

EJERCICIO ACUMULADORES



Fons Social Europeu

L'FSE inverteix en el teu futur

UNIÓ EUROPEA

Sumario

Acumuladores.....	2
-------------------	---

Acumuladores

Dados los ficheros de log de la carpeta weblogs, queremos contar cuantas veces se han enviado datos en cada uno de los formatos: png,jpg,html,css y txt.

Usa acumuladores para recorrer los ficheros y calcular cuantos hay de cada formato en total entre todos los logs.

1.- Leemos todos los ficheros de logs

```
filename="recursos/weblogs/*"
```

2.- Inicializamos todos los acumuladores que necesitamos, vamos a crear uno por cada tipo de fichero.

```
jpgcount = sc.accumulator(0)  
htmlcount = sc.accumulator(0)  
csscount = sc.accumulator(0)  
txtcount = sc.accumulator(0)  
pngcount = sc.accumulator(0)
```

3.- Vamos a crear una función que según el fichero que encuentre, sume uno a su acumulador correspondiente.

```
def countFileType(s):  
    if '.jpg' in s: jpgcount.add(1)  
    elif '.html' in s: htmlcount.add(1)  
    elif '.css' in s: csscount.add(1)  
    elif '.txt' in s: txtcount.add(1)
```

4.- Creamos un RDD con todos los ficheros de log

```
logs = sc.textFile(filename)
```

5.- Recorremos el RDD y en cada uno de los registros de log vamos a buscar el formato, y sumaremos 1 al acumulador correspondiente

```
logs.foreach(lambda line: countFileType(line))
```

6.- Mostramos el total de los acumuladores:

```
print('Ficheros totales:')  
print('.css:', csscount.value)  
print('.html:', htmlcount.value)  
print('.jpg:', jpgcount.value)  
print('.txt:', txtcount.value)  
print('.png:', pngcount.value)
```

```
Ficheros totales:  
.css: 180632  
.html: 173049  
.jpg: 24743  
.txt: 6096  
.png: 0
```