阶段一: 记录与回顾

任务:

- 1. 基于事件扩散网络分析事件的规模、扩散深度、扩散宽度、扩散速度;
- 2. 分析事件的新奇性、事件所激发的用户情感维度和强度、事件随时间扩散曲线的特征、事件的空间特征、事件参与者的特征;

思路:

- 对于数据集进行静态特征分析
- 对结果进行联合实际情况进行分析

0. 数据集

0.1 数据集的列名具体含义

维度	释义
标题/微博内容	标题或转发内容
信息属性	本条微博的信息属性判断，敏感OR非敏感
原创/转发	原创/转发
发布时间	发布时间
原微博内容	原创微博内容
全文	全文
话题	内容中涉及的微博话题
发布省份/备案地	个人信息中填写的省份/非微博网站备案省份
发布城市	个人信息中填写的城市
转发数	本条微博被转发数
评论数	本条微博被评论数
点赞数	本条微博被点赞数
精准地域	微博内容涉及省份或城市
情绪	本条微博的情绪判断
mid（MD5加密）	该微博的id
uid（MD5加密）	发布该微博用户的id
根微博id（MD5加密）	原创微博id
根微博用户id（MD5加密）	发布原创微博用户的id
父微博id（MD5加密）	本条微博转发的上一条微博的id
父微博用户id（MD5加密）	本条微博转发的上一条微博用户的id
认证类型	微博用户认证类型
性别	该微博用户性别属性
粉丝数	该微博用户粉丝数
微博数	该微博用户发布的微博数
图片URL	图片url

0.2 微博数据集

事件名称	文件大小	采集记录(无去重)	采集记录(去重)	采集时长(days/hours/minutes)
海南进入生活垃圾全焚烧时代	375.1KB	247	246	7/22/9
合安高铁进入试运行	1.1MB	1185	1170	6/14/26
广州落户门槛降低	3.4MB	2015	1988	15/15/5
安徽推行机动车检验标志电子化	10.0MB	7525	7472	69/10/25
重庆尾号888888手机号法拍85万	10.3MB	11,986	11,894	5/2/31
安徽歙县内涝严重道路无法通行	18.6MB	17,428	17,258	31/12/45
深圳推行强制休假制度	19.3MB	15,083	14,991	69/7/58
河北一幼儿园食堂现发臭肉馅	22.8MB	17,497	17,403	20/7/21
数字人民币正在雄安新区等地试点测试	27.8MB	19,718	19,549	153/18/57
重庆警方通报城管追打女商贩被砍伤	29.8MB	25,038	24,852	11/3/11
2020世界5G大会开幕	34.0MB	31,158	31,011	4/23/59
《海南自由贸易港建设总体方案》印发	35.1MB	33,625	33,285	222/13/31
上海野生动物园熊伤人致1人死亡	49.8MB	45,485	45,145	13/8/21
长征八号首飞成功	55.1MB	46,894	46,468	21/23/4
丁真回应意外走红	103.6MB	73,824	73,436	52/3/40
唐山大地震44周年	134.6MB	107,461	106,472	31/21/12
男子被浴室玻璃门割伤手	245.4MB	248,063	246,401	7/12/38
上海名媛群争议	302.7MB	284,020	281,693	31/16/31
凉山木里县境内发生森林火灾	378.6MB	303,282	301,743	5/15/31

事件名称	文件大小	采集记录(无去重)	采集记录(去重)	采集时长(days/hours/minutes)
第33届中国电影金鸡奖开幕式	403.8MB	404,671	402,870	3/23/59
腾讯状告老干妈事件-2	589.9MB	543,788	541,759	4/11/58
温岭大溪一油罐车发生爆炸	673.1MB	635,437	632,153	6/6/42
2020年广州国际车展举办	837.2MB	712,213	709,741	14/23/59
腾讯状告老干妈事件-1	870.8MB	798,031	798,031	8/23/58
福建泉州一酒店发生坍塌事故	888.0MB	762,368	755,979	11/22/51

0.3 非微博数据集

事件名称	文件大小	采集记录	参与媒体数目
合安高铁进入试运行	1.0MB	266	67
重庆尾号888888手机号法拍85万	7.0MB	2,704	342
海南进入生活垃圾全焚烧时代	8.2MB	1,973	308
河北一幼儿园食堂现发臭肉馅	19.5MB	9,198	595
重庆警方通报城管追打女商贩被砍伤	24.6MB	13,948	689
男子被浴室玻璃门割伤手	29.5MB	11,364	512
第33届中国电影金鸡奖开幕式	33.1MB	15,386	664
唐山大地震44周年	33.9MB	14,644	1,077
上海野生动物园熊伤人致1人死亡	49.4MB	19,362	860
深圳推行强制休假制度	77.5MB	20,057	1,076
2020世界5G大会开幕	87.3MB	29,485	1,802
广州落户门槛降低	94.7MB	18,370	1,257
丁真回应意外走红	95.8MB	25,738	845
安徽歙县内涝严重道路无法通行	119.0MB	32,341	1,927
长征八号首飞成功	179.6MB	53,984	2,615
凉山木里县境内发生森林火灾	193.5MB	87,842	3,642
上海名媛群引争议	236.3MB	76,551	1,358
数字人民币正在雄安新区等地试点测试	335.3MB	75,067	3,040
温岭大溪一油罐车发生爆炸	335.3MB	152,881	4,658
安徽推行机动车检验标志电子化	349.2MB	89,037	2,826
2020年广州国际车展举办-1	474.0MB	196,355	1,666
腾讯状告老干妈事件	476.6MB	186,008	None
2020年广州国际车展举办-2	549.3MB	203,166	1,748
福建泉州一酒店发生坍塌事故	599.9MB	278,470	5,200
2020年广州国际车展举办-3	619.2MB	203,961	2,230
《海南自由贸易港建设总体方案》印发	1.1GB	172,590	6,202

1. 任务一

1.1 构建事件扩散网络

- 思路:

1. 关注数据集中`MD5-mid`和`MD5-父微博ID`, 由`MD5-父微博ID` --> `MD5-mid`作为网络中的一条有向边. 重复边算作一条边. 如无父节点, 就将其作为一个节点加入(它有可能是别的微博的父节点).
2. 边有一个转发时间的属性.
3. 保存网络格式为邻接表格式(*.adjlist).

