1、字典连接

文件 sheet.txt 内容如下

00001 1

00002 2

文件 product.txt 内容如下

1 皮鞋

2 衣服

希望输出结果

00001 皮鞋

00002 衣服

2、分组去重并计算

原始数据

2010-05-04 12:50,10,10,10

2010-05-05 13:50,20,20,20

2010-05-06 14:50,30,30,30

2010-05-05 13:50,20,20,20

2010-05-06 14:50,30,30,30

2010-05-04 12:50,10,10,10

2010-05-04 11:50,10,10,10

结果数据

2010-05-04 11:50,10,10,10

2010-05-04 12:50,20,20,20

2010-05-05 13:50,40,40,40

2010-05-06 14:50,60,60,60

3、单词计数

4、词频计数 Top K

5、倒排索引

提示：按照|分隔字符串时，因为|是正则表达式的特殊字符，因此使用split(“\\|”)写法才有效。

原始数据

cx1|a,b,c,d,e,f

cx2|c,d,e,f

cx3|a,b,c,f

cx4|a,b,c,d,e,f

cx5|a,b,e,f

cx6|a,b,c,d

cx7|a,b,c,f

cx8|d,e,f

cx9|b,c,d,e,f

结果数据

d|cx1,cx2,cx4,cx6,cx8,cx9

e|cx1,cx2,cx4,cx5,cx8,cx9

a|cx1,cx3,cx4,cx5,cx6,cx7

b|cx1,cx3,cx4,cx5,cx6,cx7,cx9

f|cx1,cx2,cx3,cx4,cx5,cx7,cx8,cx9

c|cx1,cx2,cx3,cx4,cx6,cx7,cx9