谁是冷场王

天才唐

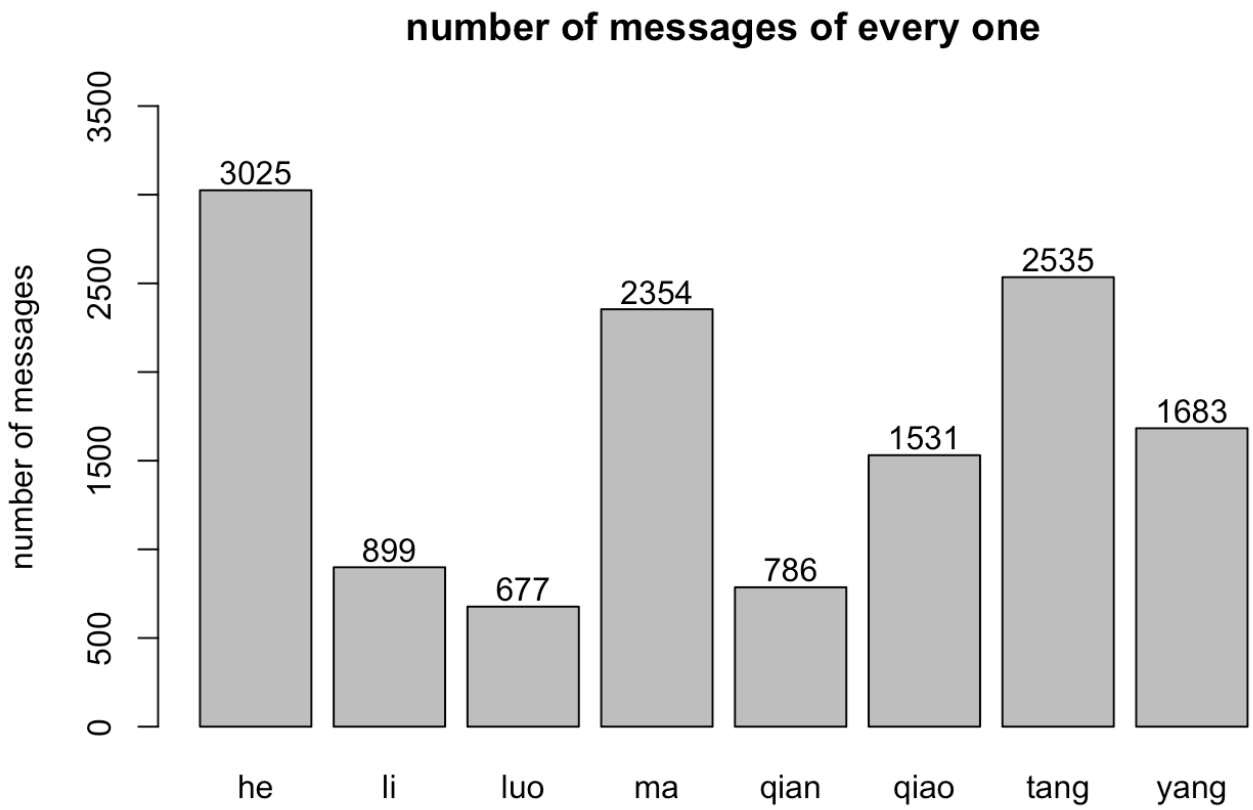
2015年11月2日

读取EMIT群的聊天记录(2014-11-02 06:12:05 ~ 2015-10-30 11:37:44)，进行预处理，将时间拆分表示，精确到分钟，并把昵称简化为英文形式，忽略聊天的具体容，将处理完成后的数据读取入表格。

