Readme

Ashok Natesan

June 5, 2016

Data Sets generated

This project is based on data compiled from the UCI HAR project, located at http://archive.ics.uci.edu/ml/datasets/Human+Activity+Recognition+Using+Smartphones. The exact algorithms used for the computation of the base metrics are extensively documented at the UCI website for your reference and these descriptions are not replicated here.

In this project, we have derived 2 tidy data sets from the base data

- A consolidated tidied data set (X_consolidated.txt) from the training and test data sets, that retains 79 (mean and std deviation) metrics of the 561 metrics compiled by the UCI UAR project.
- A second derivied tidy data set (meanBySubjectActivity.txt) shows observation data that is grouped by Subject, Activity and summarizes (using mean) the 79 metrics from the X_consolidated.txt.

Both these files are generated by running the Run_Analysis.R script in the directory, "HARtidyDS""

Code Artifacts

In this project, there are 3 R scripts

1. Run_Analysis.R: Driver script that generates the data sets. Can be invoked as Run_Analysis()
2. tidyHAR.R: Script that generates the X_consolidated.txt with many tidying steps documented in the code. For test purposes, tidyHAR() can be invoked in a stand-alone mode to generate the X_consolidate.txt by itself.
3. createHARAvgDataset.R: Scripts that derives the meanBySubjectActivity.txt data set, grouping the observation data by Subject and Activity

Tidying approach for X_consolidated.txt (tidyHAR.R)

The various tidying steps required in this script included

1. Consolidating the observation data (X_<type>.txt) with the companion Subject (subject_.txt) and ActivityCode (confusingly named y_.txt) files where <type> refers to both Training and Test datasets

2. Merging the above data set with Lookups data in terms of the Features (to get the column names for the data.frame and data.table), as well as the Activity descriptions (to add as a descriptive column in the data)

3. In the process, other smaller tidying subtasks included:

    a. Detection and de-duping (by suffixing unique ids) of duplicate entries in the feature set.

    b. Removal of metrics that were not mean or standard deviations - rearranging the column order for writing to the file, so that the UAR mettrics were at the tail end of the data set

    c. Generation of an index column which generates unique ids for each observation - Generation of a type field which tracks the association of each observation to Training or Test data set in X_consolidated.txt

The details of all of these steps are documented in reasonable detail in tidyHAR.R

This script uses data.frames primarily and the dplyr and plyr packages

Approach for meanBySubjectActivity.txt (createHARAvgDataset.R)

Since this phase starts with tidy data (X_consolidated.txt), the work in this script primarily has to do with calculating the mean of each metric grouped by subject, activity

This is done in a fairly straightfoward way using data.tables. This file is also extensively documented with comments.