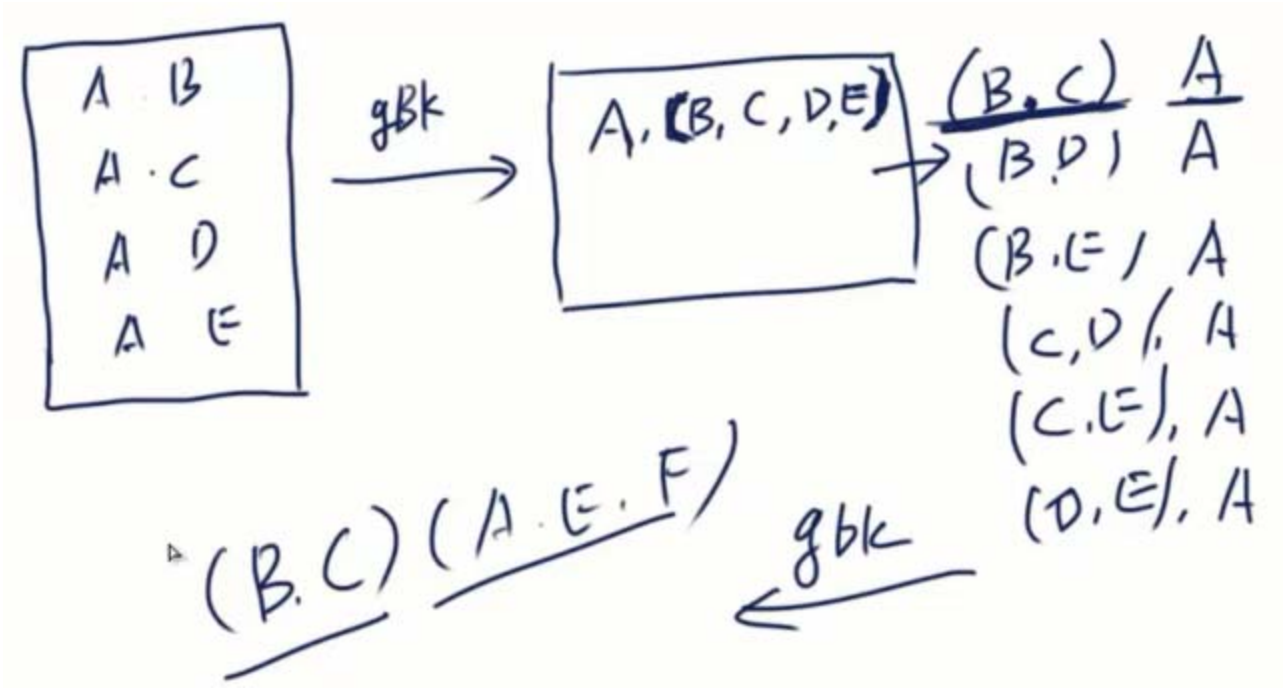
pysparkTJhw04

根据好友数据集，完成统计两个人的共同好友：
输入数据格式：
A,B
C,D

输出数据格式：
(A,B) -->(C,D,E,F)




```
from itertools import combinations

def combine_friends(key,values):
  iters = combinations(list(values),2)
  for us in iters:
    yield(us,key)

file='file:///i:/friends.txt'
rdd = sc.textFile(file)
rdd1 = rdd.map(lambda x:x.split('\t')).groupByKey()
rdd2=rdd1.flatMap(lambda x:combine_friends(x[0],x[1]))
rdd3=rdd2.groupByKey()
rdd4=rdd3.mapValues(list)
```

In [1]:

```
import pyspark
from pyspark.sql import SparkSession
```

In [2]:

```
#生成SparkSession实例
spark = SparkSession.builder \
      .master("local[*]") \
      .appName("pysparkTJhw04") \
      .config("spark.some.config.option", "some-value") \
      .getOrCreate()
```

In [3]:

```
#通过sparkSession获取上下文
sc = spark.sparkContext
```

In [6]:

```
%pwd
```

Out[6]:

```
'/home/ian/code/github/LSCJcourses/pysparkTJ'
```

In [11]:

```
rdd1 = sc.textFile('file:///home/ian/code/github/LSCJcourses/pysparkTJ/friends.txt'
```

In [13]:

```
rdd2 = rdd1.map(lambda x:x.split('\t'))
```

In [17]:

```
rdd3 = rdd2.groupByKey().map(lambda x:(x[0],list(x[1])))
```

In [18]:

```
import itertools
```

In [20]:

```
list(itertools.combinations([1,2,3],2))
```

Out[20]:

```
[(1, 2), (1, 3), (2, 3)]
```

In [23]:

```
rdd4 = rdd3.flatMap(lambda x:[(i,x[0]) for i in list(itertools.combinations(x[1],2)
```

In [24]:

```
rdd5 = rdd4.groupByKey().map(lambda x:(x[0],list(x[1])))
```

In [26]:

```
rdd5.take(5)
```

Out[26]:

```
[(('0', '136593'), ['867923']),
 (('0', '523684'), ['867923']),
 (('0', '815602'), ['867923']),
 (('0', '835220'), ['867923']),
 (('0', '857527'), ['867923', '891835'])]
```

end

end

end

end

end

In [ ]:

```

```