Programming Assignment 3

Contents

1 Introduction

- Scope: Project Work
- Course: Getting and Cleaning Data, provided by coursera.org
- Author: Giulio Giovannetti
- Date: 25/10/2020
- Main: "run_analysis.r"

2 Project requirements

The purpose of this project is to demonstrate your ability to collect, work with, and clean a data set. The goal is to prepare tidy data that can be used for later analysis. One of the most exciting areas in all of data science right now is wearable computing - see for example this article . 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 site where the data was obtained:

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

Here are the data for the project:

https://d396qusza40orc.cloudfront.net/getdata%2Fprojectfiles%2FUCI%20HAR%20Dataset.zip

You should create one R script called run_analysis.R that does the following.

1. Merges the training and the test sets to create one data set.
2. Extracts only the measurements on the mean and standard deviation for each measurement.
3. Uses descriptive activity names to name the activities in the data set
4. Appropriately labels the data set with descriptive variable names.
5. From the data set in step 4, creates a second, independent tidy data set with the average of each variable for each activity and each subject.

3   Project work

3.1   Input

- Download the data set here;
- put the data set in the same directory of the main script "run_analysis.r";
- unzip the data set.

3.2   Ouput

- The main script "run_analysis.r" produces the file "tidy_data_set.txt" formattes as required.

3.3   Code description

3.3.1   Load libraries

The 'dplyr' is required to effectively manage data set.

```
library(dplyr)
```

3.3.2   Define input folders

The whole UCI HAR data set is provided in several minimal data set, each in a different folder. These folders' path are defined.

```
training_subject_file<-"./UCI HAR Dataset/train/subject_train.txt"
test_subject_file<-"./UCI HAR Dataset/test/subject_test.txt"
training_data_file<-"./UCI HAR Dataset/train/X_train.txt"
test_data_file<-"./UCI HAR Dataset/test/X_test.txt"
training_activity_file<-"./UCI HAR Dataset/train/y_train.txt"
test_activity_file<-"./UCI HAR Dataset/test/y_test.txt"
field_names_file<-"./UCI HAR Dataset/features.txt"
activity_labels_file<-"./UCI HAR Dataset/activity_labels.txt"
```

3.3.3   Load data

The decomposed data set are loaded into R.

```
training_subject<-read.table(training_subject_file,sep="")
test_subject<-read.table(test_subject_file,sep="")
training_data_set<-read.table(training_data_file,sep="")
test_data_set<-read.table(test_data_file,sep="")
training_activity<-read.table(training_activity_file,sep="")
test_activity<-read.table(test_activity_file,sep="")
field_names<-read.table(field_names_file,sep="")
activity_labels<-read.table(activity_labels_file,sep="")
```

3.3.4   Complete data set

Complete the data set with column names.

```
field_names<-field_names$V2
names(training_data_set)<-field_names
names(test_data_set)<-field_names
names(test_activity)<-"ActivityID"
names(training_activity)<-"ActivityID"
names(training_subject)<-"Subject"
names(test_subject)<-"Subject"
```

3.3.5 Merge 'train' and 'test' data set

Merge the 'train' and the 'test' data set.

```
data_set<-rbind(test_data_set,training_data_set)
activity<-rbind(test_activity,training_activity)
subject<-rbind(test_subject,training_subject)
```

3.3.6 Filter 'mean' and 'std' data fields

Select just the column where a mean or a std value is present.

```
data_set<-select(data_set,grep("Mean|mean|Std|std",field_names))
```

3.3.7 Append 'activity id', 'activity label' and 'subject'

Append the activity description and the subject to the data set.

```
data_set<-cbind(data_set,activity)
data_set<-mutate(data_set,ActivityLabel=activity_labels[ActivityID,2])
data_set<-select(data_set,-"ActivityID")
data_set<-cbind(data_set,subject)
```

3.3.8 Prepare new data set

Prepare the final data set by groupping data by activity and subject and calculate the mean value for each column.

```
data_set<-group_by(data_set,ActivityLabel,Subject)
new_data_set<-summarise_all(data_set,mean)
```

3.3.9 Save data

Save the data as a table.

```
write.table(new_data_set,file = "tidy_data_set.txt",row.names = FALSE)
```