Clean Data

Haichen

1/5/2021

Overview

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. You will be graded by your peers on a series of yes/no questions related to the project. You will be required to submit: 1) a tidy data set as described below, 2) a link to a Github repository with your script for performing the analysis, and 3) a code book that describes the variables, the data, and any transformations or work that you performed to clean up the data called CodeBook.md. You should also include a README.md in the repo with your scripts. This repo explains how all of the scripts work and how they are connected.

Get data

```r
library(data.table)
library(plyr)

setwd("C:/Coursera/course3")
subjectTrain = read.table(’./UCI HAR Dataset/train/subject_train.txt’,header=FALSE)
xTrain = read.table(’./UCI HAR Dataset/train/x_train.txt’,header=FALSE)
yTrain = read.table(’./UCI HAR Dataset/train/y_train.txt’,header=FALSE)

subjectTest = read.table(’./UCI HAR Dataset/test/subject_test.txt’,header=FALSE)
xTest = read.table(’./UCI HAR Dataset/test/x_test.txt’,header=FALSE)
yTest = read.table(’./UCI HAR Dataset/test/y_test.txt’,header=FALSE)

features <- read.table(’./UCI HAR Dataset/features.txt’, header = FALSE)
features <- as.character(features[,2])

data.train <-  data.frame(subjectTrain, yTrain, xTrain)
names(data.train) <- c(c(’subject’, ’activity’), features)
data.test <-  data.frame(subjectTest, yTest, xTest)
names(data.test) <- c(c(’subject’, ’activity’), features)

data.all <- rbind(data.train, data.test)
```

Mean and standard deviation for each measurement

```
mean_std.locate <- grep(’mean|std’, features)
data.sub <- data.all[,c(1,2,mean_std.locate + 2)]
```

Uses descriptive activity names

```
activity.labels <- read.table(’./UCI HAR Dataset/activity_labels.txt’, header = FALSE)
activity.labels <- as.character(activity.labels[,2])
data.sub$activity <- activity.labels[data.sub$activity]
```

Appropriately labels the data set with descriptive variable names

```
name.new <- names(data.sub)
name.new <- gsub("[(][)]", "", name.new)
name.new <- gsub("^t", "TimeDomain_", name.new)
name.new <- gsub("^f", "FrequencyDomain_", name.new)
name.new <- gsub("Acc", "Accelerometer", name.new)
name.new <- gsub("Gyro", "Gyroscope", name.new)
name.new <- gsub("Mag", "Magnitude", name.new)
name.new <- gsub("-mean-", "_Mean_", name.new)
name.new <- gsub("-std-", "_StandardDeviation_", name.new)
name.new <- gsub("-", "_", name.new)
names(data.sub) <- name.new
```

Creates a second dataset

```
data.tidy <- aggregate(data.sub[,3:81], by = list(activity = data.sub$activity, subject = data.sub$subj
write.table(x = data.tidy, file = "data_tidy.txt", row.names = FALSE)
```