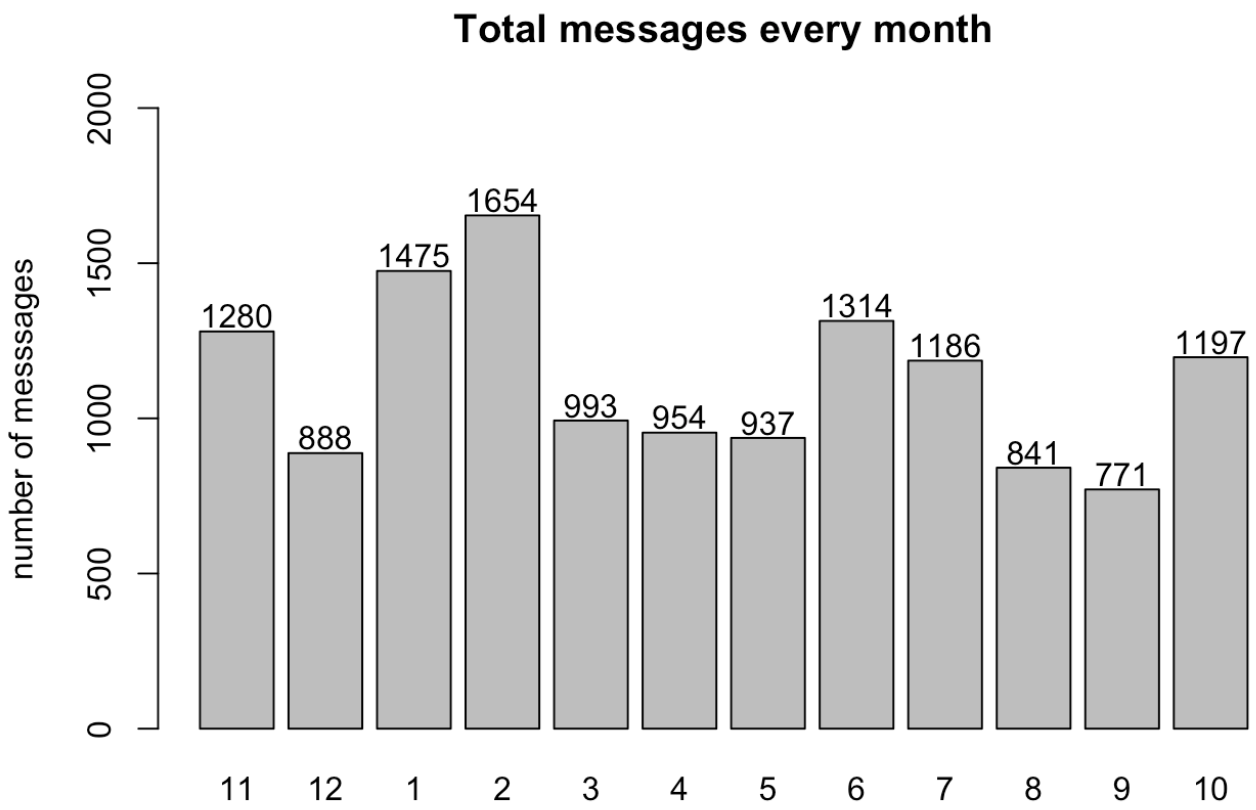
表格的前10行以及整个表格的简单总结：

```
##    year month day hour minute name
## 1  2014    11   2    6     12   he
## 2  2014    11   2    7     48 yang
## 3  2014    11   2    7     51 yang
## 4  2014    11   2    7     52 yang
## 5  2014    11   2   12     30 tang
## 6  2014    11   2   12     30 tang
## 7  2014    11   2   12     30 tang
## 8  2014    11   2   12     32  luo
## 9  2014    11   2   12     34   he
## 10 2014    11   2   12     36   he
```

