

## Contador de palabras con pyspark

RDD de palabras se convierte en un RDD de pares clave/valor donde la clave es la palabra y el valor el número de veces que aparece la palabra en el texto

```
rdd=sc.textFile("ruta_fichero")
words=rdd.flatMap(lambda x:x.split(" "))
result=words.map(lambda x:(x,1)).reduceByKey(lambda x,y:x+y)
```

```
In [24]: words = sc.textFile("book2.txt")
```

```
In [25]: words.filter(lambda w: w.startswith(" ")).take(5)
```

```
Out[25]: [u'      from The Works of Theophile Gautier Volume 19',
u' THE WORKS OF',
u' TH\x9cOPHILE GAUTIER',
u' VOLUME NINETEEN',
u' TRANSLATED AND EDITED BY']
```

```
In [26]: counts = words.flatMap(lambda line: line.split(" ")).map(lambda word: (word, 1)).reduceByKey(lambda a, b: a + b)
```

```
In [27]: counts.collect()
```

```
Out[27]: [(u'', 592),
(u'donate', 1),
(u'yellow', 1),
(u'four', 4),
(u'catch', 1),
(u'protest', 1),
(u'sleep', 3),
(u'right.\u201d', 1),
(u'appetite', 3),
```