README

wporter

May 19, 2014

Getting and Cleaning Data Project

Introduction

This is a exercise in cleaning up a messy data set for analysis as well as the creation of supporting documentation on how it was done. This Rmd file was used to create a pdf README file for documentation.

A full description of the experiment is available at the site where the data was obtained:

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

Briefly, the experiments have been carried out with a group of 30 volunteers within an age bracket of 19-48 years. Each person performed six activities (WALKING, WALKING.UPSTAIRS, WALKING.DOWNSTAIRS, SITTING, STANDING, LAYING) wearing a smartphone (Samsung Galaxy S II) on the waist. The resulting raw data was collected (see the "Inertial Signals" folders) and various basic statistical measurements were performed on it. The resulting data sets are in a multitude of individual file sets that need to be combined, cleaned up, and subsetted to meet the requirements of the assignment.

Instructions from the assignment

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

Codebook

The codebook file contains a detailed description of the variables in the data set and can be found in Code_Book.Rmd and CodeBook.pdf

Code to make a Tidy Data Set:

`run_analysis.R`

Load in libraries

```
library(plyr)
```

Load in the raw data sets

Data must be in your current working directory

```
#setwd("~/Documents/Coursera/GetCleanData/UCI HAR Dataset")
testData <- read.table("test/X_test.txt")
trainData <- read.table("train/X_train.txt")
testActivities <- read.table("test/y_test.txt")
trainActivities <- read.table("train/y_train.txt")
testSubject <- read.table("test/subject_test.txt")
trainSubject <- read.table("train/subject_train.txt")
featLabel <- read.table("features.txt")
actLabel <- read.table("activity_labels.txt")
```

Tidy up names

of all the feature labels without making them too long

```
featLabel$V2 <- sub("^t", "time", featLabel$V2)
featLabel$V2 <- sub("^f", "frequency", featLabel$V2)
featLabel$V2 <- gsub("-", "", featLabel$V2)
featLabel$V2 <- gsub("\\(", "", featLabel$V2)
featLabel$V2 <- gsub("\\)", "", featLabel$V2)
featLabel$V2 <- gsub("acc", "Accel", featLabel$V2)
featLabel$V2 <- gsub("mean", "Mean", featLabel$V2)
featLabel$V2 <- gsub("std", "Std", featLabel$V2)
featLabel$V2 <- sub(",", "", featLabel$V2)
```

Add names column to Data

```
colnames(testData) <- featLabel$V2
colnames(trainData) <- featLabel$V2
```

Add train/test columns to Data

```
testData$dataType <- as.character("test")
trainData$dataType <- as.character("train")
```

Add activities column to Data

```
testactivitieslist <- testActivities[,1]
testData$activities <- as.numeric(testactivitieslist)
trainactivitieslist <- trainActivities[,1]
trainData$activities <- as.numeric(trainactivitieslist)
```

Add subject column to Data

```
testsubjectlist <- testSubject[,1]
testData$subject <- as.numeric(testsubjectlist)
trainsubjectlist <- trainSubject[,1]
trainData$subject <- as.numeric(trainsubjectlist)
```

Combine trainData and testData

This is the complete data set with all the variables included

```
combinedData <- rbind(testData, trainData)
```

Change Activities values from numbers to labels

```
combinedData$activities <- gsub("1", "WALKING", combinedData$activities)
combinedData$activities <- gsub("2", "WALKING.UPSTAIRS", combinedData$activities)
combinedData$activities <- gsub("3", "WALKING.DOWNSTAIRS", combinedData$activities)
combinedData$activities <- gsub("4", "SITTING", combinedData$activities)
combinedData$activities <- gsub("5", "STANDING", combinedData$activities)
combinedData$activities <- gsub("6", "LAYING", combinedData$activities)
```

Save the full tidy data set

Even though it's not in the instructions, it's easier to make a complete tidy data set once in case you need to run different analysis in the future.

```
write.table(combinedData, "Tidy_Data.txt")
```

Subset out the mean and stddev data

```
combinedStd <- grep("Std", names(combinedData), value = TRUE)
combinedMean <- grep("Mean",names(combinedData), value = TRUE)
subColumns <- c("activities", "subject", combinedStd, combinedMean)
subData <- combinedData[,subColumns]
```

Calculate Means per Subject and Activity

```
subDataAvg <- aggregate(subData[,3:88], by=list(subData$activities, subData$subject), mean)
names(subDataAvg) <- gsub("Group.1", "activities", names(subDataAvg))
names(subDataAvg) <- gsub("Group.2", "subject", names(subDataAvg))
```

Save the Tidy Dataset to submit

```
write.table(subDataAvg, "average_subject_and_activity.txt", row.names = FALSE)
```