金融大数据-实验3

1、阶段一

> 任务（MapReduce）：
>
> 精简数据集：淘宝双十一用户购物数据集（100万条），见附件 million_user_log.csv.zip
>
> 基于精简数据集完成MapReduce作业：
>
> o 统计各省的双十一前十热门关注产品（“点击+添加购物车+购买+关注”总量最多前10的产品）
>
> o 统计各省的双十一前十热门销售产品（购买最多前10的产品）

1.1、实验环境

- Win10
- hadoop-3.0.0
- pycharm
- 本地运行

我们将用Python为Hadoop编写一个简单的MapReduce程序，但是不使用Jython将代码转换成Java jar文件。

参考教程:

1、如何用python编写mapreduce程序——以词频统计为例

https://www.michael-noll.com/tutorials/writing-an-hadoop-mapreduce-program-in-python/#test-your-code-cat-data--map--sort--reduce

1.2、map.py

- key：省份
- value：item_id+action

```
# coding=UTF-8
#!/usr/bin/env python
"""map.py"""

import sys
for row in sys.stdin:
    row = row.strip()
    users = line.split(',')
    print ('%s\t%s,%s' % (users[10],users[1], users[7]))//省份、商品id、行动action
```

1.3 reducer.py

- 数据结构：字典dict={}
- dict[key]=[{关注},{销售}]，key值表示一个包含双字典的列表。

```
# coding=UTF-8
# !/usr/bin/env python
"""reducer.py"""
```

```
from operator import itemgetter
import sys

key={}
# input comes from STDIN
for row in sys.stdin:
    row = row.strip()
    # from map.py
    province, value = row.split('\t')
    item_id, action = value.split(',')
    sales=int(action)
    if province not in key:
        key[province]=[{},{}]
        commodity_top10_1=key[province][0]
        commodity_top10_2=key[province][1]
        commodity_top10_1[item_id]=commodity_top10_1.get(item_id,0)+1
        if sales == 2:
            commodity_top10_2[item_id]=commodity_top10_2.get(item_id,0)+1
    else:
        commodity_top10_1=key[province][0]
        commodity_top10_2=key[province][1]
        commodity_top10_1[item_id]=commodity_top10_1.get(item_id,0)+1
        if sales == 2:
            commodity_top10_2[item_id]=commodity_top10_2.get(item_id,0)+1

for province in key:
    commodity_top10_1=key[province][0]
    commodity_top10_2=key[province][1]
    sort1 = sorted(commodity_top10_1.items(),key=lambda x: x[1], reverse=True)
    sort2 = sorted(commodity_top10_2.items(),key=lambda x: x[1], reverse=True)
    print(province+':'+'\n')
    print('前十热门关注产品:')
    for i in range(10):
        print(sort1[i], end=" ")
    print('\n')
    print('前十热门销售产品:')
    for k in range(10):
        print(sort2[k], end=" ")
    print('\n')
```

1.4、运行结果

终端运行程序：

```
type million_user_log.csv | python map.py | python reducer.py
```

运行结果见文件夹：阶段一运行结果截图。格式为：【省份+前世热门关注+前十热门销售】。

1.5 反思

中途被输出时的省份乱码坑了好久。。。转成utf-8编码也没有用。后来找到一篇教程：《解决 Excel 打开 UTF-8 编码 CSV 文件乱码的 BUG》https://blog.csdn.net/leonzhouwei/article/details/8447643

这篇教程还是不错的，照着方法2做即可。