Getting and Cleaning Data Project

Wayne Ye

September 22, 2015

```
rm(list=ls())#clean the workspace
library(dplyr)#call dplyr lib
```

```
## Warning: package 'dplyr' was built under R version 3.2.2
```

```
##
## Attaching package: 'dplyr'
##
## The following objects are masked from 'package:stats':
##
##     filter, lag
##
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

Load the data from text files

First set the working directory to the source file location

```
#load files from the master
act_labels<-read.table("./Samsung_data/activity_labels.txt")
features<-read.table("./Samsung_data/features.txt")
test_data<-read.table("./Samsung_data/test/X_test.txt")
train_data<-read.table("./Samsung_data/train/X_train.txt")
test_subject<-read.table("./Samsung_data/test/subject_test.txt" )
train_subject<-read.table("./Samsung_data/train/subject_train.txt")
test_activity<-read.table("./Samsung_data/test/y_test.txt" )
train_activity<-read.table("./Samsung_data/train/y_train.txt")
```

Merging training and test data sets

```
raw_alldata<-rbind(test_data,train_data)
names(raw_alldata)<-features[,2] #adhere feature list to colname
rm(list=c("test_data","train_data"))#remove the raw data to save some memmory
```

Extract only the mean and standard deviation data (with patterns of “mean()” or “std()”)

```
mean_features<-grep("mean\\(\\)|std\\(\\)",features$V2)#pattern recognition
my_data<- raw_alldata[,mean_features]
rm(list=c("raw_alldata"))#remove the raw data to save some memmory
#add variable to identify activity and subject
```

Add subject_id and activity_id labels to the data set

```r
activity_id<-rbind(test_activity,train_activity)
subject_id<-rbind(test_subject,train_subject)
my_data$activity_id<-activity_id$V1
my_data$subject_id<-subject_id$V1
#clean up the act_labels
names(act_labels)<-c("activity_id","activity")
```

Uses descriptive activity names to name the activities in the data set

```r
#merge act_label with my_data
new_my_data<-merge(my_data,act_labels,by="activity_id")
rm(list=c("act_labels","activity_id","subject_id"))#remove the raw data to save some memmory
rm(list=c("test_subject","train_subject","my_data"))#remove the raw data to save some memmory
```

Appropriately labels the data set with descriptive variable names

```r
feature_names<-names(new_my_data)
new_feature_names<-gsub("^t","time_",feature_names)
new_feature_names<-gsub("^f","frequency_",new_feature_names)
new_feature_names<-gsub("Acc","Accelerometer_",new_feature_names)
new_feature_names<-gsub("Gyro","Gyroscope_",new_feature_names)
new_feature_names<-gsub("Body","Body_",new_feature_names)
new_feature_names<-gsub("Jerk","Jerk_",new_feature_names)
new_feature_names<-gsub("Gravity","Gravity_",new_feature_names)
new_feature_names<-gsub("Mag","Magnitude_",new_feature_names)
new_feature_names<-gsub("-mean\\(\\)","Mean",new_feature_names)
new_feature_names<-gsub("-std\\(\\)","StD",new_feature_names)
names(new_my_data)<-new_feature_names
```

Creates a second, independent tidy data set with the average of each variable for each activity and each subject

```r
##Group by subject and activity type
mydata_grouped<-group_by(new_my_data,subject_id,activity)
mydata_2<-summarize_each(mydata_grouped,funs(mean))
mydata_2<-arrange(mydata_2,subject_id,activity_id)
```

A glimpse to the new data set