总记录天数为363天，总发言数为13490, 可以发现瑞哥的话是最多的(3025)，而罗导是最含蓄的(677)。

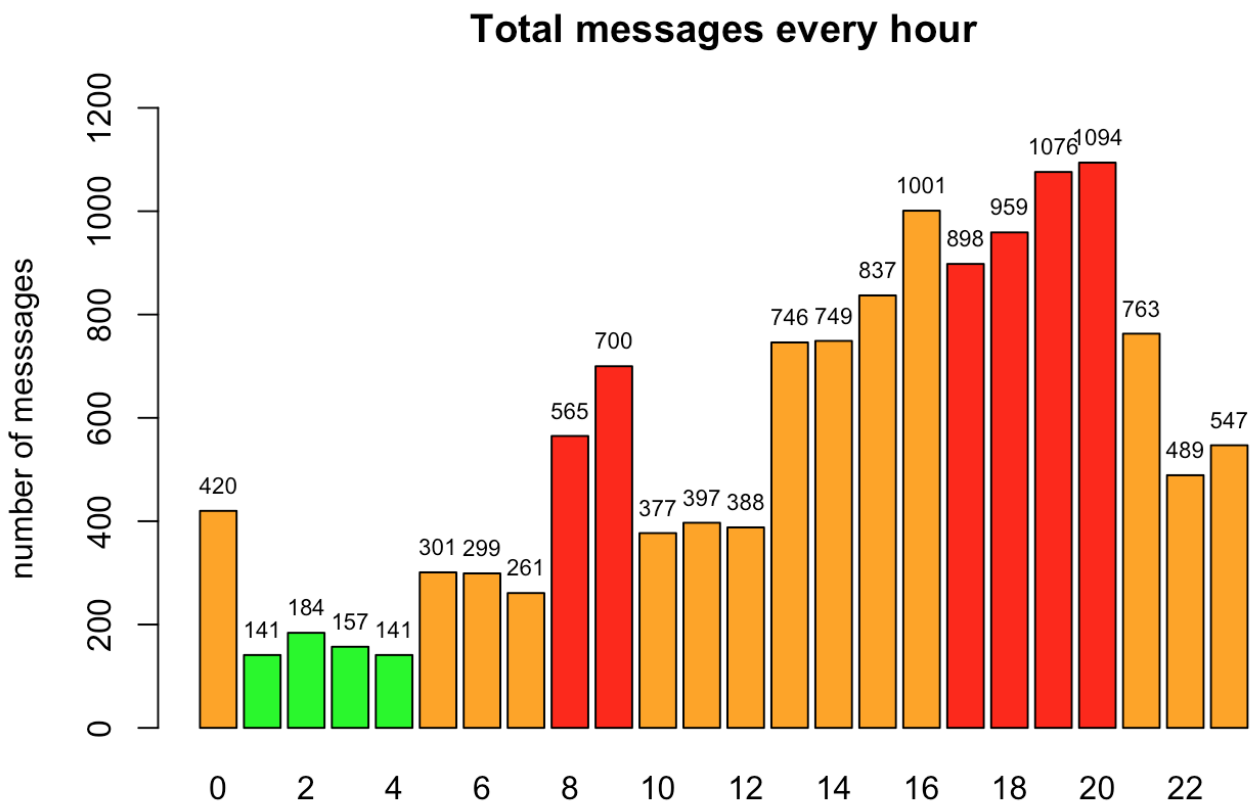


画出每个月的总消息数的柱状图，可以看出消息数多的月份并不是大家聚在一起的月份，而是在其前后的月份。比如14年12月大家聚集在LA，因此并不需要微信沟通，相反在12月前后因为要商量行程以及分离后过分思念彼此（有点肉麻）微信消息数显著上升。8, 9月份同理。



从小时方面看，每天最活跃的是太平洋时间的4pm-8pm，此时为东部时间的7pm-11pm，北京时间的7am-11am，刚好是最可能全部群成员都在线的时间。三个巨大的降幅分别出现在太平洋时间的的8pm-9pm，0am-1am，9am-10am，也对应着东部11pm-12pm，西部0am-1am，中国0am-1am。正好是三个地区的入睡时间，从入睡时间的差异也可以看出群内东部的成员睡觉比较早。如果按照所有人8点起

床，那么红色部分代表所有人都在线的时间，橘黄色代表两个时区的人在线的时间，绿色代表只有一个时间的人在线的时间，柱状图颜色的深浅与高度基本吻合。



重新整理数据，以第一条消息为起始时间，计算之后每天消息的时间，以分钟为单位，并按照('he', 'li', 'luo', 'ma', 'qian', 'qiao', 'tang', 'yang')的顺序给姓名编号，整理好后的表格前10行如下：

```
##    minute name id
## 1       0   he  1
## 2      96 yang  8
## 3      99 yang  8
## 4     100 yang  8
## 5     378 tang  7
## 6     378 tang  7
## 7     378 tang  7
## 8     380  luo  3
## 9     382   he  1
## 10    384   he  1
```

