Getting and Cleaning Data

Konstantinos Papastamos

20 March 2015

The script checks if the required dataset exists in your working directory. If not it downloads the zip file, unzips it, deletes the zip file

```
if(!file.exists("UCI HAR Dataset")){
  url="https://d396qusza40orc.cloudfront.net/"
  url = paste(url,"getdata%2Fprojectfiles%2FUCI%20HAR%20Dataset.zip",sep = "")
  download.file(url,"data")
  unzip("data")
  file.remove("data")
}
```

```
## [1] TRUE
```

Checks if package plyr is installed(we are going to need it later) and if not installs it and loads it, if yes it loads it.

```
if(require(plyr)==FALSE){
  install.packages("plyr")
}else{
  library(plyr)
}
```

```
## Loading required package: plyr
```

Reads the activity ,the feature and the subject files and stores them in the corresponding variables

```
train_activities=read.table("train\\y_train.txt",header=FALSE)
test_activities=read.table("test\\y_test.txt",header=FALSE)



train_features=read.table("train\\X_train.txt",header=FALSE)
test_features=read.table("test\\X_test.txt",header=FALSE)



train_subject=read.table("train\\subject_train.txt",header=FALSE)
test_subject=read.table("test\\subject_test.txt",header=FALSE)
```

Merges the two datasets and names the variables accordingly

```
subject=rbind(train_subject,test_subject)
features=rbind(train_features,test_features)
activities=rbind(train_activities,test_activities)

names(subject)="subjects"
names(activities)="activities"
feature_names=read.table("features.txt")
feature_names=feature_names[,2]
names(features)=feature_names
```

Creates the final dataset by adding the subject,features and activities dataframes together

```
full_dataset=cbind(subject,activities,features)
```

Reads the activity labels from the activity_labels.txt file and changes activity labels class in the dataset to factor. Then changes the factor's levels from 1,2,3,4,5,6 to the appropriate label

```
activity_labels=read.table("activity_labels.txt",header=FALSE)

full_dataset[,2]=as.factor(full_dataset[,2])

levels(full_dataset[,2])=activity_labels$V2
```

Labels the data with descriptive variable names, changing the t to time, the f to frequency, the Acc to Accelerometer, the Gyro to Gyroscope, the BodyBody to Body,the Mag to Magnitude and the tBody to timeBody(some variables did't begin with the letter t so i needed to add this last one)

```
names(full_dataset)=gsub("^f","frequency",names(full_dataset))
names(full_dataset)=gsub("^t","time",names(full_dataset))
names(full_dataset)=gsub("Acc","Accelerometer",names(full_dataset))
names(full_dataset)=gsub("Gyro","Gyroscope",names(full_dataset))
names(full_dataset)=gsub("BodyBody","Body",names(full_dataset))
names(full_dataset)=gsub("Mag","Magnitude",names(full_dataset))
names(full_dataset)=gsub("tBody","timeBody",names(full_dataset))
```

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

```
tidy_dataset=aggregate(. ~ subjects+activities,full_dataset,mean)
tidy_dataset=arrange(tidy_dataset,subjects,activities)
```