

## Code Book for run\_analysis.R

## Introduction

This code book is for the run\_analysis.R script developed for the Data Sharing course project. The goal of the project is to write a script that takes unstructured data and create a tidy data set. According to the paper "Tidy Data" written by Hadley Wickham (J. Stat. Software, 59(10), 2014), "Tidy datasets provide a standardized way to link the structure of a dataset (its physical layout) with its semantics (its meaning)." To accomplish this, Wickham further says that "A dataset is a collection of values... Values are organized in two ways. Every value belongs to a variable and an observation." To keep this standard, a data table that contains tidy data has three attributes (Wickham):

1. Each variable forms a column
2. Each observation forms a row
3. Each type of observational unit forms a table

It is following these principles that the structure of the output file that results from running the run\_analysis.R script is derived.

## Input files:

The data was provided free of charge and as is by:

Davide Anguita, Alessandro Ghio, Luca Oneto, Xavier Parra and Jorge L. Reyes-Ortiz. Human Activity Recognition on Smartphones using a Multiclass Hardware-Friendly Support Vector Machine. International Workshop of Ambient Assisted Living (IWAAL 2012). Victoria-Gasteiz, Spain. Dec 2012

To understand the operation of the script, we must first understand the structure of the input files. The run\_analysis script reads in 8 separate text files from the data source. The input files and a brief description follow:

- features.txt: List of all features.
- activity\_labels.txt: Establishes a numerical value to the activity name.
- X\_train.txt: Training set.
- y\_train.txt: Training labels.
- X\_test.txt: Test set.
- y\_test.txt: Test labels.
- subject\_test.txt: A list of the subjects (identified by a unique number)
- subject\_train.txt: A list of the subjects (identified by a unique number)

The input data is read into data tables with the following names

## features:

This is a table of 561 observation types and 2 variables. From the features\_info.txt file included with the data. The data is a collection of output from the accelerometer (ACC) and gyroscope (Gyro) features of a Samsung Galaxy II smart phone. From this data, other variables are calculated from the linear acceleration and angular velocity. The leading "t" indicates the data is collected in the time domain. The leading "f" indicates that the data has been subjected to a Fast Fourier Transform, and mathematical operation that transforms data from the time domain to the frequency domain.

## run\_analysis\_code\_book

Variables with "Acc" in the name have standard gravity units (g). Variables with "Gyro" have units radians/second.

tBodyAcc-XYZ  
tGravityAcc-XYZ :  
tBodyAccJerk-XYZ :  
tBodyGyro-XYZ :  
tBodyGyroJerk-XYZ :  
tBodyAccMag  
tGravityAccMag  
tBodyAccJerkMag  
tBodyGyroMag  
tBodyGyroJerkMag  
fBodyAcc-XYZ  
fBodyAccJerk-XYZ  
fBodyGyro-XYZ  
fBodyAccMag  
fBodyAccJerkMag  
fBodyGyroMag  
fBodyGyroJerkMag

the data labeled `"*BodyACC"` and `"*BodyGyro"` is the raw data from the accelerometer and the gyroscope. The other data is derived from these quantities. For each of the above variables, the following quantities were determined and included in the input dataset

mean(): Mean value  
std(): Standard deviation  
mad(): Median absolute deviation  
max(): Largest value in array  
min(): Smallest value in array  
sma(): Signal magnitude area  
energy(): Energy measure. Sum of the squares divided by the number of values.  
iqr(): Interquartile range  
entropy(): Signal entropy  
arCoeff(): Autoregression coefficients with Burg order equal to 4  
correlation(): correlation coefficient between two signals  
maxInds(): index of the frequency component with largest magnitude  
meanFreq(): Weighted average of the frequency components to obtain a mean frequency  
skewness(): skewness of the frequency domain signal  
kurtosis(): kurtosis of the frequency domain signal  
bandsEnergy(): Energy of a frequency interval within the 64 bins of the FFT of each window.  
angle(): Angle between two vectors.

