README.md

Brett Taylor

February 22, 2015

# Getting and Cleaning Data Course Project

This course project is focused on collecting, work with and cleaning data, and ultimately creating a tidy data set. The data source is based on a study of "Human Activity Recognition Using Smartphones Data Set". The following description is found here: [UCI Machine Learning Repository](http://archive.ics.uci.edu/ml/datasets/Human+Activity+Recognition+Using+Smartphones).  
*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.*

## Project Objectives

1. Merge the training and the test sets to create one data set.
2. Extract only the measurements on the mean and standard deviation for each measurement.
3. Use descriptive activity names to name the activities in the data set
4. Appropriately label the data set with descriptive variable names.
5. From the data set in step 4, create a second, independent tidy data set with the *average* of each variable for each activity and each subject.

Note that the actual process I used to does not follow the exact sequence shown above, and it does end up with the outcome that is shown in step 5. There are several scripts have been used to execute this project.

|  |  |  |
| --- | --- | --- |
| Script | Output | Purpose |
| downloadData.R | ./data/projdata.zip | Download required input data to ensure they are ready. |
| **run\_analysis.R** | activity\_summary.txt | Process the input data, and create a summary file based on the mean of the selected measurements. |
| README.Rmd | README.md | R markdown file that generates the README.md markdown file. It describes the processes associated with this project. This document is the output file. |
| CodeBook.Rmd | CodeBook.md | R markdown file that generates the code book, CodeBook.md, that describes the data sets, and how they were processed. The document includes embeded R code that creates the complex description of data in an automated fashion. |

**Script:** "run\_analysis.R"

This is the main script that must be run to create the tidy data set. The purpose of this script is to retrieve the data from the UCI-HAR dataset, and to then process it to create a summary file that is based on the average of each variable grouped by each activity and subject.

*Processing Overview* The script was developed to properly load the data, and process it in a simplified method.

1. **Clean Feature names**

* +Output: character vector feature.names

1. **Read Source Data**

* +Output: data.frames X\_test,X\_train,y\_test,y\_train,subject.test,subject.train

1. **Combine & Merge data frames**

* +Output: data.frame all.data

1. **Label all.data**

* +Output: data.frame all.data.labeled

1. **Select the correct variables for all.data.labeled**

* +Output: data.frame all.data.labeled

1. **Summarize all.data.labeled**

* +Output: data.frame tidy.data

1. **Write tidy.data to file**

* +Output: file "activity\_summary.txt"

*Script requirements*

Packages that must be installed:

install.packages("dplyr")

Data that must be installed:  
The UCI-HAR data set is required to process run this script. Download the data to an existing directory named data. If the directory does not exist it will be created.

# downloadData.R  
## Create the ./data directory if it does not exist.  
if(!file.exists("/data")) {  
 dir.create("data")   
}  
## Download the data from the Internet if it does not exist.  
if(!file.exists("./data/projdata.zip")) {  
 data.source.url<-"https://d396qusza40orc.cloudfront.net/getdata%2Fprojectfiles%2FUCI%20HAR%20Dataset.zip"  
 download.file(url=data.source.url,destfile="./data/projdata.zip",method="curl")  
 download.date<-date()  
##Manually unzip the file so that it can be processed.   
}

## Additonal Informaton about how the data was created

Information about the column names from the data set X\_test.txt and X\_train.txt created in the study.  
[features\_info.txt](data/UCI-HAR/features_info.txt)

Detailed documentation on the study [A Public Domain Dataset for Human Activity Recognition Using Smartphones](https://www.elen.ucl.ac.be/Proceedings/esann/esannpdf/es2013-84.pdf)

|  |
| --- |
| Human Activity Recognition Using Smartphones Dataset Version 1.0 |
| Jorge L. Reyes-Ortiz, Davide Anguita, Alessandro Ghio, Luca Oneto. Smartlab - Non Linear Complex Systems Laboratory DITEN - Universit? degli Studi di Genova. Via Opera Pia 11A, I-16145, Genoa, Italy. [activityrecognition@smartlab.ws](mailto:activityrecognition@smartlab.ws) (<http://www.smartlab.ws>) |