对"话题"以及"冷场次数"的定义：

定义一个时间常量$Cold$，同时初始化一个话题计数器$COUNTER = 0$，每当一个人发消息，计数器加1($COUNTER+ = 1$)直到在某个人发完消息后$Cold$时间内没有人回复则判定最后发言的那个人冷场，此次话题终止，$COUNTER$记录了这次话题中所有人发言的总次数，我们定义为这次话题的热度($Popularity$)。同时下一个话题开始，直到$Cold$间隔内没人发言，那么再次判定最后发言的人冷场。在此报告中指定$Cold = 90$，也就是1个半小时内没人回复定为冷场。

对表格进行遍历和统计，得到一个$ColdMatrix$，其中记录了每个人在每一个话题中的发言数，前10个话题的统计结果如下：

```
##      X1 X2 X3 X4 X5 X6 X7 X8 X9 X10
## he    1  0 16  0  2  0  0  0  0   0
## li    0  0  1  2  0  0  2  3  0   0
## luo   0  0  1  0  0  0  0  0  0   0
## ma    0  0  1  0  0  1 14  3  2   1
## qian  0  0  0  0  0  0  0  2  0   0
## qiao  0  0  1  0  0  0  0  0  0   0
## tang  0  0  8  0  2  1  7  2 11   0
## yang  0  3  3  0  0  0  0  1  5   0
```

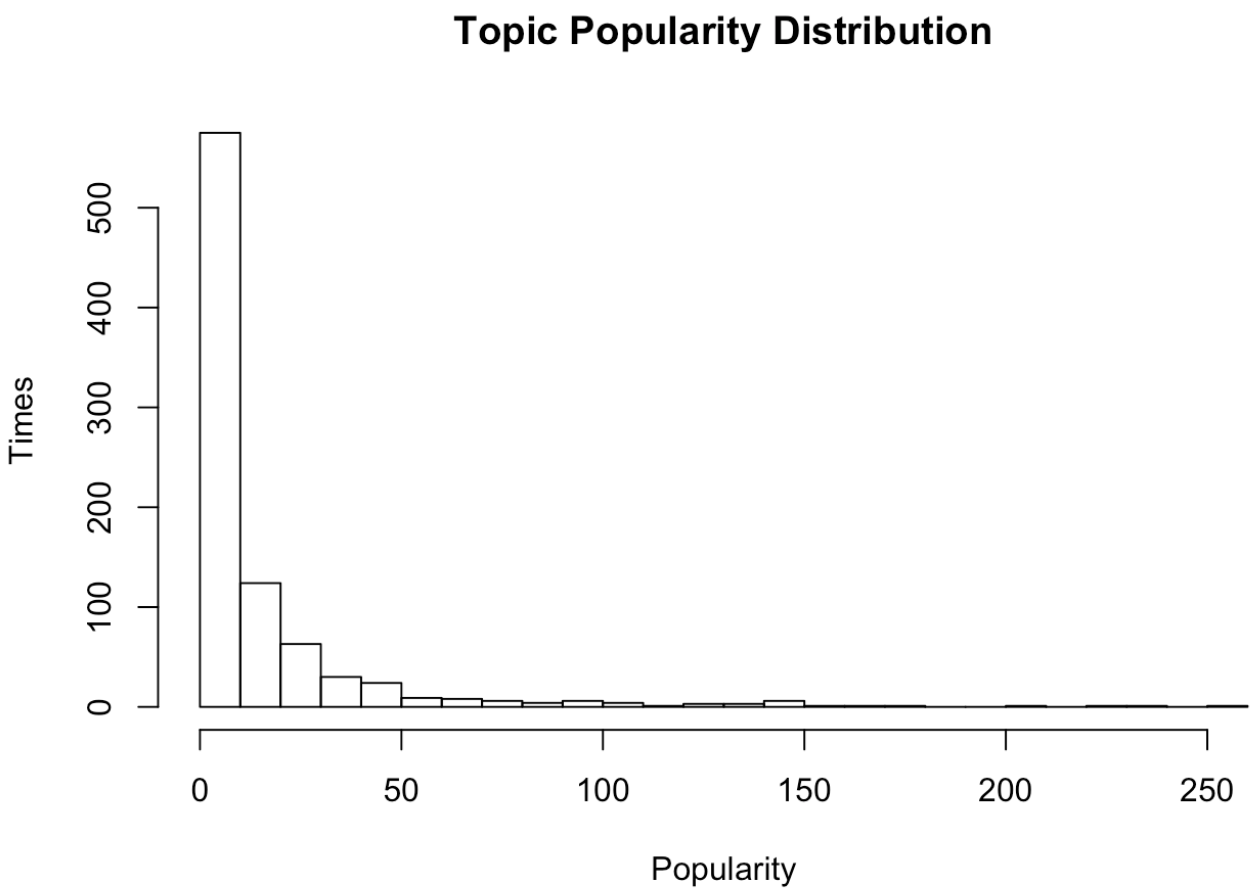
同时保存一个记录每个话题热度的数组$TopicSize$，前10个话题的热度如下：

```
##  [1]  1  3 31  2  4  2 23 11 18  1
```

