Readme for John Hopkins and Cleaning Data course project

Karol Dorniak

Sunday, January 25, 2015

The objective of the project is to create a tidy.txt data set based on the raw data collected from the accelerometers from the Samsung Galaxy S smartphone.

To fulfil the project's requirements, the R script has been developed: run\_analysis.R, which purpose was to collect, transform and tidy raw data. The final results were saved in the tidy.txt file.

Raw data used by run\_analysis.R has been downloaded from <https://d396qusza40orc.cloudfront.net/getdata%2Fprojectfiles%2FUCI%20HAR%20Dataset.zip> and extracted to "UCI HAR Dataset" folder.

The following source files from the "UCI HAR Dataset" folder were used by run\_analysis.R to produce tidy.txt:

* UCI HAR Dataset/activity\_labels.txt - an activity's user friendly desciption of a given activity id
* UCI HAR Dataset/features.txt - list of all features (variables in X\_train.txt, X\_test.txt). The label of the column Nth of the X\_train.txt/X\_test.txt can be found in the row N of features.txt.
* UCI HAR Dataset/train/X\_train.txt - training data set
* UCI HAR Dataset/train/subject\_train.txt - each row identifies the subject who performed the activity for each window sample. The data in the Nth row of training data set relates to a subject identified in the Nth row of subject\_train.txt
* UCI HAR Dataset/train/y\_train.txt - data in the Nth row of the training data set relates to an activity id identified in the Nth row of the y\_train.txt. An activity's user friendly desciption of a given activity id can be found in the activity\_labels.txt
* UCI HAR Dataset/test/X\_test.txt - test data set
* UCI HAR Dataset/test/subject\_test.txt - each row identifies the subject who performed the activity for each window sample. The data in the Nth row of test data set relates to a subject identified in the Nth row of subject\_test.txt
* UCI HAR Dataset/test/y\_test.txt - data in the Nth row of the training data set relates to an activity id identified in the Nth row of the y\_test.txt. An activity's user friendly description of a given activity id can be found in the activity\_labels.txt

**run\_analysis.R script and "UCI HAR Dataset" folder are located in the working directory (same location).**

run\_analysis.R conducts the following operations on the raw data to produce tidy.txt file:

1. Merge the training and the test sets to create one data set and provide descriptive variable names:

A. load activity ids and labels from UCI HAR Dataset/activity\_labels.txt

B. load list of variables of each feature vector from UCI HAR Dataset/features.txt

C. load train data from UCI HAR Dataset/train/X\_train.txt

D. rename columns in C using data frame from B, we want to have user friendly varable name, example: V1 becomes "tBodyAcc-mean()-X"

E. load subject ids for train data from UCI HAR Dataset/train/subject\_train.txt

F. assign user friendly variable name to a column of data frame from point E, V1 becomes SubjectID

G. load train activity label id from UCI HAR Dataset/train/y\_train.txt

H. assign user friendly variable name to a column of data frame from point G, V1 becomes ActivityID

I. create final train data table by column binding data frames from F, H, D

J. load test data from UCI HAR Dataset/test/X\_test.txt

K. rename columns in J using data frame from B, we want to have user friendly varable name, example: V1 becomes "tBodyAcc-mean()-X"

L. load subject ids for test data from UCI HAR Dataset/test/subject\_test.txt

M. assign user friendly variable name to a column of data frame from point L, V1 becomes SubjectID

N. load test activity label id from UCI HAR Dataset/test/y\_test.txt

O. assign user friendly variable name to a column of data frame from point N, V1 becomes ActivityID

P. create final test data table by column binding data frames from M, O, K

R. merge data frames from I and P by row binding

1. Extract from the new merged data set only the measurements on the mean and standard deviation for each measurement.

* In this step we are reducing the number of columns in the merge data frame to only those that we are interested in: ActivityID,SubjectID and any column that have "std.." or "mean.." in its name

1. Uses descriptive activity names to name the activities in the merged data set
2. 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 (mean of all variables group by activity and subject)
3. Save independent tidy data to tidy.txt

**run\_analysis.R code has been explained in details using script commentaries - open the script file to read code explanation**