Code Book - Course Project

Getting and Cleaning Data

2023-01-19

Code Book

This document describes the variables, data, and any work performed in the run_analysis.R script to prepare the tidy data set.

1. Downloading Data

The raw dataset was downloaded in the local working directory in .zip format, with the download.file() function and unzipped to a folder called "UCI HAR Dataset" with the unzip() function.

2. Loading data

The .txt files in the text and train folders of the UCI HAR Dataset were loaded in Rstudio environment with the read.table() functions. The files (loaded .txt files) were:

- features (features.txt) - a 561 x 2 table containing a list of all features (number of features: 561)
- labels (activity_labels.txt) - a 6 x 2 table that links the class labels (number of class labels: 6) with their activity name
- X_train (X_train.txt) - a 7352 x 561 table with the recorded feature training data
- X_test (X_test.txt) - a 2947 x 561 table with the recorded feature test data
- y_train (y_train.txt) - a 7352 x 2 table with the training data of the recorded activity names
- y_test (y_test.txt) - a 2947 x 2 table with the test data of the recorded activity names
- subject_train (subject_train.txt) - a 7352 x 1 table identifying the subject who performed the activity for each window sample in the training data set
- subject_test (subject_test.txt) - a 2947 x 1 table identifying the subject who performed the activity for each window sample in the test data set

3. Merge the training and test datasets to one dataset

X_test, subject_test and y_test were merged in one dataset Test_Data with the cbind() function. Test_Data is a 2947 x 563 table. Also X_train, subject_train and y_train were merged in one dataset Train_data with the cbind() function. Train data is a 7352 x 563 table.

Finally the final dataset 'data' was created with the rbind() function. data is a 10299 x 563 table.

4. Extract only the measurements on the mean and standard deviation for each measurement

sum_data - a 10299 x 88 table - is a subset of data containing the columns: Lab_ID, Subject_ID, and the measurements on the mean and standard deviation for each measurement/variable.

- Lab_ID is a column variable that is linked with the activity name
- Subject_ID is a column variable that identifies the subject who performed the activity for each window sample

5. Use descriptive activity names to name the activities in the data set

A new column (LabelName) is added to the sum_data table - making it a 10299 x 89 table - that links the Lab_ID (class label 1-6) with the activity name. The labels table is a 6 x 2 that matches the class label (column 1) with the activity name (column 2)

6. Appropriately label the data set with descriptive variable names

After printing the names of the sum_data with the names() function, the words that need to be labelled appropriately along with the revised labeling are:

- Acc - Accelerometer
- Gyro - Gyroscope
- mag - Magnitude
- angle - Angle
- gravity - Gravity
- BodyBody - Body
- t - Time (whenever 't' is written in the beginning of the variable name)
- Freq - Frequency
- f - Frequency (whenever 'f' is written in the beginning of the variable name)
- mean - Mean
- std - St_Dev

7. Create a second, independent tidy data set with the average of each variable for each activity and each subject

final_data - a 180 x 89 table - containing the average of each variable from the sum_data table for each activity (LabelName) and each subject (Subject_ID) The final_data is saved as a .txt file as final_data.txt in the working directoty and uploaded in the GitHub repository.