还保存一个数组$ColdMan$记录每个话题的冷场者，前10个话题的冷场者如下：

```
##  [1] "he"   "yang" "li"   "li"   "he"   "ma"   "ma"   "li"   "ma"   "ma"
```

最后统计得到的话题总数为873，话题热度的分布如下，大部分话题热度不超过50：

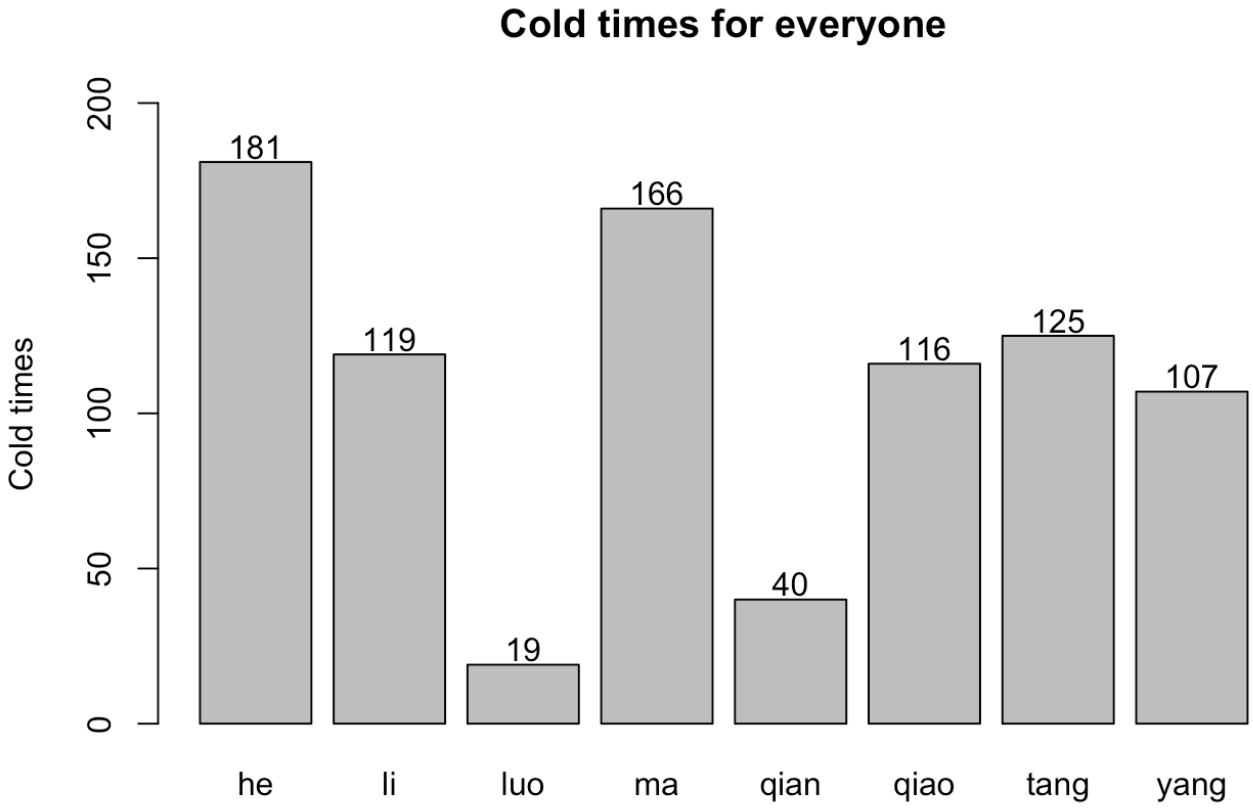


最热的三个话题的热度分别为257, 238, 223，发生时间，冷场时间，终结时间，话题起始者，话题冷场者分别为：

```
## [1] "top 1"
## [1] "Start Time:"
##      year month day hour minute name
## 4375 2015     2  17   19     40 tang
## [1] "End Time:"
##      year month day hour minute name
## 4631 2015     2  17   23     56 qiao
## [1] "------------------------------------------------------"
## [1] "top 2"
## [1] "Start Time:"
##       year month day hour minute name
## 13248 2015    10  29   17     50   li
## [1] "End Time:"
##       year month day hour minute name
## 13485 2015    10  29   19     12 tang
## [1] "------------------------------------------------------"
## [1] "top 3"
## [1] "Start Time:"
##      year month day hour minute name
## 7739 2015     5  19   13      1 tang
## [1] "End Time:"
##      year month day hour minute name
## 7961 2015     5  19   21     13 qiao
## [1] "------------------------------------------------------"
```

按照所示时间返回原文件查找话题内容，最大的话题是由2015年2月17号晚上发红包引起的，终止于Gina的一张晒红包的图片。第二大的话题是由2015年10月29号讨论初中经典瞬间引起的，终止于我的一句“sqq老是穿老马球衣”。第三大的话题起始于2015年5月19号瑞哥商量去腿哥家玩的计划，由一句“发现一个规律。。跟罗导有过故事的女人，都或多或少向娱乐圈靠拢或者已经在其中。。”发展为讨论香港三级片，最后由Gina的一句毫不相关的“第一次尝试鸡蛋仔成功”终结。

每个人的冷场次数如下，可见瑞哥冷场次数最多，但就这样认为瑞哥是冷场王并不合理，因为瑞哥发的消息数和参与的话题数也最多，所以冷场次数多也是情理之中，但并不能说瑞哥很冷场。