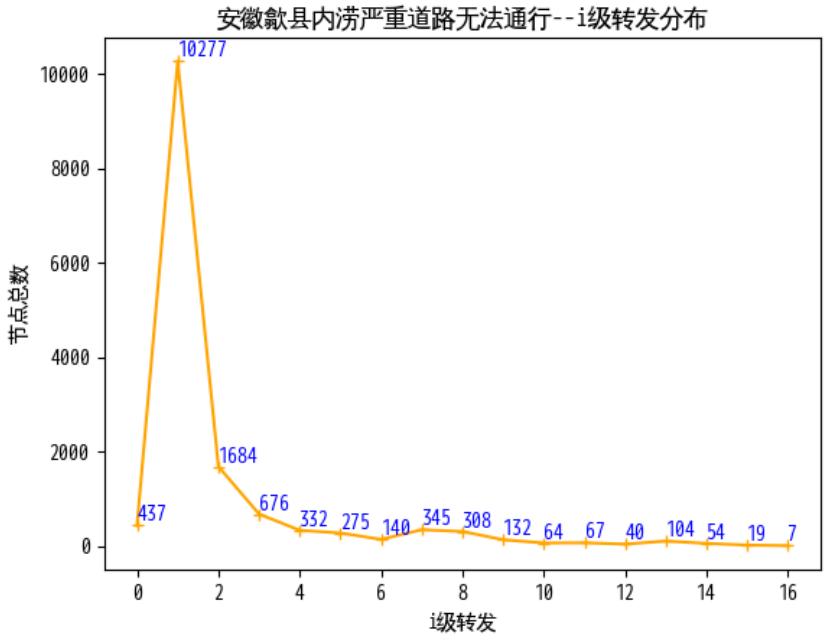
- 结果:(对与网络中第i级转发的节点数, 保存在Analysis_of_Network.txt, 扩散的深度和广度在此处有所体现)

事件名称	网络节点数/边数	孤立节点个数	传播深度
海南进入生活垃圾全焚烧时代	246/28	205	3
合安高铁进入试运行	1,174/786	320	5
广州落户门槛降低	1,995/1,492	374	7
安徽推行机动车检验标志电子化	7,491/5,097	1996	6
重庆尾号888888手机号法拍85万	12,086/6,922	4692	6
安徽歙县内涝严重道路无法通行	17,378/14,524	2417	17
深圳推行强制休假制度	15,022/11,741	2847	11
河北一幼儿园食堂现发臭肉馅	17,413/15,663	1300	13
数字人民币正在雄安新区等地试点测试	19,703/14,557	4194	15
重庆警方通报城管追打女商贩被砍伤	24,922/20,280	3854	12
2020世界5G大会开幕	31,079/28,097	2400	6
《海南自由贸易港建设总体方案》印发	33,432/16,474	15786	9
上海野生动物园熊伤人致1人死亡	45,292/35,083	8888	23
长征八号首飞成功	46,597/37,908	7188	14
丁真回应意外走红	73,607/68,443	4149	10
唐山大地震44周年	106,610/97,389	7327	17
男子被浴室玻璃门割伤手	246,809/209,507	35397	20
上海名媛群争议	283,220/239,635	39007	31
凉山木里县境内发生森林火灾	302,453/285,440	12463	37
第33届中国电影金鸡奖开幕式	403,767/389,769	10674	9
腾讯状告老干妈事件-2	570,781/487,970	49630	10
温岭大溪一油罐车发生爆炸	632,733/581,397	39536	25
2020年广州国际车展举办	710,970/659,887	42904	18
腾讯状告老干妈事件-1	823,467/719,246	69232	17
福建泉州一酒店发生坍塌事故	757,468/665,293	71379	27

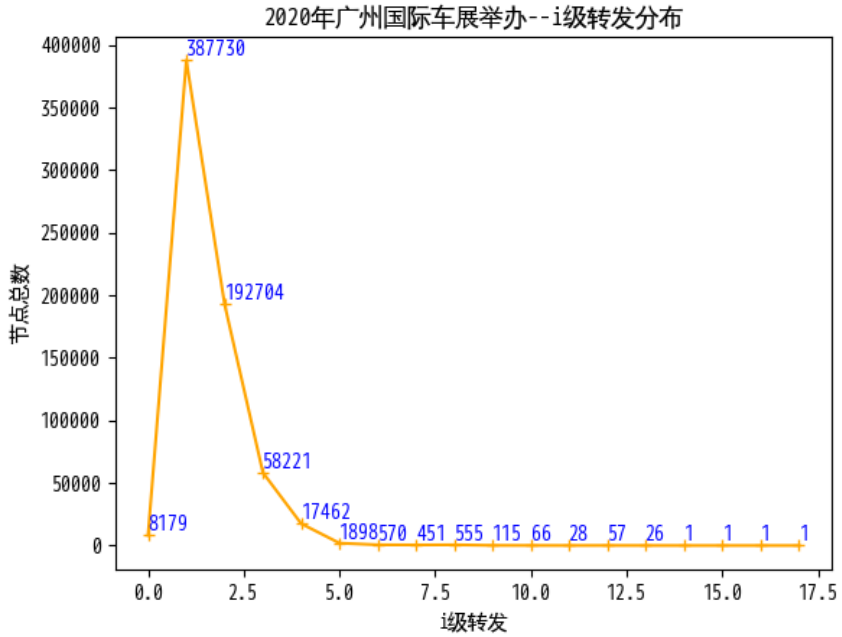
- 分析: 事件规模可以用节点和边的数量来衡量. 对于事件扩散网络, 我们要去除孤立节点, 其实也就有了一棵树, 只是树的有多个根. 节点数一般要比边数多一点, 不存在环.

1.2 扩散深度和扩散广度

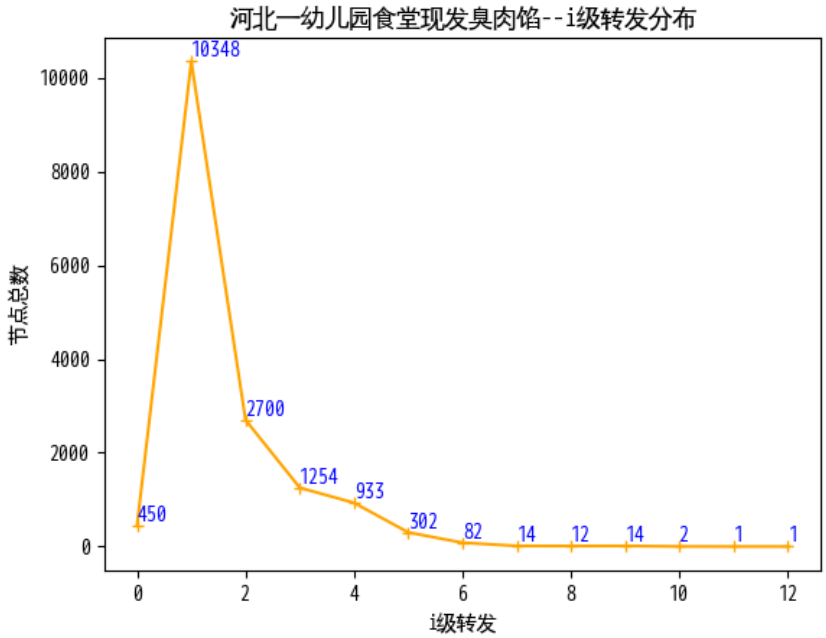
- 思路: 利用树的深度和每层节点数来代表扩散深度和扩散广度
- 结果: 三个事件的扩散深度(n级转发)和广度(第i级转发的节点数目).
    - 安徽歙县内涝严重道路无法通行



