# script06\_1.sh

#运行 ./script06\_1.sh <NICK\_NAME>

user=$1

#获得该用户全部微博

cut -f 6,9,10 weibo.top10wan | grep "^$user" | sort -u | cut -f3 | sed 's/#//g'> 'temp\_'$user'\_weibo.txt'

echo ' ' > temp\_total\_weibo.txt

#对其微博进行分词

cat 'temp\_'$user'\_weibo.txt' | while read line

do newline=${line//#/ }

wget "http://api.pullword.com/get.php?source=$newline&param1=0.6&param2=1" -O temp\_weibo.txt

cat temp\_weibo.txt | cut -d: -f1 | sort -u >> temp\_total\_weibo.txt

done

#计算词频

weibo\_cnt=$((cat 'temp\_'$user'\_weibo.txt') | wc -l)

grep -v '^\s\*$' temp\_total\_weibo.txt | sort | uniq -c | awk '{print $2"\t"$1'/'cnt}' cnt="$weibo\_cnt" | sort -k2rn -t$'\t' > hw6.1\_result

# script06\_2.sh

# 运行 ./script06\_2.sh <KEYWORD>

keyword=$1

cut -f 6,9,10 weibo.top10wan | sort -u > temp\_all\_weibo.txt

#获得user

cat temp\_all\_weibo.txt | awk -F"\t" '{if($3~key)print $1;}' key="$keyword" | sort -u > temp\_keyword\_users.txt

echo " " > temp\_user\_freq.txt

#计算keyword词频

cat temp\_keyword\_users.txt | while read line

do total=$(grep "^$line" temp\_all\_weibo.txt | wc -l)

keynum=$(grep "^$line" temp\_all\_weibo.txt | cut -f3 | grep $keyword | wc -l)

re=$(awk -v key="$keynum" -v to="$total" 'BEGIN{print key/to}')

echo -e $line"\t"$re >> temp\_user\_freq.txt

done

cat temp\_user\_freq.txt | sort -k2,2rn > hw6.2\_result

本次作业注意事项：

1. 微博去重：微博去重应按照weiboid去重，不能单纯只看微博内容。有可能出现微博内容相同但并非同一条微博的情况，如：“转发微博”。
2. 文档频率的含义是这个文档中出现了就计算一次，而不是多次。因此，在统计时，每条微博中重复的词只应计算一次。参考思路：对每一条微博进行分词。完成分词后，对本条微博中的分词去重。