## code book for assignment 3 getting and cleaning data course

Experimental Background

It involves human activity recognition using smartphones dataset.

Signal measurements from two triaxial sensors (accelerometer and gyroscope) were carried out with 30 subjects. The measurements were recorded for 6 types of activities: 1-WALKING, 2- WALKING\_UPSTAIRS, 3- WALKING\_DOWNSTAiRS, 4- SITTING, 5- STANDING, 6- LAYING. Experiments were video recorded and manually analysed.

The obtained dataset was partitioned into two sets- 70% of the subjects was used to generate the training data, the other 30% was used for test data. These measurements were manipulated to construct 561 features with time (t) and frequency (f) domain variables. The resulting data was normalized to [-1,1] range.

As a part of this assignment these features were manipulated in order to merge the partitioned datasets and to average the features representing raw data mean and standard deviation values

The dataset for this assignment was downloaded and read from:

http://archive.ics.uci.edu/ml/datasets/Human+Activity+Recognition+Using+Smartphones

Raw Data Characterization

The data is zipped to a folder and includes the following files:

/UCI HAR Dataset/

* Activity labels (as described previously)
* Features - list of 561 variables with unique nomenclature
* Features\_info.txt
* README.txt file

/train/

* /Inertial signals/- raw data for the triaxial measurements
* subject \_train.txt- ID numbers for the subjects that generated the train dataset
* X\_train.txt- normalized raw values that represent the features for 6 activities and 70% of subjects constructing the training dataset
* Y\_train.txt- activity ID variable for the training dataset

/test/

* /inertial signals/
* Subject\_test.txt
* X\_test
* y\_test

List of final variables

The final and merged data is composed of averages for 66 selected variables that represent mean and standard deviation values of the raw data. In addition to these averaged variables, the first two variables: ‘subject\_ID’ and ‘activity\_name’ are identifiers for every row in the dataset.

A list of these variables composing the final tidy data are as followed:

|  |  |  |  |
| --- | --- | --- | --- |
| 1] | subject\_id | activity\_name | average-tBodyAcc-mean-X |
| [4] | average-tBodyAcc-mean-Y | average-tBodyAcc-mean-Z | average-tBodyAcc-std-X |
| [7] | average-tBodyAcc-std-Y | average-tBodyAcc-std-Z | average-tGravityAcc-mean-X |
| [10] | average-tGravityAcc-mean-Y | average-tGravityAcc-mean-Z | average-tGravityAcc-std-X |
| [13] | average-tGravityAcc-std-Y | average-tGravityAcc-std-Z | average-tBodyAccJerk-mean-X |
| [16] | average-tBodyAccJerk-mean-Y | average-tBodyAccJerk-mean-Z | average-tBodyAccJerk-std-X |
| [19] | average-tBodyAccJerk-std-Y | average-tBodyAccJerk-std-Z | average-tBodyGyro-mean-X |
| [22] | average-tBodyGyro-mean-Y | average-tBodyGyro-mean-Z | average-tBodyGyro-std-X |
| [25] | average-tBodyGyro-std-Y | average-tBodyGyro-std-Z | average-tBodyGyroJerk-mean-X |
| [28] | average-tBodyGyroJerk-mean-Y | average-tBodyGyroJerk-mean-Z | average-tBodyGyroJerk-std-X |
| [31] | average-tBodyGyroJerk-std-Y | average-tBodyGyroJerk-std-Z | average-tBodyAccMag-mean |
| [34] | average-tBodyAccMag-std | average-tGravityAccMag-mean | average-tGravityAccMag-std |
| [37] | average-tBodyAccJerkMag-mean | average-tBodyAccJerkMag-std | average-tBodyGyroMag-mean |
| [40] | average-tBodyGyroMag-std | average-tBodyGyroJerkMag-mean | average-tBodyGyroJerkMag-std |
| [43] | average-fBodyAcc-mean-X | average-fBodyAcc-mean-Y | average-fBodyAcc-mean-Z |
| [46] | average-fBodyAcc-std-X | average-fBodyAcc-std-Y | average-fBodyAcc-std-Z |
| [49] | average-fBodyAccJerk-mean-X | average-fBodyAccJerk-mean-Y | average-fBodyAccJerk-mean-Z |
| [52] | average-fBodyAccJerk-std-X | average-fBodyAccJerk-std-Y | average-fBodyAccJerk-std-Z |
| [55] | average-fBodyGyro-mean-X | average-fBodyGyro-mean-Y | average-fBodyGyro-mean-Z |
| [58] | average-fBodyGyro-std-X | average-fBodyGyro-std-Y | average-fBodyGyro-std-Z |
| [61] | average-fBodyAccMag-mean | average-fBodyAccMag-std | average-fBodyAccJerkMag-mean |
| [64] | average-fBodyAccJerkMag-std | average-fBodyGyroMag-mean | average-fBodyGyroMag-std |
| [67] | average-fBodyGyroJerkMag-mean | average-fBodyGyroJerkMag-std |  |

Names interpretation:

* Acc- acceleration signal
* Gyr- gyroscope signal
* t- time domain
* f- frequency domain (time transformed by FFT)
* Body- body acceleration/gyroscope signal
* Gravity- gravity acceleration/gyroscope signal
* Mag- magnitude calculation for the 3D signals
* Jerk- Jerk signals

Study Design

The script “run\_Analysis.R” was established in order to get the final tidy data. The process by which this script was undertaken was as follows:

The script uses the packages ‘dplyr’ and ‘data.table’

* Merge the training and the test sets to create one data set

Download and unzip files to the working repository. Store data files as data table.

Join the following files: x\_test, x\_train, subject\_test, subject\_train, y\_test and y\_train and construct one complete dataset that binds train and test data with the subjects that produced it.

* Extract only the measurements on the mean and standard deviation for each measurement

Find the target features by inspecting the variables name, extract only features that represent mean or standard deviation values as well as the subject\_ID and activity\_ID identifiers and store them in a separate data table.

* Use descriptive activity names to name the activities in the data set

Replace the ID activity identifier with name identifier by joining the ‘activity\_labels’ table that contains the descriptive activity names with the merged table created in the previous step.

* Appropriately label the data set with descriptive variable names.

Extract a vector with the merged table label names, remove a parenthesis symbol ‘()’ from the labels in order to create more descriptive and clear labels, fix a typo that exists in some labels- “BodyBody” was renamed as “Body” and set this vector as a new label vector for the merged data table.

* create a second, independent tidy data set with the average of each variable for each activity and each subject.

Group the resulting data set by subject\_ID and activity\_name, summarize average values for all rows by subject and activity, name averaged variables by adding the phrase ‘average-’ to their name and set the new names as labels of a new summarized data table- ‘final\_tidy\_data’.

Finally, write the data into a text file and store it in the working directory