    - 2020年广州国际车展举办



    - 河北一幼儿园食堂现发臭肉馅

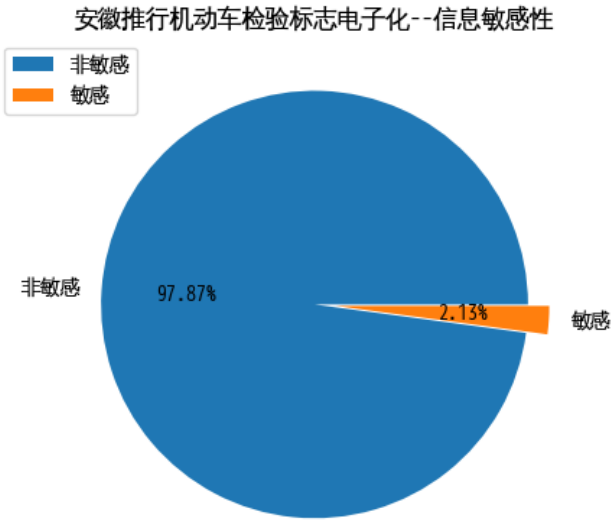


- 分析: 上面三个事件的传播的广度和深度就体现在图中, 可以得出以下结论:
  1. 一级转发节点占据最多, 在不同事件中,占比基本保持一致在`55%-70%`之间.
  2. 5级转发之前的节点占据网络节点总数`90%`左右, 也就是任一事件扩散规模主要是由前5级转发节点构成的
  3. 不忽略任意一个转发的话, 每个事件的传播深度是不一样的.

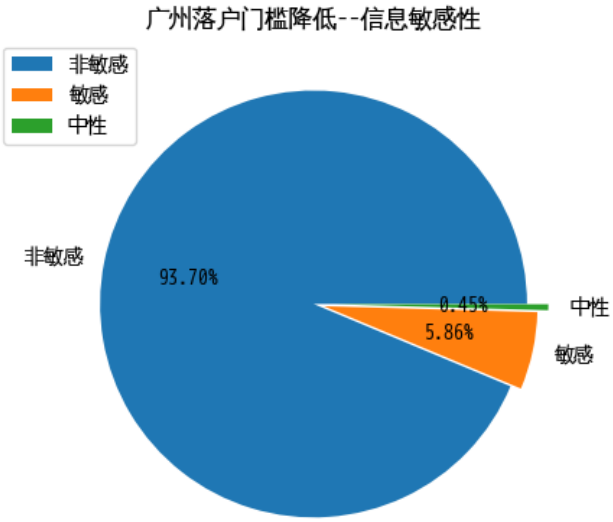
2. 任务二

2.1 分析事件的新奇性

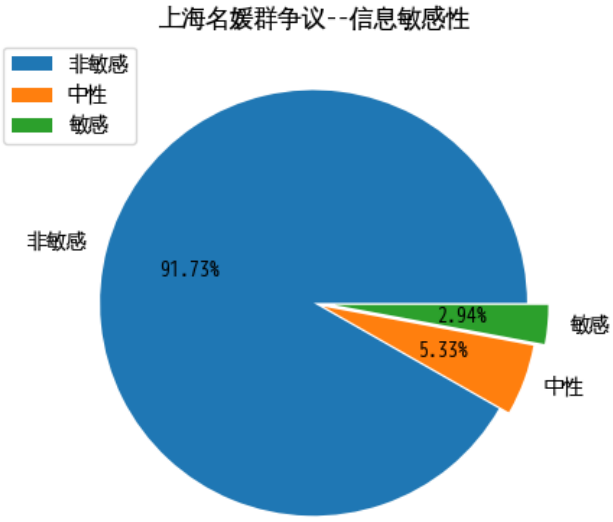
- 思路: 该特征暂时用数据集中的`信息属性`代替
- 结果: 列举三个事件
  - 安徽推行机动车检验标志电子化



  - 广州落户门槛降低



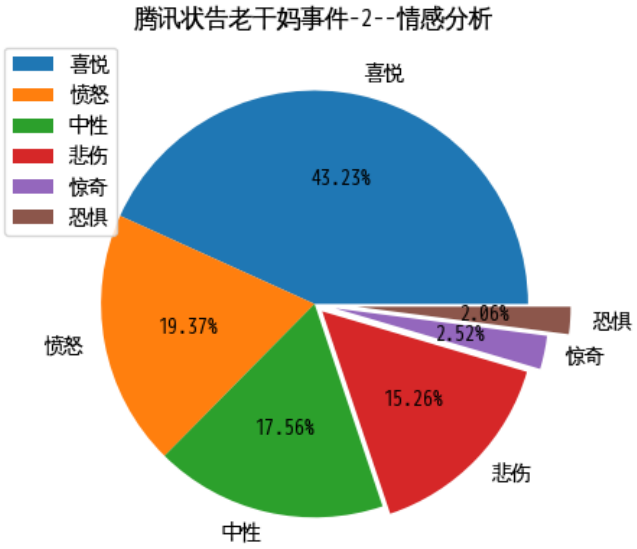
- 上海名媛争议



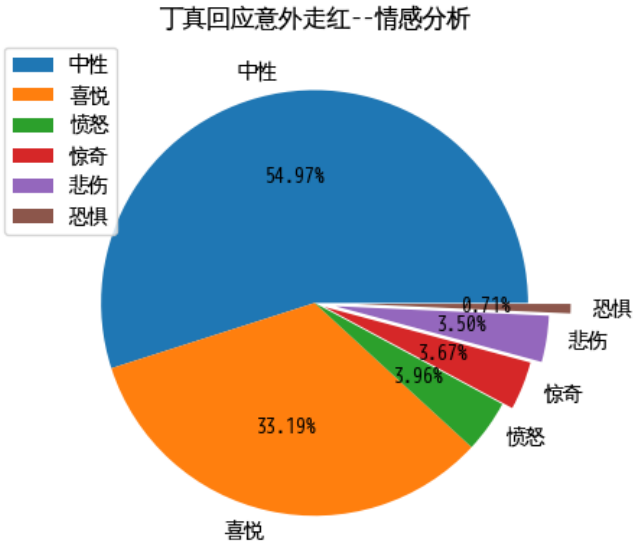
- 分析
  1. 事件的非敏感性质占比普遍在90%以上, 说明诸多信息对大众来说并不是很敏感.
  2. 中性和敏感对于有些事件来说, 并不能很好确定, 但是两个占比较小.

