Cleaning Accelerometer and Gyroscope Data

This repo contains my project work for the "Getting and Cleaning Data" course, which is the 3rd course of the Johns Hopkins University Data Science Specialization on Coursera.

The purpose of this project is to demonstrate the ability to collect, work with, and clean a data set. The goal is to prepare Tidy Data that can be used for later analysis.

Tidy Data

More info on the concept of Tidy Data can be found in this Wikipedia article.

An in-depth treatise by Hadley Wickham about Tidy Data is presented in this paper.

Contents of repo

This repo contains:

- this **README.md** file;
- a **run_analysis.R** sript;
- the **tidy_data.txt** output file;
- a **Codebook.md**, corresponding to the **tidy_data.txt** file;

Description of project

One of the most exciting areas in all of data science right now is wearable computing. Companies like Fitbit, Nike, and Jawbone Up are racing to develop the most advanced algorithms to attract new users. The data linked to from the course website represent data collected from the accelerometers from the Samsung Galaxy S smartphone.

A full description is available at the website where the data was obtained.

How the run_analysis.R script works

Packages used by the script

Before running the script, make sure that packages **dplyr**, **stringr** and **rebus** are installed on your system.

Step 1. Download and unzip

The script downloads a **HAR.zip** file and unzips it to a **UCI HAR Dataset** folder.

The UCI HAR Dataset folder

The unzipped **UCI HAR Dataset** folder contains the following files:

- activities (dim: 6 by 2): 6 types of activity;
- features.txt (dim: 561 by 2): 561 measurement features;
- test/subject_test.txt (dim: 2947 by 1): subjectIDs;
- test/y_test.txt (dim: 2947 by 1): activityIDs;
- test/X_test.txt (dim: 2947 by 561): measurement features;

- train/subject_train.txt (dim: 7352 by 1): subjectIDs;
- train/y_train.txt (dim: 7352 by 1): activityIDs;
- train/X_train.txt (dim: 7352 by 561): measurement features;

*Note that the folder contains two **Inertial Signals** folders that can be safely ignored for the purpose of this script.*

Step 2. Read activities and features and do some cleanup

- transform activity values, e.g. from **"WALKING_UPSTAIRS"** to **"Walking upstairs"**;
- transform feature values, e.g. from **"fBodyBodyGyroJerkMag-std()"** to **"FreqBodyGyroJerkMagStd"**;
- only keep the 66 features corresponding to **mean** or **std** measurements.

Step 3. Read test and train files and merge them

The 6 test and train files are imported. In the **X_test** and **X_train** data frames, only the columns corresponding to **mean** and **std** measurements are kept.

First, the test and train data files are merged to respectively **test_data** (dim: 2946 by 68) and **train_test** (dim: 7352 by 68) data frames. Then both are combined to a data frame called **data** (dim: 10299 by 69), where the variable names are copied from the values of the **features** data set.

Lastly, the activity labels are merged into the **data** data frame.

Step 4. Make a tidy data set

To create a tidy data set, the data is first aggregated by subject and type of activity. Then means are calculated and arranged by subject and activity.

In the tidy dataset, called **tidy_data.txt**, the column names corresponding to the 66 measurement features are transformed to have **"mean"** at the front of the column name, e.g. from **"TimeBodyAccMeanX"** to **"meanTimeBodyAccMeanX"**

At the end of the script, the **tidy_data.txt** file is written to file.

Tidyness of the constructed dataset

The dataset is tidy, because:

- each observation (combination of **subjectID** and **activityType**) is in one row. There are 180 rows: 6 types of activity for each of 30 subjects;
- each column has exactly one unique type of measurement feature.