README

Rasmus Thingbak Larsen

First reads train data, and binds them together. Therafter loads test data and binds them together. Afer this the two datasets created are merged using rowbind, and the unused data sets are cleansed from memory.

setwd("C:/Users/Rasmus/Desktop/Coursera")  
X\_train <- read.csv("UCI HAR Dataset/train/X\_train.txt", sep="", header=FALSE)  
Y\_train <- read.csv("UCI HAR Dataset/train/Y\_train.txt", sep="", header=FALSE)  
subject\_train <- read.csv("UCI HAR Dataset/train/subject\_train.txt", sep="", header=FALSE)  
train <- cbind(X\_train, Y\_train, subject\_train)  
  
X\_test <- read.csv("UCI HAR Dataset/test/X\_test.txt", sep="", header=FALSE)  
Y\_test <- read.csv("UCI HAR Dataset/test/Y\_test.txt", sep="", header=FALSE)  
subject\_test = read.csv("UCI HAR Dataset/test/subject\_test.txt", sep="", header=FALSE)  
  
test <- cbind(X\_test, Y\_test, subject\_test)  
  
data <- rbind(train, test)  
  
rm(X\_train, Y\_train, subject\_train, X\_test, Y\_test, subject\_test, train, test)

Labels and features are now read from the working directory, and formatting is converted to CamelCase, to make it friendler to R and the user.

setwd("C:/Users/Rasmus/Desktop/Coursera")  
labels <- read.csv("UCI HAR Dataset/activity\_labels.txt", sep="", header=FALSE)  
features <- read.csv("UCI HAR Dataset/features.txt", sep="", header=FALSE)  
features[,2] <- gsub("-mean","-Mean", ignore.case=FALSE, features[,2])  
features[,2] <- gsub("-std","-Std", ignore.case=FALSE, features[,2])  
features[,2] <- gsub("[\\(\\),\\-]", "", ignore.case=FALSE, features[,2])

The scope of the task is defined using scopeCols, where only the columns incorporating Means or Std is kept. Everything else is removed.

scopeCols <- grep(".\*Mean\*|.\*Std.\*", features[,2])  
features <- features[scopeCols,]  
scopeCols <- c(scopeCols, 562, 563) # 562: activity, 563: subject, from data  
data <- data[,scopeCols]

Columns are renamed.

colnames(data) <- c(features$V2, "activity", "subject")

A lookup function is basically performed using gsub, to substitute the activity number with its associated name.

activity = 1  
for (i in labels$V2) {  
 data$activity <- gsub(activity, i, data$activity)  
 activity <- activity + 1  
}

The tidy data set is formed using aggregate. The data set is aggregated by the index created, which allow the mean and the std to be calculated for each type of activity, for each subject, and for each variable.

byIndex <- list(activity = data$activity, subject=data$subject)  
tidyData <- aggregate(data, by=byIndex, mean)  
tidyData <- tidyData[order(tidyData[,2], decreasing=FALSE),]  
tidyData <- tidyData[,1:88]

The tidyData file is created in the working directory.

write.table(tidyData, "tidyData.txt", row.names=FALSE)