Codebook

Hiro

2021/10/29

Summary

- This codebook explains the data source (original text files), variables, function, names of the measurement variables in the final output "selectedGroupedSummary", data table transformation, and tidiness of the data tables.

- Data table, "dataAll" is the 1st deliverable of combined one set of data.

- "selected" is the subset table, name of which includes "-mean()" or "-std()".

- "selectedGrouped" is the grouped table by "Subject" and "Activity".

- Final output is the "selectedGroupedSummary", which shows the averages of each selected measurement variables.

- Tidy data must own the properties: (1). Each variable must have its own column, (2). Each observation must have its own row., and (3). Each value must have its own cell.
- "dataAll" is a tidy data table, since it meets all of the tree conditions of tidiness of data. Specifically, (1) each variables, "Subject", "Activity", and 561 measurement variables, e.g."tBodyAcc-mean()-X" has its own column. (2) There are 10,299 observations having its own row. (3) Each value has its own cell.

- Since "dataAll" is a tidy dataset, derived datasets, namely, "selected", "selectegGrouped", and final output, "selectedGroupedSummary" are all "tidy".

Specifics

- Original files were obtained from the web-site ("WEBSITE") below.

- https://d396qusza40orc.cloudfront.net/getdata%2Fprojectfiles%2FUCI%20HAR%20Dataset.zip

Text files collected from the "WEBSITE"
X_test.txt : Test set
y_test.txt : Test labels
subject_test.txt : Links the subject with Test set
X_train.txt : Training set
y_train.txt : Training labels
subject_train.txt : Links the subject with Train set
activity_labels.txt : Links the class labels with their activity name
features.txt : List of all features

Text files stored as data.table
data_X_test : X_test.txt
data_y_test : y_test.txt
data_subject_test : subject_test.txt
data_X_train : X_train.txt
data_y_train : y_train.txt
data_subject_train : subject_train.txt
data_activity_labels : activity_labels.txt
data_features : features.txt

Data transformation
test : temporary created data.table
train : temporary created data.table
dataAll : combined table consisting of "test" and "train" table
selected : subset data.table with variables, names of which are "Subject", "Activity", and those containing "-mean()" or "-std()"
selectedGrouped : "selected" data.table was transformed being grouped by "Subject" and "Activity"
selectedGroupedSummary : final tidy data set containing averages of each measurement of "selectedGrouped"

Other variables
g : extracted characters ending as "mean()" or "std()" in the "features" data.
activityNames : descriptive names of activities
valueNames : descriptive names of measurement variables

Function
f : created to to replace values of "Activity" with descriptive activity names
1: WALKING
2: WALKING_UP
3: WALKING_DW
4: SITTING
5: STANDING
6: LAYING

Fixed variables
(Subject)
1:30 : subject 1 to 30

(Activity)
WALKING : Activity
WALKING_UP : Activity
WALKING_DW : Activity
SITTING : Activity STANDING : Activity
LAYING : Activity

Name of the variables(average) in the final output ("selectedGroupedSummary")
mean(tBodyAccMagMean)
mean(tBodyAccMagStd)
mean(tGravityAccMagMean)
mean(tGravityAccMagStd)
mean(tBodyAccJerkMagMean)
mean(tBodyAccJerkMagStd)
mean(tBodyGyroMagMean)
mean(tBodyGyroMagStd)
mean(tBodyGyroJerkMagMean)
mean(tBodyGyroJerkMagStd)
mean(fBodyAccMagMean)
mean(fBodyAccMagStd)

mean(fBodyBodyAccJerkMagMean)
mean(fBodyBodyAccJerkMagStd)
mean(fBodyBodyGyroMagMean)
mean(fBodyBodyGyroMagStd)
mean(fBodyBodyGyroJerkMagMean)
mean(fBodyBodyGyroJerkMagStd)

Data Transformations

- Original data files obtained from the WEBSITE were first stored in data, e.g. “data_X_test”.
- “test” and “train” are temporary tables, that merges “Subject” and “Activity”, horizontally.
- “dataAll” is one data set, created by merging “train” and “test”, vertically.
- “selected” is a subset table from “dataAll”, selecting “Subject”, “Activity”, and all of the columns name of which contains “mean()” or “std()” at the end. In this process, “grep” function was applied to extract them from the “data_features”, and stored in variable “g”.
- At this timing, “Activity” values were transformed as descriptive names, e.g. “WALKING”. In this process, customary made function “f” was used as well as “lappy” and “unlist” functions.
- Next, lengthy variable names containing “-mean()” or“-std()” were appropriately adjusted by applying “sub” function and “regular expression”.
- Then, “selectedGrouped” table was created by grouping “Subject”, “Activity”, byh applying “grouped_by” function.
- Finally, “selectedGrouped” was summarized by calculating average or mean of values of each variables. The Number of the variables applied is eighteen(18). The resulted name of the tidy data variables are automatically assigned, e.g.“mean(tBodyAccMagMean)”.
- The final output was named as “selectedGroupedSummary”, consisting of 180 rows and 20 columns including two fixed variables “Subject”and “Activity”, and 18 variables, e.g. “mean(tBodyAccMagMean)”.