CDAP – Práctica 10
Apache Spark

Apartado único (obligatorio)

Al igual que en la práctica anterior, se proporcionan una serie de ficheros que contienen datos de ventas de novelas ofrecidas por distintas librerías:

- Los ficheros **`join2_book?.txt`** consisten en listas de novelas y, para cada una, la librería que la ofrece.
- Los ficheros **`join2_num?.txt`** contienen también listas de novelas y, para cada una, la cifra de ejemplares vendidos en una de las librerías.

El objetivo de este apartado es implementar un programa en python para Apache Spark que proporcione respuesta a la siguiente pregunta:

¿Dada una librería, cuál ha sido el número de ejemplares vendidos (por cualquiera de las librerías) de las novelas que ofrece dicha librería?

La implementación se realizará usando el intérprete de python para Apache Spark. Inicialmente se puede intentar implementar el programa para una determinada librería y a continuación modificarlo para que el programa admita una librería como primer argumento.

El programa debe mostrar el resultado por el terminal y escribir el resultado en un fichero **`output.txt`**. Por ejemplo, las líneas que se obtendrán si la librería elegida es "*FNAC*" son:

```
Obteniendo ventas de novelas ofrecidas por FNAC
Una_Playa_Salvaje: 54208
Una_Mirada_Mortal: 46834
,,,,
Una_Playa_Impactante: 50604
Una_Playa_Grande: 48283
Una_Decepcion_Triste: 51604
TOTAL:  3031762
```

NOTA1: para poder utilizar el ejecutable spark-submit con un programa python **`EjemplaresVendidosNovelasLibrerias.py`**, las primeras líneas de éste deben ser de la forma:

```
from pyspark import SparkContext
sc = SparkContext(appName="novelasLibrerias")
sc.setLogLevel("WARN")

from functions import *
import sys

libreria=sys.argv[1]

print ("Obteniendo ventas de novelas ofrecidas por "+libreria)
```

Y el programa se invocará con una línea:

```
$ spark-submit ./EjemplaresVendidosNovelasLibrerias.py  FNAC
```

NOTA2: en Apache Spark, se puede crear un RDD a partir de varios ficheros de texto. Para ello, se usarán comodines en la lectura:

```
novelas_ventas_files = sc.textFile("join2_num*.txt")
```

Aspectos a tener en cuenta:

- Se valorará que el código esté libre de errores y warnings y bien comentado y estructurado.
- El programa deberá ser subido al ejercicio de faitic (junto con cualquier otro fichero que se desee añadir) en formato .zip o .tar.gz. El nombre del fichero será **PR10GR#.zip** ó **PR10GR#.tar.gz** (# es el número del grupo). **Debe ser subido por todos los miembros del grupo de prácticas**.