README

Getting and cleaning data

Course Project

Instructions:

1.- setwd(to the directory where you cloned this repo)

Download file at https://d396qusza40orc.cloudfront.net/getdata%2Fprojectfiles%2FUCI%20HAR%20Dataset.zip (https://d396qusza40orc.cloudfront.net/getdata%2Fprojectfiles%2FUCI%20HAR%20Dataset.zip) and unzip it in this directory. It should create a subdirectory “UCI HAR Dataset”.

2.- Run the run_analysis.R script either directly with R or using: source(‘./run_analysis.R’)

Its quite chatty and can take a little while, but if you scroll up when it finishes, you will see it executes each requirement for the course in order, print()ing helpful comments as to what it was doing at the time it was running.

The resulting dataset is in the by_activity_and_subject.txt file.

Codebook:

The codebook.txt file contains the description of the resulting dataset.

Analysis:

At first, mostly bash was what was needed to check for the main things:

```
ls -l "./UCI HAR Dataset/test"
```

```
## total 51712
## drwxr-xr-x@ 11 alex  staff       374 Jun  4 18:41 Inertial Signals
## -rw-r--r--@  1 alex  staff  26458166 Nov 29  2012 X_test.txt
## -rw-r--r--@  1 alex  staff      7934 Nov 29  2012 subject_test.txt
## -rw-r--r--@  1 alex  staff      5894 Nov 29  2012 y_test.txt
```

```
wc -l "./UCI HAR Dataset/test/X_test.txt" "./UCI HAR Dataset/train/X_train.txt"
```

```
##     2947 ./UCI HAR Dataset/test/X_test.txt
##     7352 ./UCI HAR Dataset/train/X_train.txt
##    10299 total
```

So about 30 to 70 percent in each file, as expected. But how are things ordered?

```
head -n 1 "./UCI HAR Dataset/test/X_test.txt" | wc -w
```

```
##      561
```

Ah… each line, a vector of 561 variables. That is what the documentation of the dataset says.

If this is true, then the subject file and the y_test files (each identifying the performing subject and the activity being performed by each observation), must also have the same length:

```
wc -l "./UCI HAR Dataset/test/y_test.txt" "./UCI HAR Dataset/test/subject_test.tx
t"
```

```
##     2947 ./UCI HAR Dataset/test/y_test.txt
##     2947 ./UCI HAR Dataset/test/subject_test.txt
##     5894 total
```

Yes. Same number of lines in each. Does this also happen in the train directory?

```
wc -l "./UCI HAR Dataset/train/y_train.txt" "./UCI HAR Dataset/train/subject_trai
n.txt" "./UCI HAR Dataset/train/X_train.txt"
```

```
##     7352 ./UCI HAR Dataset/train/y_train.txt
##     7352 ./UCI HAR Dataset/train/subject_train.txt
##     7352 ./UCI HAR Dataset/train/X_train.txt
##    22056 total
```

It appears so!

Ok, so we also have a file to translate activity numbers into activity names:

```
cat "./UCI HAR Dataset/activity_labels.txt"
```

```
## 1 WALKING
## 2 WALKING_UPSTAIRS
## 3 WALKING_DOWNSTAIRS
## 4 SITTING
## 5 STANDING
## 6 LAYING
```

Then that means activity files should have 1-6 values allways:

```
echo train
sort -u "./UCI HAR Dataset/train/y_train.txt"
echo test
sort -u "./UCI HAR Dataset/test/y_test.txt"
```

```
## train
## 1
## 2
## 3
## 4
## 5
## 6
## test
## 1
## 2
## 3
## 4
## 5
## 6
```

Yup. So with that info, i proceded to create the functions to load activities and subjects and then one for the full X_dirname.txt file. This last one is the most critical one and its called featureVector(dname), where dname is either “test” or “train”

All anyone would need to know about how this data was loaded and transformed is in that function.