EJERCICIO PERSISTENCIA

Sumario

Persistencia

Vamos a probar el funcionamiento de la persistencia, lo que se pretende es hacer alguna operación sobre un RDD a partir de un fichero grande y comprobar el tiempo que tarda con persistencia y sin persistencia

1.- Creamos un fichero pesado en este caso hemos creado un fichero a partir del libro del El quijote, replicando el mismo varias veces

#Creamos el RDD a partir del ficheros

```
from pyspark.sql import SparkSession
sc = SparkContext.getOrCreate()
el_quijote= sc.textFile('el_quijote.txt')#Fichero modificado para que pese mas
palabras_quijote = el_quijote.flatMap(lambda palabra:palabra.split(' '))
```

#Nos quedamos con las palabras que sean mayor que 10

```
palabras = el_quijote.flatMap(lambda palabra:palabra.split('
')).filter(lambda p: len(p)>10)
```

#Con la funcion datetime.now() obtenemos la hora de inicio y la hora de fin, para restar y saber el tiempo que ha tardado sin cachear.

```
from datetime import datetime
ini = datetime.now()
contar = palabras.count()
fin = datetime.now()
print(str(contar) + ' ' + str(fin-ini) )
```

#Cacheamos ahora el proceso

```
palabras = el_quijote.flatMap(lambda
palabra:palabra.split(' ')).filter(lambda p:
len(p)>10).cache()
```

#Repetimos la prueba,ahora con la información previamente cacheada

```
ini = datetime.now()
contar = palabras.count()
fin = datetime.now()
print(str(contar) + ' ' + str(fin-ini) )
```

RESULTADO DE LA PRUEBA:

Tiempo sin cachear:

```
In [8]: #Con la funcion datetime.now() obtenemos la hora de inicio y la hora de fin,
        #para restar y saber el tiempo que ha tardado sin cachear.
        ini = datetime.now()
        contar = palabras.count()
        fin = datetime.now()
        print(str(contar) + ' ' + str(fin-ini) )

        9917012 0:01:08.867471
```

Tiempo con cache

```
In [9]: #Repetimos la prueba,ahora con la información previamente cacheada
        ini = datetime.now()
        contar = palabras.count()
        fin = datetime.now()
        print(str(contar) + ' ' + str(fin-ini) )

        9917012 0:00:26.007728
```