

Universidad Internacional de Valencia

MASTER EN BIG DATA Y DATA SCIENCE

PROBLEMA 2: CATEGORÍA DE VIDEO MENOS VISTA



Procesamiento de datos masivos: Spark

Autor:
Adrián Hernández Padrón
Julio 2022

1. Código

Una vez hemos leído los datos de entrada, vamos a mapearlos con un flatMap. Debido a la naturaleza del archivo no es necesario leer las filas una a una, podemos añadir las columnas deseadas con un .split y el índice de la columna. Es necesario hacer un try-except para evitar almacenar datos vacíos, con esto solucionamos que el reduceByKey de problemas ya que no sabe como actuar con filas vacías.

```
def PrepareData(linea):
    clavevalor=[]
    #debido a la naturaleza del fichero, simplemente haciendo el .split de la
    #Es necesario hacer un try-except porque la presencia de datos vacíos ar
    #Si los datos existen los guardamos en un array, si existen los ignoramos
    try:
        clave, valor = (linea.split("\t")[3]), int(linea.split("\t")[5])
        clavevalor.append((clave, valor))
    except Exception:
        pass
    return clavevalor
```

Figura 1: Devolvemos la clave-valor con la categoría-tiempo de reproducción.

El reduceByKey que hace la suma de todos los minutos de reproducción para cada categoría, lo hacemos con lambda y después calculamos el valor mínimo entre todas las categorías.

```
#Vamos a resolver el problema con el map de la función y el reduceByKey usando lambda, después
suma = datosEntrada.flatMap(PrepareData).reduceByKey(lambda x, y: x + y).min(lambda x: x[1])
#Al hacer el .min() la variable suma pierde su estructura RDD, tras ponerlo de la forma deseada
```

El valor devuelto, que es el mínimo, pierde su estructura RDD, por tanto una vez lo coloquemos con la estructura deseada de salida, usamos .parallelize para transformarlo a RDD y poder usar saveAsTextFile.

```
resultado = [str(suma[0]) + ';' + str(suma[1])]
#Transformamos la variable resultado a RDD usando parallelize y después exportamos
spark.sparkContext.parallelize(list(resultado)).repartition(1).saveAsTextFile(salida)
```

2. Ejecución y resultados

Para lanzar el programa escribimos lo siguiente por la línea de comandos, en donde 0222 es la carpeta con todos los archivos de texto y el *.txt indica que queremos leer todos los archivos txt de la carpeta. *spark –submitCategoriaDeVideosMenosVista.py' file : /Users/adrihp/Master/MBID03/scriptsSpark/0222/*.txt' file : /Users/adrihp/Master/MBID03/scriptsSpa*
No es necesario eliminar el archivo txt que no contiene datos puesto que con el try-except se ignora sin generar ningún problema.

La entrada tiene la siguiente estructura y la salida se guarda en la carpeta salida2 y tiene la siguiente estructura.



Figura 2: Carpeta con la entrada.

LKn7ZAJ4hW0	TheReceptionist	653	Entertainment	424	I30Z1	4.34	I305	744
DjdA-5oKYFQ	NxTDln0uybo	c-8VuICzXtU	DH56yrI05nI	W1Uo5D0Ttzc	E-3zXq_r4w0		KRHfMQqSHpk	
1TCeoRPg5dE	yAR26yhuYNY	2ZgXx72XmoE	-7C1Go-YgZ0	vmdP00d6cxI	u02kj6_D8B4			
pIMp0RZthYw	1tUDz0p10pk	heqocRij5P0	XIuvoH6rUq	LGvU5DsezE0				
xiDqywcDQRM	uX81lMev6_o							
7D0Mf4Kn4Xk	periurban	583	Music	201	6508	4.19	687	312
yu06yjlVXe8	VqpnWBo-R4E	bdDskrr8jRY	y3IDp2n7B48	JngPWhfCb2M	KQaUvH5oi04			
NSzrwv5MCwc	NHB0a0xtlgU	DlRodd4s86s	EzKw0YLh-S0	eUIfRyrgwp8	AK8Wtfwe-1k			
Eg4hGkIqBGw	N1lkLaLJHlc	-uIffs-DHKM	zpTorUhCd8Y	AvSK0qPw7EU	WX5KLMqY4bM			
VKFqgoeMdiw								
n1cEq1C8oq0	Pipistrello	525	Comedy	125	1687	4.01	363	141
i30NkJ0rak	2XtLgZol5wI	3nH5Tccz8E0	bSPVayE0NhE	sEqCkwPmQ_w	hut3VRL5XRE			
bwLPSLUT-6U	ds8T05LExr0	7PSvpPXppXA	yLup8wjbSIo	lb4d1pZI9c	uRQYan-CTQ			
gnpvEvuiFoQ	F2_5K0nSsFI	DINu35v3eMU	9uSiyn7t_0o	YfShxdbAJ58	ssdfqTwZXY0			
z5wDjq8o60c								
OHkEzL4Unck	ichannel	638	Comedy	299	8043	4.4	518	371
FDIH1GNQXQE	Wtj31off8-I	mDiwzhc8dQ0	N4EYgXReBzM	NyC_0Z6zoUk	4DxyF39Myto			
aiYwo5K0VWg	M12NaXU6gms	d0VYKbEbXQ8	LQUV_XGzHmA	80mL_BJRLRw	geCFW97-f0A			
DVNwUKAuB3I	FMWYExDEJk	rE7TuuXkk4E	bWicrzq2ApQ	jh6EpXnMb18	9JhU2jE02gg			
nfBfC8hif1Y								

Figura 3: Entrada: ejemplo de un fichero

```
(base) adrihp@MacBook-Air-de-Adrian-2 Problema2 % cat salida2/part-000000
Travel & Places;57748080
```

Figura 4: Salida: Categoría de video;Tiempo de reproducción total