因为大文件没下完，使用的是小文件

<PATH\_TO\_weibo**.**top10wan> 指代文件 weibo.top10wan 的绝对路径

### 计算出微博 1 天的数据总量：

awk 'END{print NR}' <PATH\_TO\_weibo**.**top10wan>

输出：100000

###  微博发文user\_id总量:

awk 'BEGIN{FS="\t";total=0}{sum[$5] ++}; END{ for (id in sum) {total++} {print total}}' <PATH\_TO\_weibo**.**top10wan>

输出：47661

###  微博文章总量:

|  |
| --- |
| awk 'BEGIN{FS="\t";total=0}{sum[$9] ++}; END{ for (id in sum) {total++} {print total}}' <PATH\_TO\_weibo**.**top10wan> |

输出：96688

### 计算发文(weibo\_id)最多的top 100个用户：

|  |
| --- |
| awk 'BEGIN{FS="\t"}{if(!seen[$5,$9]++) sum[$5]++;}; END{ for (id in sum) {print id, sum[id]} }' <PATH\_TO\_weibo**.**top10wan> | sort -r -nk2 | head -n 100 > test1\_3**.**txt |

输出(前面是 user\_id, 后面是该 user\_id 的 distinct count(weibo\_id))：

6529707947 38

2112339023 36

1750349294 30

5549358119 22

2275883675 22

6017112148 21

5666958393 21

5470635923 21

3275115274 21

2028810631 21

2000961721 21

1789250890 21

1664176597 21

1644114654 21

1402977920 21

6547519225 20

6546803187 20

6543339781 20

6541521551 20

6541519982 20

6540602382 20

6540449397 20

6540256162 20

6539227175 20

6539222506 20

6539204661 20

6539199106 20

6539196906 20

6539188269 20

6539184460 20

6539182738 20

6539179327 20

6539175885 20

6539122291 20

6539086042 20

6539081082 20

6539068477 20

6539059540 20

6539009621 20

6539008560 20

6538994856 20

6538992295 20

6538975374 20

6538950043 20

6538947265 20

6538943889 20

6538939385 20

6538933249 20

6538929464 20

6538929039 20

6538925576 20

6538918894 20

6538916848 20

6538911838 20

6538907040 20

6538899479 20

6538896294 20

6538868185 20

6538865611 20

6538820010 20

6538800040 20

6538774841 20

6538765187 20

6538745552 20

6538732433 20

6538731105 20

6538704716 20

6538674993 20

6538669514 20

6538655676 20

6538651778 20

6538650509 20

6538650054 20

6538648594 20

6538646649 20

6538641102 20

6538627901 20

6538624610 20

6538612491 20

6538612120 20

6538604110 20

6538603144 20

6538602090 20

6538595961 20

6538590143 20

6538588969 20

6538586954 20

6538583938 20

6538579937 20

6538578005 20

6538573966 20

6538572591 20

6538570906 20

6538564887 20

6538562653 20

6538548741 20

6538531114 20

6538531080 20

6538530100 20

6538529674 20