#1、最热门的国家签到排行（除中国外）

#1.1 API请求出经纬对应的国家省份城市信息，得到文件：lat\_lon\_country\_uniq

#1.2 合并2018-5.1-5.5数据到weibo\_locate.2018-05.1-5，把国家等信息拓展weibo\_locate.2018-05.1-5的12，13，14列

awk -F "\t" 'NR==FNR{a[$1$2]=$3"\t"$4"\t"$5;next}{if(a[$8$9]) print $0"\t"a[$8$9]; else print $0"\t""NA""\t""NA""\t""NA"}' lat\_lon\_country\_uniq weibo\_locate.2018-05.1-5 > new\_weibo\_locate

#得到文件是源数据上添加三列：国家-省份-城市（国外则仅有国家，有的经纬度API请求不到数据拓展列设置为“NA”）

#1.3

#提取出userid和国家

cat new\_weibo\_locate | cut –f5,12 | awk '{if ($2!="中国") print $1"\t"$2}' | sort| uniq >userid\_country

#统计国家热门排行榜

cat userid\_country | cut -f2 | sort | uniq -c | sort -k1rn > cnt\_country

**#最热门的国家签到排行榜如下：**

**#6728 Japan**

**#4714 United States**

**#3932 Thailand**

**#2931 Australia**

**#2615 South Korea**

**#2534 United Kingdom**

**#…**

#2、最热门的中国城市排行

#2.1

#提取出userid和中国省份信息

cat new\_weibo\_locate | cut -f5,12,13 | awk '{if ($2=="中国") print $1"\t"$3}' | sort | uniq > userid\_province

#2.2

#统计热门省份（去掉直辖市）

cat userid\_province | cut -f2 | sort | uniq -c | sort -k1rn | awk '{if (!match ($2,".\*市")) print $1"\t"$2}'>cnt\_province

**#热门省份排行榜：**

**#56571 广东省**

**#46351 山东省**

**#42992 江苏省**

**#37447 浙江省**

**#32144 四川省**

**#31570 河南省**

**#...**

#3、微博用户签到城市数量排名和签到数量排名，join到一张表，按照城市签到数量(用户)倒叙排列

#3.1

#提取出userid和对应城市信息

cat new\_weibo\_locate | cut -f5,12,14 | awk '{if ($2=="中国") print $1"\t"$3}' | sort | uniq > userid\_city

#统计热门城市（按用户签到数量）

cat userid\_city | cut -f2 | sort | uniq -c | awk '{print $2"\t"$1}' > cntUserid\_city

#3.2

#提取出weiboid和对应城市信息

cat new\_weibo\_locate | cut -f4,12,14 | awk '{if ($2=="中国") print $1"\t"$3}' | sort | uniq > weiboid\_city

#统计热门城市（按微博签到数量）

cat weiboid\_city | cut -f2 | sort | uniq -c | awk '{print $2"\t"$1}' > cntWeiboid\_city

#3.3

#join到一张表排序

join cntUserid\_city cntWeiboid\_city  | sort -k2rn > join\_city\_userid\_weiboid

#热门城市排行榜：

**#city userid\_cnt weiboid\_cnt**

**#北京市 40842 53557**

**#上海市 26919 34426**

**#成都市 20204 26270**

**#西安市 17666 22745**

**#广州市 17304 21973**

**#深圳市 15702 20270**

**#...**

#4、找到在5月份（实际统计1-5号这五天）有跨国旅行的人，计算这部分人在微博总用户的比例

#4.1

#计算跨国旅行的人数（userid\_country 组合去重）

cat new\_weibo\_locate | cut -f5,12 | awk -F"\t" '{if ($2!="中国") print $1"\t"$2}' | sort | uniq | wc –l

#返回：41334

#4.2

#总人数计算

cat new\_weibo\_locate | cut -f5,12 | sort | uniq | wc -l

#返回：656676

**#跨国旅行比例: 41334/656676= 6.3%**

#5、跨国旅行的人喜欢用什么手机，首先需要找到有过跨国旅行的人，然后去5月17日那一天的微博数据中去兑出他们可能的手持设备，计算结果是这样一个排行榜，同类型设备需要归并，比如小米各种型号要归并成小米（要用到正则），苹果各种型号也要归并到苹果，其他杂牌手机不要求合并，只合并小米、苹果、华为、Oppo和Vivo这五个。

#5.1

#提取出跨国旅行人的userid

cat new\_weibo\_locate | cut -f5,12 | awk -F"\t" '{if ($2!="中国") print $1"\t"$2}' | sort | uniq > userid\_travel\_country

#5.2

#去之前一份总数据中提取出userid\_device信息（存为文件：userid\_device\_uniq 大概有500万userid）

#然后兑出userid的手持设备信息

awk -F "\t" 'NR==FNR{a[$1]=$2;next}{if(a[$1]) print $1"\t"a[$1]; else print $1"\t""NA"}' userid\_device\_uniq userid\_travel\_country > userid\_device\_travel\_abroad

#5.3

#归并5个指定品牌，统计用户数量（运行 count\_top5.sh）

#!/bin/bash

for kw in "小米|红米"  "苹果|iPhone"  "华为|荣耀|huawei"  "Oppo"  "Vivo"

do

cmd0="echo -n `echo \"$kw\" | cut -d '|' -f1`\$'\t' >> device\_travel\_abroad\_Top5"

eval  $cmd0

cmd1="grep -c -E -i \"$kw\" device\_travel\_abroad >> device\_travel\_abroad\_Top5"

eval $cmd1

done

#排除5个指定品牌,统计其余用户数量

grep -v -E -i "小米|红米|苹果|iPhone|华为|荣耀|huawei|Oppo|vivo" device\_travel\_abroad | sort | uniq -c |sort -k 1rn | awk '{print $2"\t"$1}' > device\_travel\_abroad\_withoutTop5\_cnt

#5.4

#合并排序

cat device\_travel\_abroad\_withoutTop5\_cnt device\_travel\_abroad\_Top5 | sort -k2rn > device\_travel\_ranking

#跨国旅行的人热门手机排行榜（去掉了手机设备为NA的情况，因为有部分userid从5.17号大文件中兑不出相应的设备信息）

**#手机品牌 用户数量**

**#苹果 9939**

**#华为 595**

**#微博 541**

**#微博故事 152**

**#小米 143**

**#Vivo 131**

**#三星android智能手机 106**

**#Oppo 96**

**#...**

PS: 该作业缺失任务4。