run\_analysis README

Frank Shuster

Tuesday, November 18, 2014

### Introduction

This repository contains 2 files:

**avgdata.txt** - The output file requested in the project for Getting and Cleaning Data

**run\_analysis.R** - The R Script for creating avgdata.txt.

### Loading and running run\_analysis.R

The script only requires that you place run\_analysis.R in the working directory you have set for this project. To run run\_analysis.R on your machine

1. Save the script in your working directory
2. Run run\_analysis.R. This script will
   * Download the zip file and extract it and its files into a data sub-folder in your working directory
   * Builds the training and test tables and "stacks" them together
   * Creates friendly names for the activities being measured
   * Creates tidy names for the measured variables (see tidy discussion below for specifics)
   * Creates a new table with just mean and standard deviation measures
   * From this table the final table avgdata.txt is created that calculates the average of each of the measured columns (columns 3-75) for each (subject,activityname) pair.
   * The script then writes avg.data to your working directory

### Load avgdata.txt

You can access the output file, avgdata.txt, by either running run\_analysis in your working directory (following the steps above) or you can download avgdata.txt from the repository.

Once you have avgdata.txt in your working directory. you can inspect it by running the code below on your R commandline with the appropriate file path. Once the file is in your working directory, execute the following commands to load it into R and view it.

data <- read.table(file\_path, header = TRUE, sep = ",")

View(data)

### General Discussion and Approach for creating run\_analysis.R; creating the tidy data set

The Four components of tidy are:

1. Allow access to the raw data set
2. Provide the finished tidy dataset
3. Provide a codebook
4. Explicit steps to get from the raw to the tidy data set.

#### Allow Access to Raw Data Set

Lines 11 through 21 of run\_analysis.R download the zip file your working directory, extract all the files to a sub directory called data. This is the raw data.

#### Proivde Access to the Finished Tidy Data Set

You can access the final tidy dataset, avgdata.txt, via download from the repo or by running run\_analysis.R in your working directory.

#### Provide a Codebook

Please see the file codebook.html in the repo for the codebook for this project

#### Explicit Steps to get from the raw to the tidy dataset

The code is found in run\_analysis.R. Here is a description of the steps taken to create the tidy dataset avgdata.txt.

1. Get working directory and download zip file to /data subdirectory
2. Unzip download file into data subdirectory
3. Create training subject table
4. Create training X variables table
5. Create training response (y) table
6. Create test subject table
7. Create test X variables table
8. Create test response (y) table
9. Create feature names table
10. Create Tidy names for features table

* Convert all letters to lowercase
* Remove dashes, left & right partheneses, and commas
* Replace the starting "t" with "time"
* Replace the starting "f" with "freq"
* Replace the occurence of "bodybody" with "body"

1. Combine x and y training variables (cbind)
2. Combine x and y test variables (cbind)
3. Create Tidy names for "subject" and "activitynumber"
4. Merge dataset with activity labels to create tidy, more descriptive activity descriptions
5. Create a new table isolating columns in dataset with "mean" and "std" but not "meanfreq" along with subject and activitylabel Note: I chose to exclude the meanfreq variable as it did not meet the criteria for inclusion as a "mean" of a measurement. In this case the mean IS the measurement so it is excluded from this analysis.
6. Create the final tidy dataset, avgdata.txt, by taking the average of all the measurement variables for each (subject,activitylabel) combination
7. Write file to working directory

avgdata.txt is a tidy dataset because: 1. Each variable measured has it's own column (columns 3-75 of avgdata.txt) 2. Each observation has it's own row. An observation in this project is a (subject,activitylabel) pair. 3. We have only 1 table for analysis