2.2 事件所激发的用户情感维度和强度

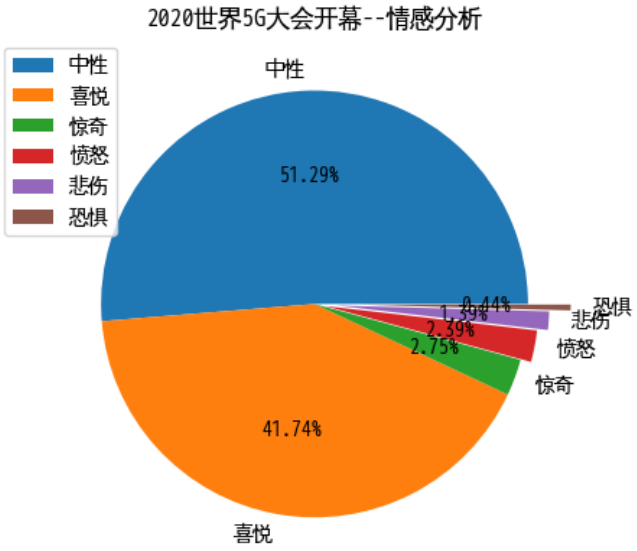
- 思路: 用`微博情绪`来表示用户的情感(微博已标注, 是否准确?)
- 结果: 依旧列举三个事件
  - 腾讯状告老干妈事件-2



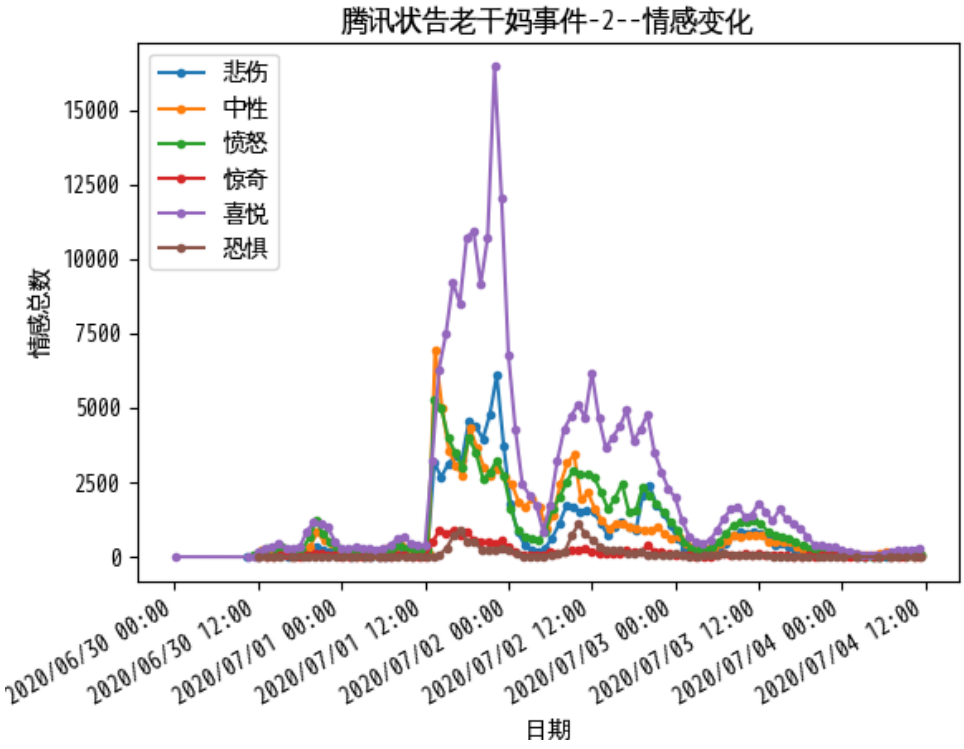
- 丁真回应意外走红



- 2020世界5G大会开幕



- 分析:
  1. 微博用户对待事件的情绪, 大部分还是中性.
  2. 根据事件的语义上带来的情感不同会有一定情感倾向, `腾讯状告老干妈事件`更多人是喜悦, 愤怒或者悲伤. 情感倾向也是一个动态的过程, 该事件在发展过程中会有一些新内容出现导致情感走势不一样. 如下图,是`腾讯状告老干妈事件`的情感走势.

