**BÀI THỰC HÀNH 4: THU THẬP VÀ TIỀN XỬ LÝ DỮ LIỆU TỔNG HỢP.**

Bộ dữ liệu: **Human Activity Recognition Using Smartphones Data Set**

Link dữ liệu: [https://d396qusza40orc.cloudfront.net/getdata%2Fprojectfiles%2FUCI HAR Dataset.zip](https://d396qusza40orc.cloudfront.net/getdata%2Fprojectfiles%2FUCI%20HAR%20Dataset.zip)

1) Tải dữ liệu về và đọc dữ liệu.

2) Đọc file README.md và cho biết các thông tin sau:

a) Chức năng (nội dung) của bộ dữ liệu là gì?

b) Các tác giả của bộ dữ liệu?

c) Cấu trúc các tập tin của bộ dữ liệu? Ý nghĩa từng tập tin.

d) Các thông tin nào mà các record lưu trữ?

3) Merge bộ dữ liệu train và test lại với nhau.

4) Với các thuộc tính dữ liệu, chỉ lấy ra các thuộc tính giá trị trung bình (mean) và giá trị độ lệch chuẩn (standard deviation).

5) Đặt tên header cho bộ dữ liệu sao cho tên có ý nghĩa.

6) Lưu bộ dữ liệu tidy lại và đặt tên là: "tidy\_data.csv".

**Nộp bài: Gồm 3 phần**

*1) File báo cáo PDF trả lời cho câu 2.*

*2) Từ câu 3 tới câu 6: viết script R hoặc Python và đặt tên là: instruction\_list.R (hoặc instruction\_list.py nếu dùng Python).*

*3) Bộ dữ liệu tidy data. Đặt tên như câu 6.*

Nén lại và đặt tên theo cú pháp **<MSSV>\_<Họ tên>\_BT5.rar**. Nộp qua course (Giảng viên sẽ tạo submission sau).

*Chúc tất cả các bạn học tốt*

# **Hướng dẫn thực hiện:**

**Bước 1:** Chuẩn bị dữ liệu

- Sử dụng R để download dữ liệu về. Dùng hàm download.file trong R. File tải về là file nén .zip, đặt tên là human\_activity.zip.

- Giải nén file: sử dụng hàm unzip trong R.

**Bước 2:** Đọc dữ liệu

- Đọc dữ liệu train X\_train.txt, y\_train.txt và subject\_train.txt.

- Đọc dữ liệu test: X\_test.txt, y\_test.txt và subject\_test.txt.

**Bước 3**: Gán tên thuộc tính cho dữ liệu.

- Đọc dữ liệu từ file **featrues.txt**. Lấy cột tên các thuộc tính ra. (Sử dụng toán tử **$** để lấy được dữ liệu dạng vector). Tổng cộng có **561** thuộc tính.

- Gán tên thuộc tính lần lược cho X\_train và X\_test. Sử dụng hàm **names()**.

**Bước 4**: Merge dữ liệu:

- Merge dữ liệu train: merge column giữa X\_train, y\_train và subject\_train. Đặt tên biến là *X\_train\_data*.

Sử dụng hàm **cbind** theo thứ tự: (y\_train, subject\_train, X\_train).

Đổi tên cột *y\_train* thành "**Activity**", cột *subject\_train* thành "**Subject"**. Sử dụng hàm names(X\_train\_data)[1] <- "**Activity** ".

- Merge dữ liệu test: merge column giữa X\_test, y\_test và subject\_test. Đặt tên biến là *X\_test\_data*

*Thực hiện tương tự như merge dữ liệu train.*

- Merge dữ liệu giữa tập train và test lại với nhau.

Sử dụng hàm **rbind** - gộp dòng. Gộp dữ liệu giữa 2 bộ *X\_train\_data* và *X\_test\_data* lại với nhau.

**Bước 5:** Lấy thuộc tinh mean và standard deviation ra.

- Sử dụng lệnh **colnames()** để lấy danh sách tên cột của dữ liệu X\_data ra.

- Sử dụng lệnh **grep()** để tìm và trả về vị trí của các thuộc tính mà trong chuỗi tên có dạng **mean** và **std**. Pattern cho 2 chuỗi trên lần lượt là: "**mean\\(\\)**" và " **std\\(\\)**".

- Chỉ lấy ra các cột (thuộc tính) có dạng mean và std như trên, kèm theo 2 thuộc tính vừa thêm vào là "Activity" và "Subject" ở Bước 4.

Lệnh ví dụ: X\_mean\_std <- X\_data[,c(col\_mean,col\_std)].

col\_mean là các vị trí của thuộc tính có chứa chuỗi **mean**, col\_std là các vị trí của thuộc tính có chứa chuỗi **std**.

- Sau khi lọc ra các thuộc tính có dạng ***mean*** và ***std***, số thuộc tính (số cột) của bộ data sẽ giảm xuống nhỏ hơn 561.

**Bước 6:** Đặt tên lại cho các thuộc tính sao cho tên có ý nghĩa. Các biến cần đặt tên lại như sau:

*Bảng 1: Mô tả tên đầy đủ của các thuộc tính*

|  |  |
| --- | --- |
| **Tên pattern trước** | **Tên pattern sau khi đổi** |
| ^t | Time domain: |
| ^f | Frequency domain: |
| - | , |
| mean\\(\\) | mean value |
| std\\(\\) | standard deviation value |
| -X | on X axis |
| -Y | on Y axis |
| -Z | on Z axis |
| AccJerk | acceleration jerk |
| Acc | acceleration |
| GyroJerk | angular velocity jerk |
| Gyro | angular velocity |
| Mag | magnitude |

Sử dụng hàm **colname()** để lấy tên các thuộc tính ra. Sau đó dùng hàm **sub()** để thay thế các giá trị trong tên thuộc tính thành các tên đầy đủ như mô tả trong bảng.

**Bước 7**: Tính giá trị trung bình của từng giá trị các thuộc tính trong bộ dữ liệu theo 2 thuộc tính là *Activity* và *Subject*

Sử dụng hàm **aggregate** với các thuộc tính trong bộ dữ liệu (trừ thuộc tính **Activity** và **Subject**). Tính theo giá trị của 2 thuộc tính ID\_Activity và Subject.

**Bước 8:** Đổi tên các biến Activity.

- Các biến "Activity" (thực chất là giá trị của y\_train và y\_test) là dạng số, gọi là ID. Ứng với mỗi ID của Activity sẽ có tên mô tả tương ứng. Tên mô tả theo từng ID của biến Activity được lưu trong file **activity\_labels.txt**.

Có 2 cách làm để đổi tên các thuộc tính trong Activity:

- Cách 1: Lấy cột "Activity" trong bộ dữ liệu ra, duyệt từng phần tử và thay thế nó bằng tên tương ứng (sử dụng for loop để duyệt).

- Cách 2: Sử dụng hàm merge để merge dữ liệu giữa cột Activity và cột ID trong file **activity\_labels.txt.** Cách này sẽ tạo ra 1 cột thứ 2 trong bộ dữ liệu, ứng với mỗi ID trong cột Activity là tên của Activity tương ứng (tương tự như lệnh kết bảng - inner join trong SQL).

**Bước 9**: Lưu dữ liệu tidy thành CSV. Sử dụng hàm **write.csv()**.

*Kết quả thu được tidy data với 68 thuộc tính và 180 dòng dữ liệu (nếu không bỏ đi thuộc tính ID\_Activity thì sẽ là 69 thuộc tính).*