Data-Cleaning Project Codebook

The input data in this project came in several files, mainly organized in two folders, called “train” and “test”. The data-cleaning steps required in this project are implemented in the R-script “run_analysis.R”. This script is organized in 5 main sections (numbered 1..5 in the script), and a preamble section that loads the input data from the files.

Section 1: combines the training and test datasets to construct a combined dataset

Section 2: extracts the columns corresponding to the “mean” and “std” stats in the combined dataset

Section 3: processes the numeric activity-labels corresponding to each row of the combined dataset, and generates a corresponding text-label for each row

Section 4: processes the given column names, to clean them up, mainly by removing non-alphanumeric characters

Section 5: generates a new file which contains the mean feature-values for each subject for each activity

Each section of the script includes explanatory comments.

Data-Format Column Names:

The combined dataset generated in Section 1, as well as the dataset generated in Section 5, both have 561 features corresponding the input features in the original training and test data. The feature-names have been edited in Section 4. The list of the new column-names is given in the file “result_features.txt”, one name per line (561 lines).

The data-frame generated in Section 5 is written out in the file: “ActivitySubjectMeans.txt”. Each column contains floating-point data, generated by the colMeans() function in R. The file contains 180 rows, as follows:

-Rows 1:30 - correspond to 30 users for activity 1 (Walking)

-Rows 31:60 - correspond to 30 users for activity 2 (Walking up)

-Rows 61:90 - correspond to 30 users for activity 3 (Walking down)

-Rows 91:120 - correspond to 30 users for activity 4 (Sitting)

-Rows 121:150 - correspond to 30 users for activity 5 (Standing)

-Rows 151:180 - correspond to 30 users for activity 6 (Laying)

The data for each row consists of space-separated columns.