Code Book

Disovankiri Boung

6/17/2020

The *run_analysis.R* script performs the data cleaning and maangement

- First, the code check if the data is ready in user's drive, if not, download from the website and unzip it.
- Second, data is read into dataframes
  - **features** <- features.txt : 561 rows, 2 columns The features selected for this database come from the accelerometer and gyroscope 3-axial raw signals tAcc-XYZ and tGyro-XYZ.
  - **activities** <- activity_labels.txt : 6 rows, 2 columns List of activities performed when the corresponding measurements were taken and its codes (labels)
  - **subject_test** <- test/subject_test.txt : 2947 rows, 1 column contains test data of 9/30 volunteer test subjects being observed
  - **x_test** <- test/X_test.txt : 2947 rows, 561 columns contains recorded features test data
  - **y_test** <- test/y_test.txt : 2947 rows, 1 columns contains test data of activities'code labels
  - **subject_train** <- test/subject_train.txt : 7352 rows, 1 column contains train data of 21/30 volunteer subjects being observed
  - **x_train** <- test/X_train.txt : 7352 rows, 561 columns contains recorded features train data
  - **y_train** <- test/y_train.txt : 7352 rows, 1 columns contains train data of activities'code labels
- **Step 1**, **dfX, dfY and subject** are created by merging the train and test dataset using *rbind()*. **new_df** merges dfX, dfY and subject together.
- **Step 2**, data on mean and standard deviation are extracted by **new_df** by using ***select()*** with its special ***contains()** argument. The tidy dateset is stored in **new_tidy_df**
- **Step 3**, the code in column 2 is labelled with descriptive names by comparing the code to the assigned activity in **activities data.frame**
- **Step 4**, labels on new_df is labelled in a more readable manne using names(), and grep() functions
  - Column 2 is renamed to activity
  - All Acc in column's name replaced by Accelerometer
  - All Gyro in column's name replaced by Gyroscope
  - All BodyBody in column's name replaced by Body
  - All Mag in column's name replaced by Magnitude
  - All start with character f in column's name replaced by Frequency
  - All start with character t in column's name replaced by Time
- **Step 5**, from the data set in step 4, **second_tidy_df** is created by taking the average (mean) of each variables by the subject and activity using group_by() and summarise_all(). The final dataset is written into Data.txt and Data.csv.