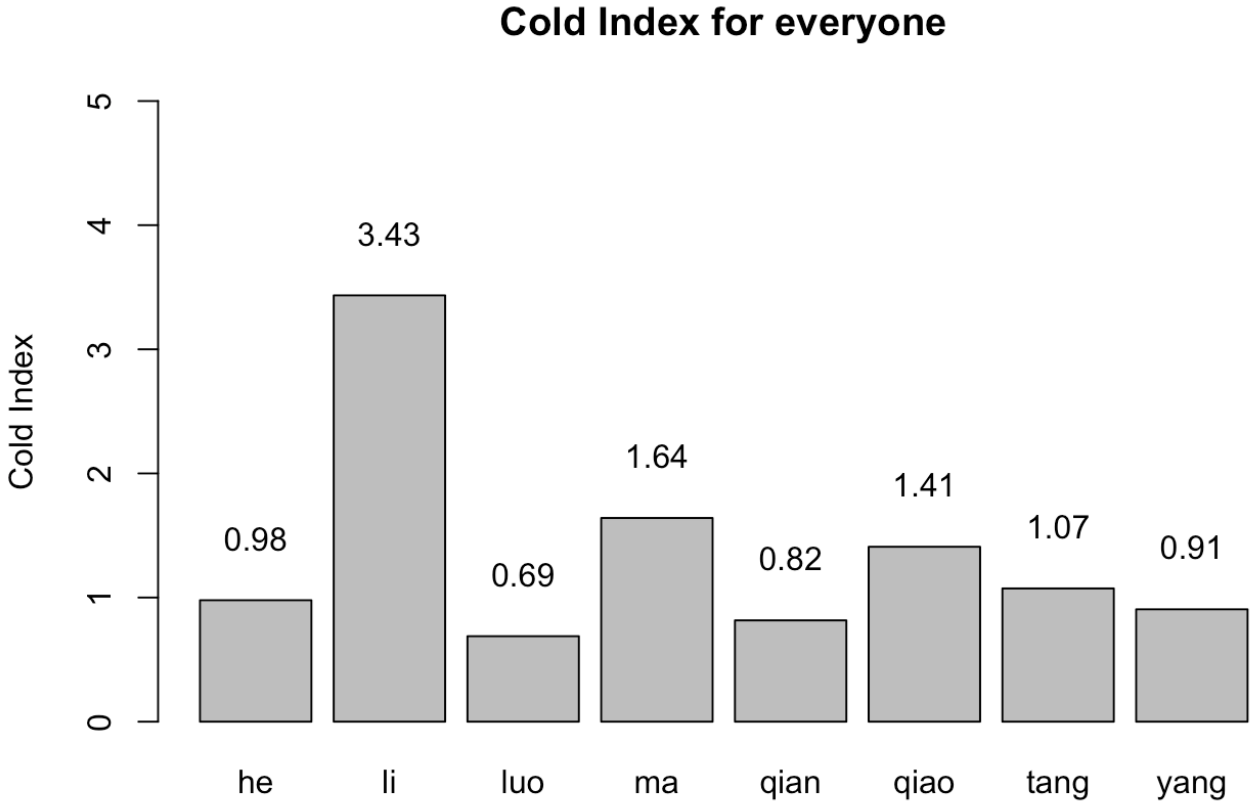


因此我们需要定义一个冷场指数 $ColdIndex$ 来综合考虑每个人的冷场次数，冷场话题的热度(冷场的话题越热，冷场指数越高)，他在话题中的发言数（如果一个人在一个话题中只说了1句话就成功终结了这个话题，那么我只能说你太冷了），以及他参与的总话题数。

$$ColdIndex_i = \frac{\sum_{j=1}^{873} \frac{Popularity_j}{ColdMatrix(i,j)} I_{\{ColdMan(j)=i\}}}{TopicTimes(i)}$$

其中$Popularity_j$表示第j个话题的热度；$ColdMatrix(i,j)$表示第i个人在第j个话题中的发言数；$I_{\{ColdMan(j)=i\}} = 1$如果第j个话题是由i终结的，要不然为0；$TopicTimes(i)$表示i参与话题的总数。

每个人的冷场指数如下，李导一鸣惊人，你猜到了吗？对$ColdIndex$的直观理解：如果某人$ColdIndex$为3表示在他参与的每个话题中，他平均只需要说n句话就能终结一个热度为$3n$的话题。



恭喜李导成功当选为冷场王！！！！

我们不妨再来分析一些别的好玩的东西，既然我们已经有了$ColdMatrix$记录了每个人在每个话题中的发言数，那么我们就可以知道哪些人更倾向于在同一个话题中发言，也就是说两个人参与话题的相似程度。计算$ColdMatrix$中每两行之间的余弦系数($CosineSimilarity$)。得到下表，数字越大表示这两个人在每个话题中的参与度更相似。同时也计算每个人和其他所有人的话题参与度的余弦系数（“others”列）。数字越大表示这个人是群内话题的引领者，数字越小表示这个人喜欢自言自语。

```