Additional vectors obtained by averaging the signals in a signal window sample. These are used on the angle() variable:

gravityMean  
tBodyAccMean  
tBodyAccJerkMean  
tBodyGyroMean  
tBodyGyroJerkMean

values in this table are normalized

activity\_label:

This table associates an integer with each of the types of activities that the subjects conducted in during the measurements. These are:

1 WALKING

## run\_analysis\_code\_book

2 WALKING\_UPSTAIRS  
3 WALKING\_DOWNSTAIRS  
4 SITTING  
5 STANDING  
6 LAYING

X\_train:

This is a table of 7352 observations of 561 variables. The 561 variables correspond to the 561 observation types in the features table, and the 7352 observations correspond to multiple measurements of the features conducted on 21 subjects performing the 6 activities. Each row corresponds to a single activity performed by one subject. This table has no column or row labels. The subjects in the train group are unique from those of the test group and were adults ages 19-48 randomly selected.

X\_test:

This is a table of 2947 observations of 561 variables. The 561 variables correspond to the 561 observation types in the features table, and the 2947 observations correspond to multiple measurements of the features conducted on 9 subjects performing the 6 activities. Each row corresponds to a single activity performed by one subject. This table has no column or row labels. The subjects in the test group are unique from those of the train group and were adults ages 19-48 randomly selected.

y\_train:

This is a dataset of 7352 observations of 1 variable. The 7352 observations correspond to the activity type (e.g. 1 == WALKING) that the subject was participating in when the observations in the rows of the X\_train table were being collected. Values in this table are limited to 1 - 6.

y\_test:

This is a dataset of 2947 observations of 1 variable. The 2947 observations correspond to the activity type (e.g. 1 == WALKING) that the subject was participating in when the observations in the rows of the X\_test table were being collected. Values in this table are limited to 1 - 6.

subject\_train:

This is a dataset of 7352 observations of 1 variable. The 7352 observations correspond to the subject who was participating in the activity that lead to the observations in the rows of the X\_train table.

subject\_test:

This is a dataset of 2947 observations of 1 variable. The 2947 observations correspond to the subject who was participating in the activity that lead to the observations in the rows of the X\_test table.

Input Summary:

Based on the size of the input files, the script is able to align the activity type and the unique subject with each of the 7352 and 2947 observations in the train and the test groups. This accumulation of observations with unique variable results in a table with 563 columns (e.g. variables) and 10299 rows (e.g. observations) - a total of 5,798,377 elements of the table.

Script operation summary:

The run\_analysis.R script reads the data into data tables. The script first adds a

#### run\_analysis\_code\_book

header to the X\_test and X\_train data tables to provide a unique name to each variable. The column names are from the features data table. It then adds a column to the X\_test and X\_train data tables that corresponds to the activity labels (from y\_test and y\_train respectively) to identify the type of activity the subject was performing while a particular observation was made. Similarly, another column (from subject\_test and subject\_train, respectively) is added that adds the subject labels to the X\_test and X\_train data table to identify the subject who performed the activity while the observation was being made.

The X\_test and X\_train data tables are then combined using the rbind() function to create a single data table (complete\_data) of all the observations of all the variables.

The data table is trimmed by selecting only the data for variables that have "mean" and "std" (mean and standard deviations). This trims the data table to 10299 observations of only 82 variables - a total of 844,518 elements in the table.

Finally, the script obtains an average value for each variable for each activity and each subject. There are 6 activities and 30 subjects. This results in a table (complete\_data\_average) that has 81 variables and 180 unique observations - for a total of 14,580 total elements in the table. This represents a data size that is 0.25% the size of the input file.

