0000      Max.    :1.0000      Max.    :1.0000      Max.    :1.0000
##
##      external           student.id           activity           observer
## Min.      :0.0000      Length:4319      Length:4319      Length:4319
## 1st Qu.:0.0000      Class :character      Class :character      Class :character
## Median :0.0000      Mode  :character      Mode  :character      Mode  :character
## Mean   :0.2753
## 3rd Qu.:1.0000
## Max.    :1.0000
##
##      project           date           global.id
## Length:4319      Min.      :2017-08-11      Length:4319
## Class :character      1st Qu.:2017-10-11      Class :character
## Mode  :character      Median :2017-11-22      Mode  :character
##                                           Mean   :2017-11-06
```

```
##          3rd Qu.:2017-12-13
##          Max.    :2018-01-10
##
```

```
str(processed_data)
```

```
## 'data.frame':    4319 obs. of  15 variables:
## $ timestamp : chr  "11/10/2017 10:18:07" "11/10/2017 10:25:51" "11/10/2017 10:28:37" "11/10/2017 10
## $ group      : chr  "Group 1" "Group 1" "Group 1" "Group 1" ...
## $ student    : Factor w/ 11 levels "Student A","Student B",...: 1 1 1 1 1 1 1 1 1 1 ...
## $ disengaged: num   0 0 0 0 0 0 0 0 0 0 ...
## $ looking    : num   0 0 0 0 0 0 0 0 0 0 ...
## $ talking     : num   1 1 0 1 1 1 0 0 0 0 ...
## $ technology: num   0 1 1 1 0 0 1 1 1 0 ...
## $ resources   : num   0 1 1 1 0 0 1 1 1 0 ...
## $ external    : num   0 0 0 0 0 0 0 0 0 1 ...
## $ student.id: chr   "Group 1 Student A" "Group 1 Student A" "Group 1 Student A" "Group 1 Student A"
## $ activity    : chr   "Standard" "Standard" "Standard" "Standard" ...
## $ observer    : chr   "1-A" "1-A" "1-A" "1-A" ...
## $ project     : chr   "Linnaruum" "Linnaruum" "Linnaruum" "Linnaruum" ...
## $ date        : Date, format: "2017-10-11" "2017-10-11" ...
## $ global.id   : chr   "Linnaruum 2017-10-11 Group 1 Student A" "Linnaruum 2017-10-11 Group 1 Student A"
```

Spot checking the data

Some tables to see what kind of data we have

```
table(processed_data$date, processed_data$student)
```

```
##
##          Student A Student B Student C Student D Student E Student F
## 2017-08-11         160         161         160         161          87          0
## 2017-10-11         148         144         132         125         143          2
## 2017-10-18          28          28          29          24           6         14
## 2017-11-22         164         164         163         164           0          0
## 2017-12-06         169         169         169         169         166          0
## 2017-12-13         184         182         178         182         150          0
## 2018-01-10           0          72          72          72          72          0
##
##          Student G ann Student H Student I Student J
## 2017-08-11          1   0          0          0          0
## 2017-10-11          2   0          0          0          0
## 2017-10-18         14   0         12         10          5
## 2017-11-22          0   0          0          0          0
## 2017-12-06          0  62          0          0          0
## 2017-12-13          0   0          0          0          0
## 2018-01-10          0   0          0          0          0
```

@Lewis ... what is this “ann” student??

MCA

....