```r
head(mydata_2[,c(1,2,4:ncol(mydata_2))])
```

```
## Source: local data frame [6 x 68]
## Groups: subject_id [1]
##
##   subject_id           activity time_Body_Accelerometer_Mean-X
##        (int)             (fctr)                          (dbl)
## 1          1            WALKING                      0.2773308
## 2          1   WALKING_UPSTAIRS                      0.2554617
## 3          1 WALKING_DOWNSTAIRS                      0.2891883
## 4          1            SITTING                      0.2612376
## 5          1           STANDING                      0.2789176
## 6          1             LAYING                      0.2215982
## Variables not shown: time_Body_Accelerometer_Mean-Y (dbl),
##   time_Body_Accelerometer_Mean-Z (dbl), time_Body_Accelerometer_StD-X
##   (dbl), time_Body_Accelerometer_StD-Y (dbl),
##   time_Body_Accelerometer_StD-Z (dbl), time_Gravity_Accelerometer_Mean-X
##   (dbl), time_Gravity_Accelerometer_Mean-Y (dbl),
##   time_Gravity_Accelerometer_Mean-Z (dbl),
##   time_Gravity_Accelerometer_StD-X (dbl), time_Gravity_Accelerometer_StD-Y
##   (dbl), time_Gravity_Accelerometer_StD-Z (dbl),
##   time_Body_Accelerometer_Jerk_Mean-X (dbl),
##   time_Body_Accelerometer_Jerk_Mean-Y (dbl),
##   time_Body_Accelerometer_Jerk_Mean-Z (dbl),
##   time_Body_Accelerometer_Jerk_StD-X (dbl),
##   time_Body_Accelerometer_Jerk_StD-Y (dbl),
##   time_Body_Accelerometer_Jerk_StD-Z (dbl), time_Body_Gyroscope_Mean-X
##   (dbl), time_Body_Gyroscope_Mean-Y (dbl), time_Body_Gyroscope_Mean-Z
##   (dbl), time_Body_Gyroscope_StD-X (dbl), time_Body_Gyroscope_StD-Y (dbl),
##   time_Body_Gyroscope_StD-Z (dbl), time_Body_Gyroscope_Jerk_Mean-X (dbl),
##   time_Body_Gyroscope_Jerk_Mean-Y (dbl), time_Body_Gyroscope_Jerk_Mean-Z
##   (dbl), time_Body_Gyroscope_Jerk_StD-X (dbl),
##   time_Body_Gyroscope_Jerk_StD-Y (dbl), time_Body_Gyroscope_Jerk_StD-Z
##   (dbl), time_Body_Accelerometer_Magnitude_Mean (dbl),
##   time_Body_Accelerometer_Magnitude_StD (dbl),
##   time_Gravity_Accelerometer_Magnitude_Mean (dbl),
##   time_Gravity_Accelerometer_Magnitude_StD (dbl),
##   time_Body_Accelerometer_Jerk_Magnitude_Mean (dbl),
##   time_Body_Accelerometer_Jerk_Magnitude_StD (dbl),
##   time_Body_Gyroscope_Magnitude_Mean (dbl),
##   time_Body_Gyroscope_Magnitude_StD (dbl),
##   time_Body_Gyroscope_Jerk_Magnitude_Mean (dbl),
##   time_Body_Gyroscope_Jerk_Magnitude_StD (dbl),
##   frequency_Body_Accelerometer_Mean-X (dbl),
##   frequency_Body_Accelerometer_Mean-Y (dbl),
##   frequency_Body_Accelerometer_Mean-Z (dbl),
##   frequency_Body_Accelerometer_StD-X (dbl),
##   frequency_Body_Accelerometer_StD-Y (dbl),
##   frequency_Body_Accelerometer_StD-Z (dbl),
##   frequency_Body_Accelerometer_Jerk_Mean-X (dbl),
##   frequency_Body_Accelerometer_Jerk_Mean-Y (dbl),
##   frequency_Body_Accelerometer_Jerk_Mean-Z (dbl),
##   frequency_Body_Accelerometer_Jerk_StD-X (dbl),
##   frequency_Body_Accelerometer_Jerk_StD-Y (dbl),
##   frequency_Body_Accelerometer_Jerk_StD-Z (dbl),
##   frequency_Body_Gyroscope_Mean-X (dbl), frequency_Body_Gyroscope_Mean-Y
##   (dbl), frequency_Body_Gyroscope_Mean-Z (dbl),
##   frequency_Body_Gyroscope_StD-X (dbl), frequency_Body_Gyroscope_StD-Y
##   (dbl), frequency_Body_Gyroscope_StD-Z (dbl),
##   frequency_Body_Accelerometer_Magnitude_Mean (dbl),
##   frequency_Body_Accelerometer_Magnitude_StD (dbl),
##   frequency_Body_Body_Accelerometer_Jerk_Magnitude_Mean (dbl),
##   frequency_Body_Body_Accelerometer_Jerk_Magnitude_StD (dbl),
##   frequency_Body_Body_Gyroscope_Magnitude_Mean (dbl),
##   frequency_Body_Body_Gyroscope_Magnitude_StD (dbl),
##   frequency_Body_Body_Gyroscope_Jerk_Magnitude_Mean (dbl),
##   frequency_Body_Body_Gyroscope_Jerk_Magnitude_StD (dbl)
```

Save to text file

```
write.table(mydata_2,file="clean_data.txt",row.names = FALSE) # save table
```