Output file, Samsung\_data\_tidy.txt

The output file, Samsung\_data\_tidy.txt is the output file that is written to disk by the script. This file should be read back into the R environment using the read.table("./Samsung\_data\_tidy.txt", header=TRUE). This operation will result in a data table with 180 observations of 81 variables. The variables are:

```
"1" "Subject"
"2" "Activity"
"3" "tBodyAccmeanX"
"4" "tBodyAccmeanY"
"5" "tBodyAccmeanZ"
"6" "tGravityAccmeanX"
"7" "tGravityAccmeanY"
"8" "tGravityAccmeanZ"
"9" "tBodyAccJerkmeanX"
"10" "tBodyAccJerkmeanY"
"11" "tBodyAccJerkmeanZ"
"12" "tBodyGyromeanX"
"13" "tBodyGyromeanY"
"14" "tBodyGyromeanZ"
"15" "tBodyGyroJerkmeanX"
"16" "tBodyGyroJerkmeanY"
"17" "tBodyGyroJerkmeanZ"
"18" "tBodyAccMagmean"
"19" "tGravityAccMagmean"
"20" "tBodyAccJerkMagmean"
"21" "tBodyGyroMagmean"
"22" "tBodyGyroJerkMagmean"
"23" "fBodyAccmeanX"
"24" "fBodyAccmeanY"
"25" "fBodyAccmeanZ"
"26" "fBodyAccmeanFreqX"
"27" "fBodyAccmeanFreqY"
"28" "fBodyAccmeanFreqZ"
"29" "fBodyAccJerkmeanX"
"30" "fBodyAccJerkmeanY"
"31" "fBodyAccJerkmeanZ"
"32" "fBodyAccJerkmeanFreqX"
"33" "fBodyAccJerkmeanFreqY"
```

```

"34" "fBodyAccJerkmeanFreqZ"
"35" "fBodyGyromeanX"
"36" "fBodyGyromeanY"
"37" "fBodyGyromeanZ"
"38" "fBodyGyromeanFreqX"
"39" "fBodyGyromeanFreqY"
"40" "fBodyGyromeanFreqZ"
"41" "fBodyAccMagmean"
"42" "fBodyAccMagmeanFreq"
"43" "fBodyBodyAccJerkMagmean"
"44" "fBodyBodyAccJerkMagmeanFreq"
"45" "fBodyBodyGyroMagmean"
"46" "fBodyBodyGyroMagmeanFreq"
"47" "fBodyBodyGyroJerkMagmean"
"48" "fBodyBodyGyroJerkMagmeanFreq"
"49" "tBodyAccstdX"
"50" "tBodyAccstdY"
"51" "tBodyAccstdZ"
"52" "tGravi tyAccstdX"
"53" "tGravi tyAccstdY"
"54" "tGravi tyAccstdZ"
"55" "tBodyAccJerkstdX"
"56" "tBodyAccJerkstdY"
"57" "tBodyAccJerkstdZ"
"58" "tBodyGyrostdX"
"59" "tBodyGyrostdY"
"60" "tBodyGyrostdZ"
"61" "tBodyGyroJerkstdX"
"62" "tBodyGyroJerkstdY"
"63" "tBodyGyroJerkstdZ"
"64" "tBodyAccMagstd"
"65" "tGravi tyAccMagstd"
"66" "tBodyAccJerkMagstd"
"67" "tBodyGyroMagstd"
"68" "tBodyGyroJerkMagstd"
"69" "fBodyAccstdX"
"70" "fBodyAccstdY"
"71" "fBodyAccstdZ"
"72" "fBodyAccJerkstdX"
"73" "fBodyAccJerkstdY"
"74" "fBodyAccJerkstdZ"
"75" "fBodyGyrostdX"
"76" "fBodyGyrostdY"
"77" "fBodyGyrostdZ"
"78" "fBodyAccMagstd"
"79" "fBodyBodyAccJerkMagstd"
"80" "fBodyBodyGyroMagstd"
"81" "fBodyBodyGyroJerkMagstd"

```

Those variables with Acc in the name have units of standard gravity (g), and those with Gyro in the name have units of radians/second.