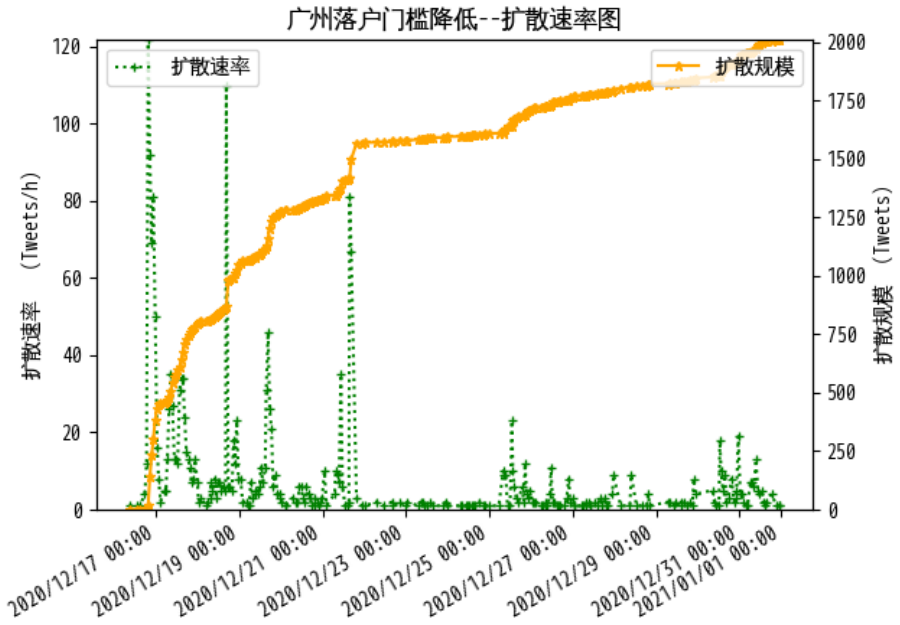


可以看到用户的情感变化主要有三个波峰, 逐次减小

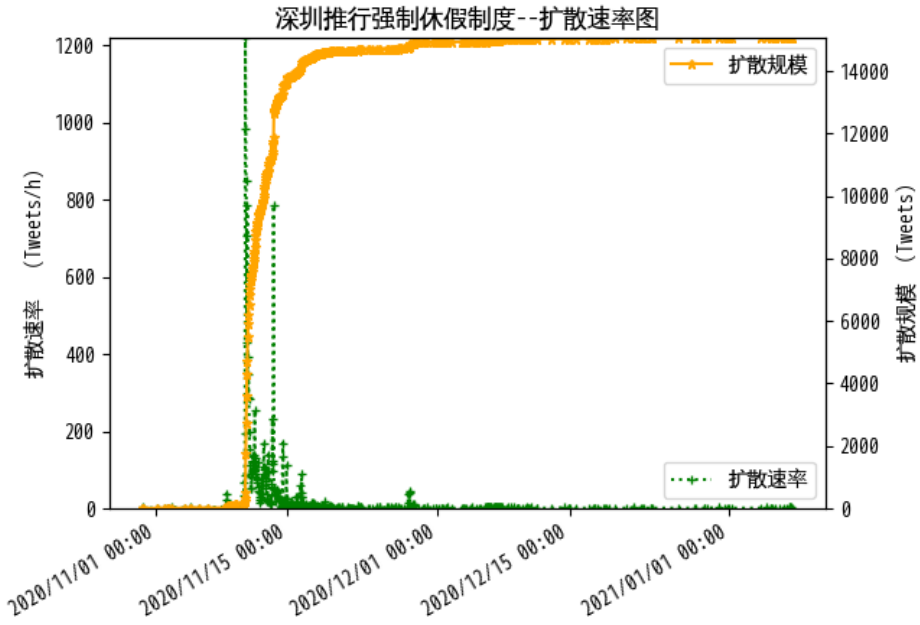
- 2020年7月1日12:00左右, `老干妈回应腾讯起诉老干妈`引起用户情感大幅度激增. 这段时间内的情感占比为喜悦>中性>愤怒>悲伤等, 也可以看出来这段时间, 情绪是有一个愤怒和中性向悲伤演进的.
- 2020年7月2日8:00左右, `老干妈称腾讯从来没有催收过`因此用户情感中等幅度激增.
- 2020年7月3日9:00左右, `老干妈上架1000瓶辣椒酱回应腾讯`以及`字节跳动副总裁吐槽腾讯`等两个时事件引起用户情感小幅度激增

2.3 事件随时间扩散曲线的特征

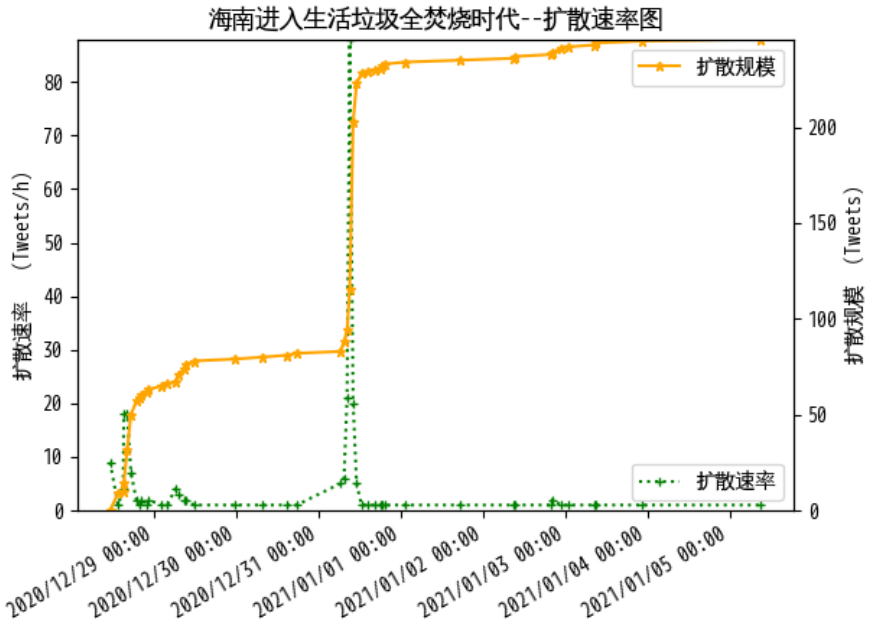
- 思路: 按照每小时产生的微博数目作为速率, 对速率进行积分可以分析某时刻的传播规模.
- 结果: 三个事件为例
  - 广州落户门槛降低



  - 深圳推行强制休假制度



  - 海南进入生活垃圾全焚烧时代

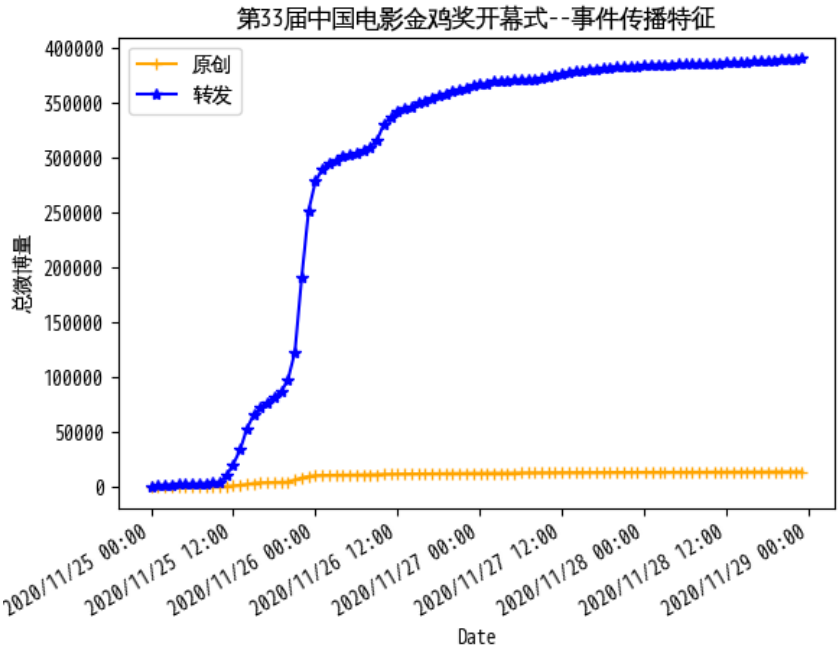


- 分析:
  1. 绿线代表扩散速率, 橙线代表扩散规模, 两条线都是以每个小时为单位进行统计.
  2. 橙线和绿线是"求导"关系
  3. 以`广州落户门槛降低事件`为例, 观察其扩散速率(绿线)有四个较高波峰, 分别可以对应该事件中四个小事件
     1. 2020年12月16日19:00左右`财经网发布【史上最宽松，大专学历即可落户这座一线城市！#广州拟出台差别化入户政策#】`权威意见领袖引领这次传播扩散
     2. 2020年12月18日16:30左右`@广州榜哥 #广州# 拟降低七区落户门槛！白云、黄埔、花都、番禺、南沙、增城、从化，只要大专，社保满一年，28岁以下就能落户！广州放松人才落户政策，对于楼市而言无疑是一大利好.`这个应该是一个水账号, 很短的时间内激增.
     3. 2020年12月19日15:38左右`@央视财经 【#广州拟放宽落户条件#：#28岁以内大专生可落户广州#】昨天，广州市就即将出台的落户政策公开征求意见.....`
     4. 2020年12月21日15:30左右, `@广州校园君 【#广州拟放宽落户条件#：#28岁以内大专生可落户广州#】` `小葳学长 【#广州拟放宽落户条件#：#28岁以内大专生可落户广州#】` `@太原校园君 【#广州拟放宽落户条件#：#28岁以内大专生可落户广州#】` `@重庆大学城科微校园 【#广州拟放宽落户条件#：#28岁以内大专生可落户广州#】`等好几个事件, 这是在高校微博之间进行传播.
  4. 有些事件中的时间扩散速率可能比较单一, 比如上例中的`海南进入生活垃圾全焚烧时代`事件, 也就两次小事件推动整个事件的传播.
  5. 有些事件传播周期长, 具有阶段性传播的特性, 如`广州落户门槛降低事件`. 而有些事件传播周期短, 很短的时间内就达到了扩散的最大规模,如`深圳推行强制休假制度`

