**Code Book**

This document describes the code inside run\_analysis.R.

**Purpose:** The code processes a number of individual but related data files related to six different physical activities of 30 volunteers captured by Samsung Galaxy S II smartphone. The purpose of the code is to read and organize the data elements stored in several files into one tidy data set (and one summarized data set) which are easy to read, interpret and process by further analysis on certain characteristics of the experiment.

**Study Design**

The details of the input data can be found at: <http://archive.ics.uci.edu/ml/datasets/Human+Activity+Recognition+Using+Smartphones>

The information is basically about several measurements (561 variables) sensed and recorded by the smartphone for each individual (30 instances) and each type of activity (6 different activities).

The code will filter the 561 measurements to 66 variables (related to only mean and standard deviation values of certain variables) and provides a more readable and compiled version of the data.

**Input:** 8 individual data sets as described in the link above and summarized below:

Two files (features.txt and activity\_labels.txt) are storing description of the coded variables in other six files. So they can be considered as “Reference” data.

Six files are grouped into two groups: Three files for Train data and three files for Test data. The structure for each corresponding file for both train and test files are identical. However the contents are not. Train data sets are related to the data captured as “train” data set for 21 individuals and Test data sets are related to the data captured as “test” data set for 9 other individuals.

**Output:** There are two data frames created by the code and two equivalent .csv files stored in the working directory

* final\_train\_test: Subject and activity level data for both train and test data
* final\_train\_test\_summary: Aggregated data set with average of all measurement

**A summary about the information:**

Table below displays the frequency of combination of Type (train/test), subject (individual code) and activity (six different activities) for the resulting data set. Freq = 0 is filtered.

One interesting observation is that for various subjects only one activity is captured. And only a few number of subjects have 2 activities captured (for example subject 6 has two activities as “WALKING” and “WALKING\_UPSTAIRS”.

Note: This table can be created by the code below running on the output data set created:

x = as.data.frame(table(final\_train\_test$type,final\_train\_test$subject,final\_train\_test$activity\_desc))

colnames(x) = c("Type","subject","activity","Freq")

y = x[x$Freq > 0,]

write.csv(y,"summary.csv")

|  |  |  |  |
| --- | --- | --- | --- |
| Type | subject | activity | Freq |
| Train | 1 | WALKING | 347 |
| Train | 3 | WALKING | 341 |
| Train | 5 | WALKING | 302 |
| Train | 6 | WALKING | 236 |
| Train | 6 | WALKING\_UPSTAIRS | 89 |
| Train | 7 | WALKING\_UPSTAIRS | 308 |
| Train | 8 | WALKING\_UPSTAIRS | 281 |
| Train | 11 | WALKING\_UPSTAIRS | 316 |
| Train | 14 | WALKING\_DOWNSTAIRS | 244 |
| Train | 14 | WALKING\_UPSTAIRS | 79 |
| Train | 15 | WALKING\_DOWNSTAIRS | 328 |
| Train | 16 | WALKING\_DOWNSTAIRS | 366 |
| Train | 17 | SITTING | 320 |
| Train | 17 | WALKING\_DOWNSTAIRS | 48 |
| Train | 19 | SITTING | 360 |
| Train | 21 | SITTING | 408 |
| Train | 22 | SITTING | 198 |
| Train | 22 | STANDING | 123 |
| Train | 23 | STANDING | 372 |
| Train | 25 | STANDING | 409 |
| Train | 26 | STANDING | 392 |
| Train | 27 | LAYING | 298 |
| Train | 27 | STANDING | 78 |
| Train | 28 | LAYING | 382 |
| Train | 29 | LAYING | 344 |
| Train | 30 | LAYING | 383 |
| Test | 2 | WALKING | 302 |
| Test | 4 | WALKING | 194 |
| Test | 4 | WALKING\_UPSTAIRS | 123 |
| Test | 9 | WALKING\_UPSTAIRS | 288 |
| Test | 10 | WALKING\_DOWNSTAIRS | 234 |
| Test | 10 | WALKING\_UPSTAIRS | 60 |
| Test | 12 | SITTING | 134 |
| Test | 12 | WALKING\_DOWNSTAIRS | 186 |
| Test | 13 | SITTING | 327 |
| Test | 18 | SITTING | 30 |
| Test | 18 | STANDING | 334 |
| Test | 20 | LAYING | 156 |
| Test | 20 | STANDING | 198 |
| Test | 24 | LAYING | 381 |

**Code profile:**

Developed in R Studio 0.98

Platform (using output of version function):

i386-w64-mingw32

i386

mingw32

i386, mingw32

3

1.3

2015

03

09

67962

R

R version 3.1.3 (2015-03-09)

Smooth Sidewalk

**Process:**

**High Level:**

# Step 1 - Read the required Data Sets

# Step 2 - Create most granular data set at subject/activity level

# Step 2.1 - Create a tidy data set for train data

# Step 2.2 - Create a tidy data set for test data

# Step 2.3 - Merge the train and test data sets into one data set

# Step 3 - Create summary average data set at subject/activity level

**Detailed Level:**

#############

# Step1 - Read the required Data Sets

# Check if the source zip file exists. If not download the file

# Check if the zip file is unzipped already. If not, unzip the file

# Note: The directory name in the zipped file is retained.

# So this phase checks if the directory exists and does not check

# the content of the directory.

# For Train records

# - subject\_train from subject\_train.txt

# - x\_train from train/X\_train.txt

# - y\_train from train/y\_train.txt

# For Test records

# - subject\_test from subject\_test.txt

# - x\_test from test/X\_test.txt

# - y\_test from test/y\_test.txt

# For reference records

# - features from features.txt

# - activity from activity\_labels.txt

#############

Note: The script writes the two data frames into .csv files at the working directory:

* **final\_train\_test.csv**
* **final\_train\_test\_summ.csv**

**Data Dictionary of the output data sets:**

* **final\_train\_test**
  + Two groups of columns are produced:
    - Factor variables indicating the observation categorical info:
      * type: Indicator of “Train” or “Test” as per source of the observation. Possible values: Train, Test
      * subject: the “id” of the subject (volunteer individual participating in the experiment) – possible values are 1 to 30 indicating 30 distinct participants. The value is directly copied from the source
      * activity\_id: the “id” of the activity (activity type code) – possible values are 1 to 6 indicating any of six possible activities captured by the device. The value is directly copied from the source
      * activity\_desc: descriptive value of the coded activity (corresponding to activity\_id). Possible values are description of 1 to 6 activities such as “WALKING”, “LAYING”, etc.
      * 66 variables as measures associated with each observation. The values are directly moved from source. Values are real numbers with the range of [-1,1]. These 66 variables are categorized into “mean” and “standard deviation” pairs. Therefore for each 33 measurements there is a pair of values one as mean value and one as standard deviationof the measurement for a particular aspect of an activity.
* **final\_train\_test\_summ**
  + This data set is a summary of **final\_train\_test** grouping the data into “subject” and the “activity” reporting the mean of all 66 measuring variables for each combination of groups (e.g. subject 2 and activity WALKING, subject 17 and activity SITTING, etc.). Therefore the data set has the same number of numeric measurement columns but the values are now the mean of the group.
  + The data set has 40 records indicating 40 distinct groups of “subject”, “activity” pair.