##          he     li    luo     ma   qian   qiao   tang   yang others
## he   1.0000 0.3097 0.4079 0.5562 0.3833 0.3520 0.5116 0.5163 0.3435
## li   0.3097 1.0000 0.3671 0.3016 0.4286 0.2022 0.2967 0.4113 0.3399
## luo  0.4079 0.3671 1.0000 0.2970 0.5947 0.4323 0.4867 0.3708 0.4813
## ma   0.5562 0.3016 0.2970 1.0000 0.2653 0.3606 0.5256 0.4447 0.3796
## qian 0.3833 0.4286 0.5947 0.2653 1.0000 0.4634 0.4925 0.3922 0.5010
## qiao 0.3520 0.2022 0.4323 0.3606 0.4634 1.0000 0.5480 0.3399 0.3519
## tang 0.5116 0.2967 0.4867 0.5256 0.4925 0.5480 1.0000 0.4012 0.4896
## yang 0.5163 0.4113 0.3708 0.4447 0.3922 0.3399 0.4012 1.0000 0.4610
```

由该表可以看出话题最相似的几对（瑞哥，小马），（瑞哥，帅哥），（腿哥，太子），（小马，我），（Gina，我）。感觉除了（腿哥，太子）这一对以外都很有道理，但他们这一对我是真冒想明白为什么相似度这么高。

观察“others”列，我们可以发现瑞哥，小马，Gina，李导都低于0.4，属于自言自语爱好者，而太子，我，腿腿和帅哥则是话题的引领者，毕竟咱是Emit乐队正规成员。

所有分析到此结束，纯属娱乐，请勿深究和打我。

本文仅供Emit微信群内部分享交流，请勿外流。