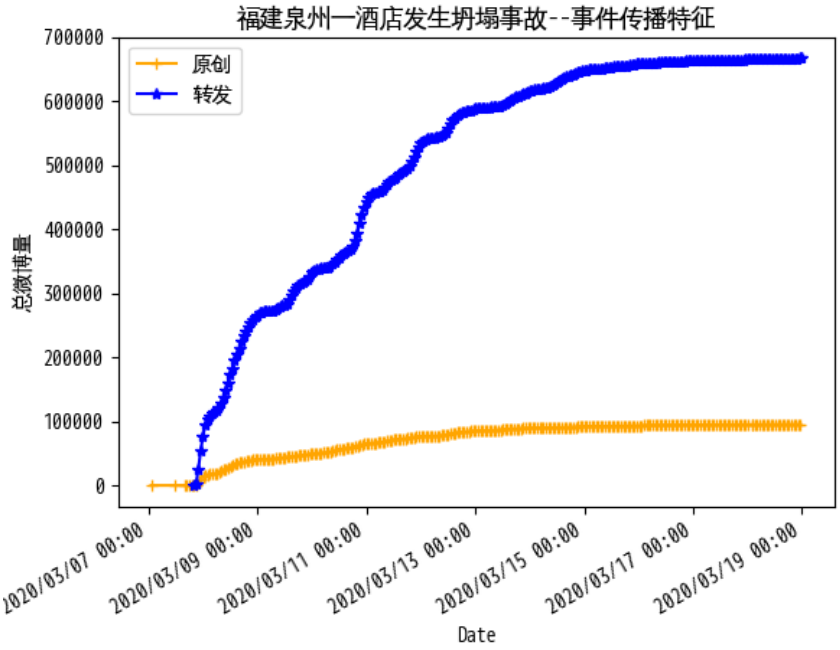
2.4 事件的空间特征

2.4.1 事件传播线(分转发和原创)

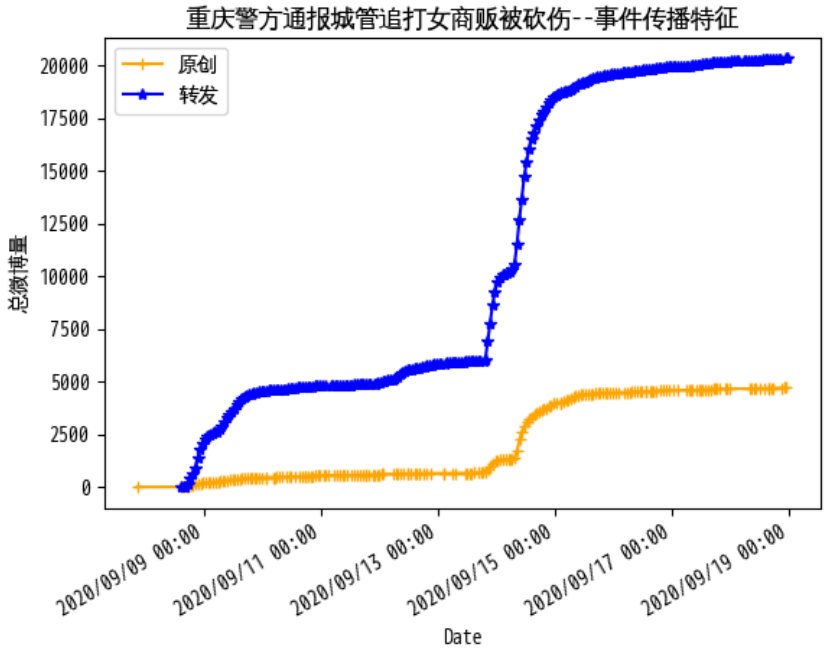
- 思路: 跟2.3类似, 只不过要分开转发和原创, 暂只考虑规模
- 结果: (三个事件为例)
  - 第33届中国电影金鸡奖开幕式



- 福建泉州一酒店发生坍塌事故



- 重庆警方通报城管追打女商贩被砍伤

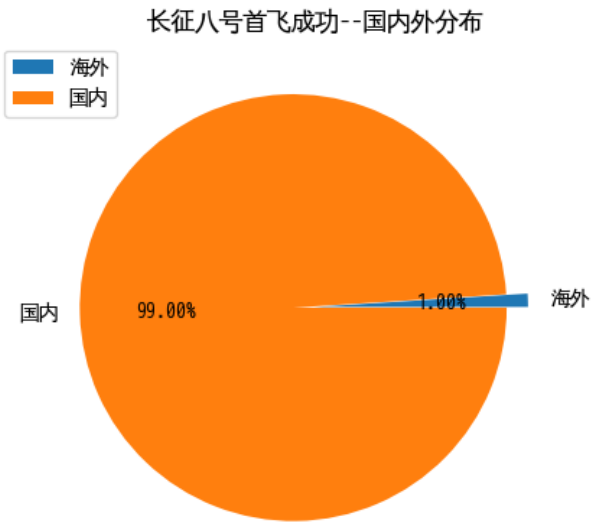


- 分析:

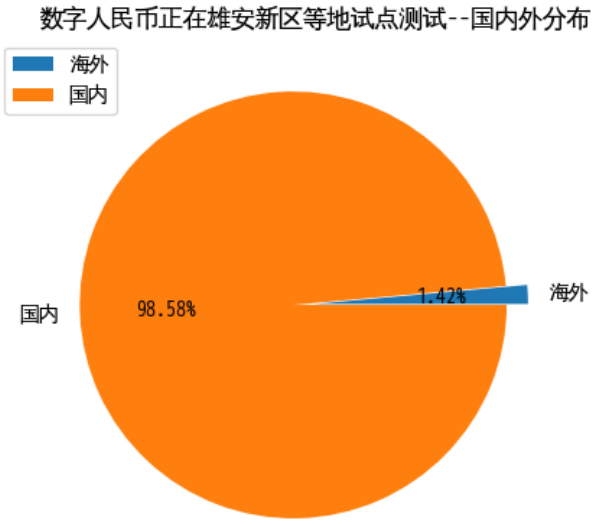
    1. 事件的传播方式主要以转发为主, 这也与微博这类社交网络传播特点有关, 意见领袖发表意见(比如粉丝数多, 微博认证达人, 权威机构等) 多数微博用户只是担任了转发载体的角色.
    2. 从最后一个图中, 即 `重庆警方通报城管追打女商贩被砍伤` 事件在2020年9月13日19:00左右, 有一些原创微博激增, 随后十分钟后, 转发微博开始激增. 这也符合一些常规认知.

