



Chapter 3: Spark RDDs

Ex2: Pair RDDs - 5000 points

Cho dữ liệu 5000_points.txt

1. Đọc dữ liệu => data. Có bao nhiêu element trong data? In 5 element đầu tiên.
2. Tạo PairRDD có tên là pair_data từ data trên, với mỗi element của data sẽ tạo thành 1 PairRDD là tuple có 2 phần tử kiểu int được tách ra bởi dấu phân tách "\t". In 5 element đầu tiên.
3. Tạo pair_data_sort từ pair_data với key được sắp giảm dần. In 5 element đầu tiên.
4. Với pair_data_sort, hãy đếm số lượng các item theo key. In ra những key có số item >1
5. Tạo pair_data_groupby từ pair_data_sort bằng cách nhóm các value có cùng key. In ra các key có số item >1
6. Tạo pair_data_reduce từ pair_data_sort, với các value có cùng key thì lấy giá trị lớn nhất trong các value. Cho biết số phần tử của pair_data_reduce này.

```
In [1]: import findspark
findspark.init()
```

```
In [2]: import pyspark
from pyspark import SparkContext
from pyspark.sql import SparkSession
```

```
In [3]: sc = SparkContext()
```

```
In [4]: #1.
data = sc.textFile("5000_points.txt", minPartitions=3)
print("The type of data is", type(data))
```

The type of data is <class 'pyspark.rdd.RDD'>

```
In [5]: print("Number of elements:", data.count())
```

Number of elements: 5000

```
In [6]: data.take(5)
```

```
Out[6]: ['664159\t550946',
'665845\t557965',
'597173\t575538',
'618600\t551446',
'635690\t608046']
```

```
In [7]: #2.
pair_data = data.map(lambda s: (int(s.split('\t')[0]), int(s.split('\t')[1])))
```



```
In [8]: pair_data.take(5)
```

```
Out[8]: [(664159, 550946),
         (665845, 557965),
         (597173, 575538),
         (618600, 551446),
         (635690, 608046)]
```

```
In [9]: #3.
pair_data_sort = pair_data.sortByKey(ascending=False)
```

```
In [10]: pair_data_sort.take(5)
```

```
Out[10]: [(961951, 577029),
         (937823, 540173),
         (932662, 537069),
         (927170, 250593),
         (925732, 210388)]
```

```
In [11]: #4.
for key, val in pair_data_sort.countByKey().items():
    if val > 1:
        print(key, val)
```

```
871544 2
865489 2
838036 2
826192 2
805089 2
804251 2
620330 2
618869 2
393758 2
341966 2
338474 2
186380 2
166973 2
```

```
In [12]: #5.
pair_data_groupby = pair_data.groupByKey().collect()
```



```
In [13]: for x, y in pair_data_groupby:
          if len(list(y)) > 1:
              print(x, list(y))
```

```
805089 [762512, 96540]
838036 [749612, 542999]
826192 [172906, 577784]
865489 [161762, 548920]
618869 [577243, 398187]
804251 [329956, 331304]
393758 [750953, 439738]
871544 [144135, 592403]
338474 [563350, 564151]
341966 [561987, 586266]
186380 [363938, 487625]
166973 [341636, 334724]
620330 [398630, 396776]
```

```
In [14]: #6.
          import math
          pair_data_reduce = pair_data.reduceByKey(lambda x, y: max(x,y))
```

```
In [15]: print("Number of elements:", pair_data_reduce.count())
```

```
Number of elements: 4987
```

```
In [16]: # check
          for x, y in pair_data_reduce.collect():
              if x==620330:
                  print(x, y)
```

```
620330 398630
```