README

Sam Vennell

27 Aug 2016

Course Project: Getting and Cleaning Data, CourseRA course taught by John Hopkins University

README

Sam Vennell 27/08/2016

The method used to produce the tidy data set is outlined below:

All aspects of the test and train data, namely the X component of the data (eg "X_test.txt") which contains the variables and their values, the y component (eg "y_test.txt") which contains the activity numbers for these and the subject IDs (eg "subject_test.txt") are all loaded into R. These aspects of the testing and training data are combined into the "wide format" data frames "testdata" and "traindata" respectively, adding a column "datatype" which indicates whether the data is testing or training data.

These two data frames are subsequently combined into the dataframe "combdata". The measurements pertaining the means and standard deviations of the variables are then extracted into the "wide format" data frame "UCIHAR_meansd_wide" by using regex expressions. We also choose to include the "angle" means, eg "angle(Z,gravityMean)", as well as the mean and standard devation variables in the standard format (eg "fBodyGyro-std()-Y").

These are then "gathered" into a tidyer, "narrow" format data frame called "UCIHAR_meansd".

Finally, the average (mean) of each variable for each activity and subject is calculated using the {dplyr} package and stored in the data frame "UCIHAR_meansd_avg". Note that the training and test data are combined in this calculation.

Notes: * I have opted to have the final data frame in "narrow" format, and to keep the original variable names. I consider that this qualifies as