

Práctica 1: Introducción a RDDs: transformaciones y acciones

Esta práctica es una introducción al uso de RDDs con spark. Trabajaremos con colecciones de datos sin esquema (RDD) y utilizaremos transformaciones y acciones para procesarlos.

Ejercicios básicos

1. Calcular la media y la desviación típica de un fichero con alturas (alturas_v0.csv). Cada fila tiene una altura (en cm). Sin embargo, algunas filas tienen errores y pone -100. Asimismo, en algunas filas las alturas están en metros en vez de en centímetros. Antes de calcular la media y la desviación, hay que filtrar los datos erróneos y corregir los datos en otras unidades.

Transformaciones y acciones sugeridas: `textFile`, `map`, `reduce`, `float(str)` (Convierte una cadena a float), `filter`, `count`, etc.

2. Calcular la media y la desviación típica de un fichero con alturas (alturas.csv) por género. Cada fila tiene género y altura (en cm). Se presentan los mismos errores que en el caso precedente.

Transformaciones y acciones sugeridas: `textFile`, `map`, `reduce`, `float(str)` (Convierte una cadena a float), `filter`, `count`, `split`, etc.

Procesado de un fichero de datos de sensores

Vamos a trabajar con el conjunto de datos del *Heterogeneity Dataset for Human Activity Recognition* (HHAR) que contiene información de los sensores de movimientos de teléfonos y relojes. El enlace a los datos es: <https://archive.ics.uci.edu/ml/datasets/Heterogeneity+Activity+Recognition+Data+Set>.

Los datos contienen mediciones de sensores de movimiento mientras los usuarios realizaban determinadas acciones. El objetivo del conjunto de datos es el de reconocer las acciones que los usuarios realizan. Aunque en esta práctica nos limitaremos a procesar el fichero, que sería en cualquier caso el paso previo necesario al reconocimiento. Las posibles acciones que contiene el fichero son: 'Biking', 'Sitting', 'Standing', 'Walking', 'Stair Up' y 'Stair down'. Los sensores medidos son: giróscopo y acelerómetro. Los tipos de dispositivos son teléfonos y relojes.

Los ficheros que vamos a utilizar son: `Phones_accelerometer.csv`, `Phones_gyroscope.csv`, `Watch_accelerometer.csv` y `Watch_gyroscope.csv`. Las columnas de los ficheros son: 'Index', 'Arrival_Time', 'Creation_Time', 'x', 'y', 'z', 'User', 'Model', 'Device', 'gt'. El contenido de las distintas columnas es:

- Index: El identificador del registro.
- Arrival_Time: el tiempo de la medición cuando la medida llega a la aplicación.
- Creation_Time: Timestamp dado por el SO.
- X,y,z: Valores de la medición dados por en los ejes: x,y,z.
- User: Identificador del usuario que realiza la acción con valores de 'a' a 'i'.
- Model: Modelo del teléfono/reloj.
- Device: El aparato concreto que toma las mediciones. Para un mismo modelo pueden tener varios aparatos.

- Gt: Actividad que el usuario está realizando de entre: bike sit, stand, walk, stairsup, stairsdown and null.

Para cada ejecución de una acción por parte de un usuario, los ficheros contienen una serie de filas (mediciones) que describen el movimiento. El objetivo será el agregar usando como clave primaria la terna usuario (User), modelo (Model) y movimiento ejecutado (gt). En concreto, hay que crear un RDD (por cada fichero) con un registro por cada usuario, modelo y clase con la media, desviación estándar y valor máximo y mínimo de la secuencia del movimiento ejecutado. Una vez hecho esto, se deberá concatenar mediante join los registros de giróscopo y acelerómetro de los relojes por un lado y de los teléfonos por otro. Finalmente se creará un RDD único (mediante union) con los RDDs de teléfonos y relojes.

Ejemplo del fichero:

```
Index,Arrival_Time,Creation_Time,x,y,z,User,Model,Device,gt
0,1424696633908,1424696631913248572,-5.958191,0.6880646,8.135345,a,nexus4,nexus4_1,stand
1,1424696633909,1424696631918283972,-5.95224,0.6702118,8.136536,a,nexus4,nexus4_1,stand
2,1424696633918,1424696631923288855,-5.9950867,0.6535491999999999,8.204376,a,nexus4,nexus4_1,stand
3,1424696633919,1424696631928385290,-5.9427185,0.6761626999999999,8.128204,a,nexus4,nexus4_1,stand
```

Ejemplo de juguete:

```
0,1424696633908,1424696631913248572,-1.0,0.6,8.2,a,nexus4,nexus4_1,stand
1,1424696633909,1424696631918283972,-5.0,0.8,8.2,a,nexus4,nexus4_1,stand
```

RDD de salida posible tras procesar el fichero (pueden tener un formato distinto pero debe incluir esta información):

```
User,Model,gt,media(x,y,z),desviacion(x,y,z),max(x,y,z),min(x,y,z)
a,nexus4,stand,-3.0,0.7,8.2,2.8,0.14,0.0,-1.0,0.8,8.2,-5.0,0.6,8.2
```

Analiza los tiempos de ejecución variando el número de cores a utilizar.

Entrega

Se deberá entregar el fin de semana del 3 de marzo. La entrega se hará a través de la página del curso y será un fichero zip que contenga un notebook para los ejercicios básicos y otro para el procesamiento de datos. Los comentarios a los ejercicios se deben incluir en los propios notebooks. No se deben incluir los ficheros de datos. Se valorará tanto el correcto funcionamiento del código como su generalidad y la no repetición de código. Las funciones deben estar documentadas.