2.4.2 事件国内外分布

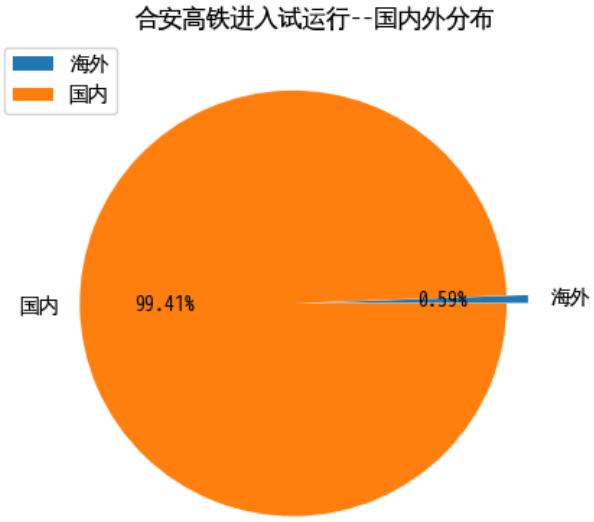
- 思路: 统计 `地域` , 区分国内外
- 结果: 三个事件为例
    - 长征八号首飞成功



    - 数字人民币正在雄安新区等地试点测试



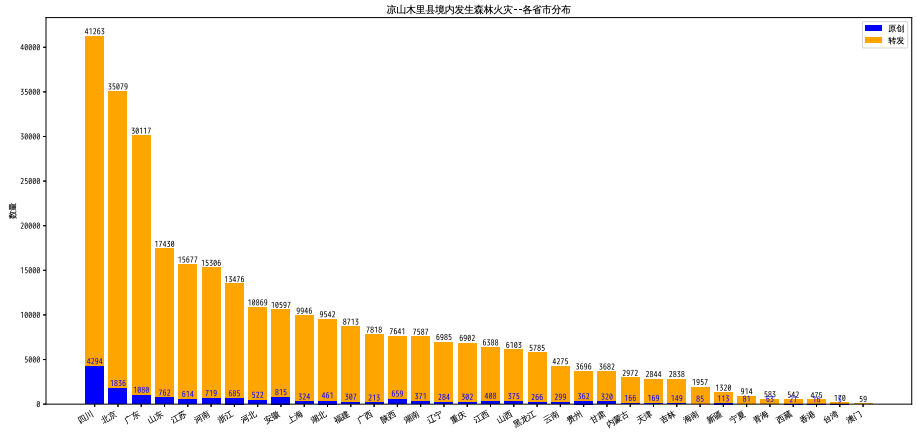
- 合安高铁进入试运行



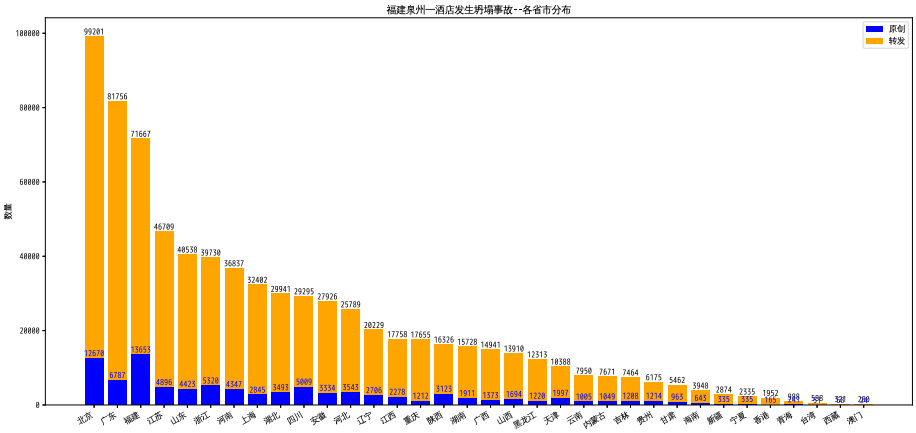
- 分析:
  1. 微博作为一个国产社交平台, 国内账户使用占据98%以上, 根据事件不同,国外用户占比也不尽相同
  2. 国外用户大部分还是一些中国用户,只是地域分布在国外.

2.4.3 事件国内各省市分布

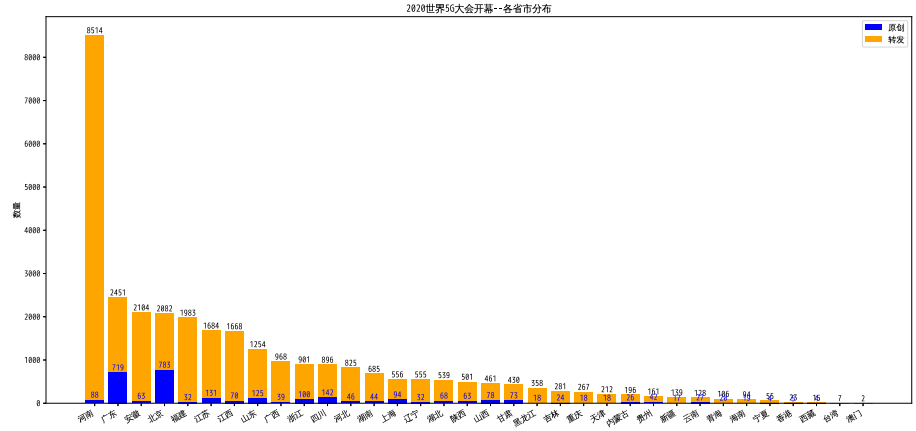
- 思路: 统计`地域`和`原创和转发`两列来统计出国内各省市分布情况
- 结果: 三个事件为例
  - 各省市分布_凉山木里县境内发生森林火灾



- 各省市分布_福建泉州一酒店发生坍塌事故



- 各省市分布_2020世界5G大会开幕



- 分析:

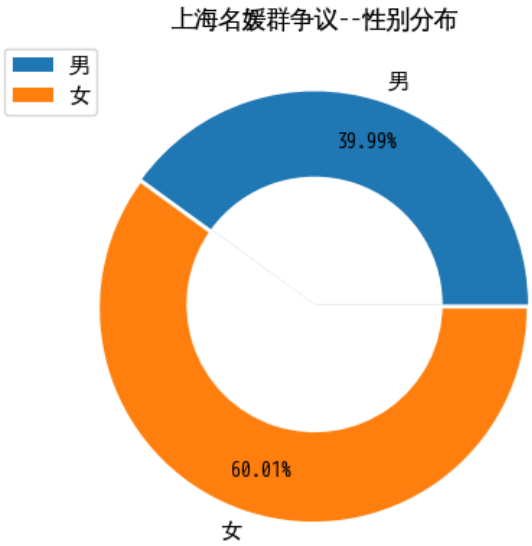
1. 多数情况下, 事件的起源地或者关联地是微博传播的主要区域, 事件区域化比较显著, 比如事件`凉山木里县境内发生森林火灾`主要有四川省的微博用户来进行原创和转发.
2. 此外, 跟一些城市的发达程度也是有关的, 比如北京微博用户对诸多事件中传播做出重要贡献. 一些西部地区,如新疆西藏等相对而言人数较少, 微博相对不活跃.

