Getting and Cleaning Data

Johns Hopkins via Coursera

Jeff Leek, PhD, Roger D. Peng, PhD, Brian Caffo, PhD

Submitted by: Mel Haynes Sr

Date submitted: 5/25/2014

**Study Design**

The experiments have been carried out with a group of 30 volunteers within an age bracket of 19-48 years. Each person performed six activities (WALKING, WALKING\_UPSTAIRS, WALKING\_DOWNSTAIRS, SITTING, STANDING, LAYING) wearing a smartphone (Samsung Galaxy S II) on the waist. Using its embedded accelerometer and gyroscope, we captured 3-axial linear acceleration and 3-axial angular velocity at a constant rate of 50Hz. The experiments have been video-recorded to label the data manually. The obtained dataset has been randomly partitioned into two sets, where 70% of the volunteers was selected for generating the training data and 30% the test data.   
  
The sensor signals (accelerometer and gyroscope) were pre-processed by applying noise filters and then sampled in fixed-width sliding windows of 2.56 sec and 50% overlap (128 readings/window). The sensor acceleration signal, which has gravitational and body motion components, was separated using a Butterworth low-pass filter into body acceleration and gravity. The gravitational force is assumed to have only low frequency components, therefore a filter with 0.3 Hz cutoff frequency was used. From each window, a vector of features was obtained by calculating variables from the time and frequency domain. 

**Raw Data Information:**

**The raw data for the exercise was downloaded from:** <https://d396qusza40orc.cloudfront.net/getdata%2Fprojectfiles%2FUCI%20HAR%20Dataset.zip>

**A full description of the raw data can be found at:** <http://archive.ics.uci.edu/ml/datasets/Human+Activity+Recognition+Using+Smartphones>

For each record in the dataset it is provided:   
- Triaxial acceleration from the accelerometer (total acceleration) and the estimated body acceleration.   
- Triaxial Angular velocity from the gyroscope.   
- A 561-feature vector with time and frequency domain variables.   
- Its activity label.   
- An identifier of the subject who carried out the experiment.

**The purpose of the project is described as:**

The purpose of this project is to demonstrate your ability to collect, work with, and clean a data set. The goal is to prepare tidy data that can be used for later analysis. You will be graded by your peers on a series of yes/no questions related to the project. You will be required to submit:

* A tidy data set
* A link to a Github repository with your script for performing the analysis
* A code book that describes the variables, the data, and any transformations or work that you performed to clean up the data called CodeBook.md.
* A README.md in the repo with your scripts. This repo explains how all of the scripts work and how they are connected.

**Code Book**

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable Name** | **Signal Type** | **Data Type** | **Value Type** |
| tbodyaccjerkmeanx | Time | Numeric | Mean Value |
| tbodyaccjerkmeany | Time | Numeric | Mean Value |
| tbodyaccjerkmeanz | Time | Numeric | Mean Value |
| tbodyaccjerkstdx | Time | Numeric | Standard Deviation |
| tbodyaccjerkstdy | Time | Numeric | Standard Deviation |
| tbodyaccjerkstdz | Time | Numeric | Standard Deviation |
| tbodyaccjerkmagmean | Time | Numeric | Mean Value |
| tbodyaccjerkmagstd | Time | Numeric | Standard Deviation |
| fbodyaccjerkmeanx | FFT | Numeric | Mean Value |
| fbodyaccjerkmeany | FFT | Numeric | Mean Value |
| fbodyaccjerkmeanz | FFT | Numeric | Mean Value |
| fbodyaccjerkstdx | FFT | Numeric | Standard Deviation |
| fbodyaccjerkstdy | FFT | Numeric | Standard Deviation |
| fbodyaccjerkstdz | FFT | Numeric | Standard Deviation |
| fbodyaccjerkmeanfreqx | FFT | Numeric | Mean Value |
| fbodyaccjerkmeanfreqy | FFT | Numeric | Mean Value |
| fbodyaccjerkmeanfreqz | FFT | Numeric | Mean Value |
| fbodybodyaccjerkmagmean | FFT | Numeric | Mean Value |
| fbodybodyaccjerkmagstd | FFT | Numeric | Standard Deviation |
| fbodybodyaccjerkmagmeanfreq | FFT | Numeric | Mean Value |
| angletbodyaccjerkmeangravitymean | Angle Between 2 vectors | Numeric | Mean Value |

**Process**

The scripts were written using RStudio version 0.98.501 and R version 3.03 on Windows 7 (64 Bit).

* The following data were loaded into R using read.table with header = FALSE
  + Activity\_labels.txt
  + Features.txt
  + Subject.txt
* The following test data were loaded into R using read.tables with stringsAsFactors = FALSE and header = FALSE
  + Subject\_test.txt
  + X\_test.txt
  + y\_test.txt
* The following training data were loaded into R using read.tables with stringsAsFactors = FALSE and header = FALSE
  + Subject\_train.txt
  + X\_train.txt
  + y\_train.txt
* The test and training subject data frames were combined using rbind
* The test and training activity data frames were combined using rbind
* The test and training data frames were combined
* The supplied features were cleansed by
  + Changing to all lower case
  + Removing all punctuation and non alphabetic characters from the names
* The cleansed names were added to the data frame
* Using grep, the data frame was filtered to select only Mean and std values per the project instructions.
* The data frame was further filtered to include only jerk and accelerate values to support the design.
* The activity and activitylabels were combined using sqldf to create a simple data frame of activity and activity label.
* The activity data frame was used to add subject and activity columns to the project data frame using cbind and the column names changed to subject and activity.
* The data frame was aggregated using the mean of each value on subject and activity using aggregate and the sorted bu subject using order.
* The resultant data frame was aggregated by Subject and Activity with the mean value of each variable.
* The data fram was written to jerkyAcceleration.txt in the project database
* The following values were written to github
  + Run\_analysis.R
  + README.md
  + CodeBook.md
  + jerkyAcceleration.txt