3. 很多事件的各省市分布的前几名基本都是北上广, 江苏, 浙江等区域.

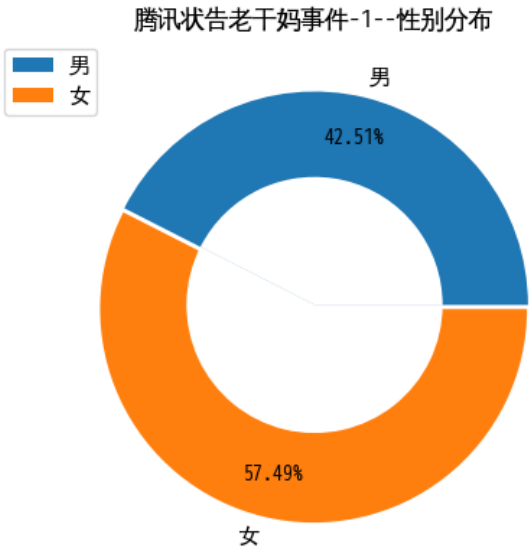
2.5 事件参与者的特征

2.5.1 用户性别分布

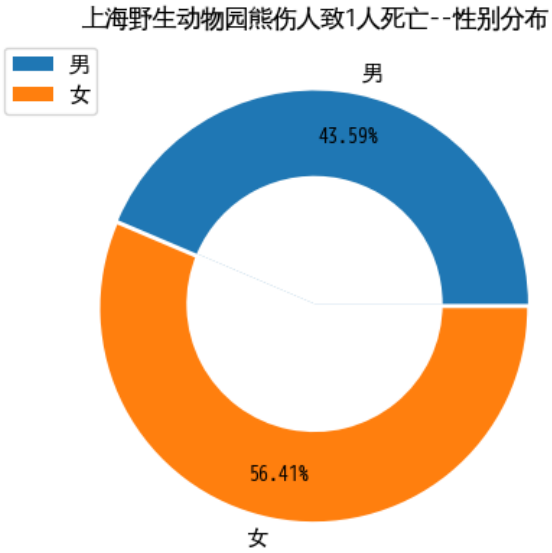
- 思路: 统计用户的性别分布(暂时没有去重)
- 结果:
  - 用户性别分布_上海名媛争议



  - 用户性别分布_腾讯状告老干妈事件-1



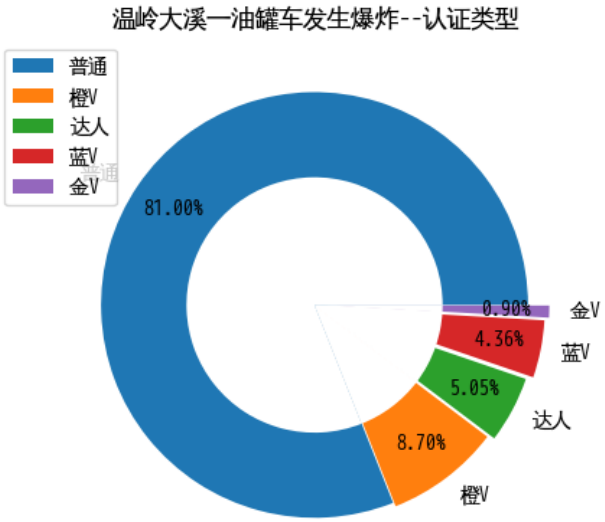
  - 用户性别分布上海野生动物园熊伤人致1人死亡



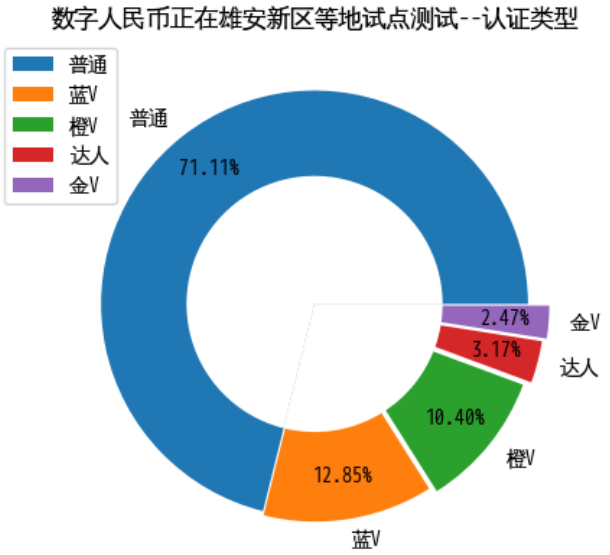
- 分析:
    1. 微博用户在大多数事件中都是女性占比高于男性.

2.5.1 认证类型分布

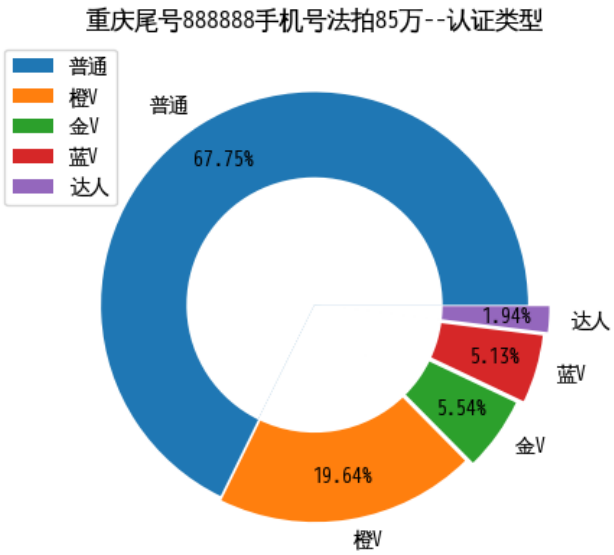
- 思路: 统计`认证类型`列, 统计划分出来不同的认证类型
- 结果: 三个事件
    - 认证类型分布_温岭大溪一油罐车发生爆炸



    - 认证类型分布_数字人民币正在雄安新区等地试点测试



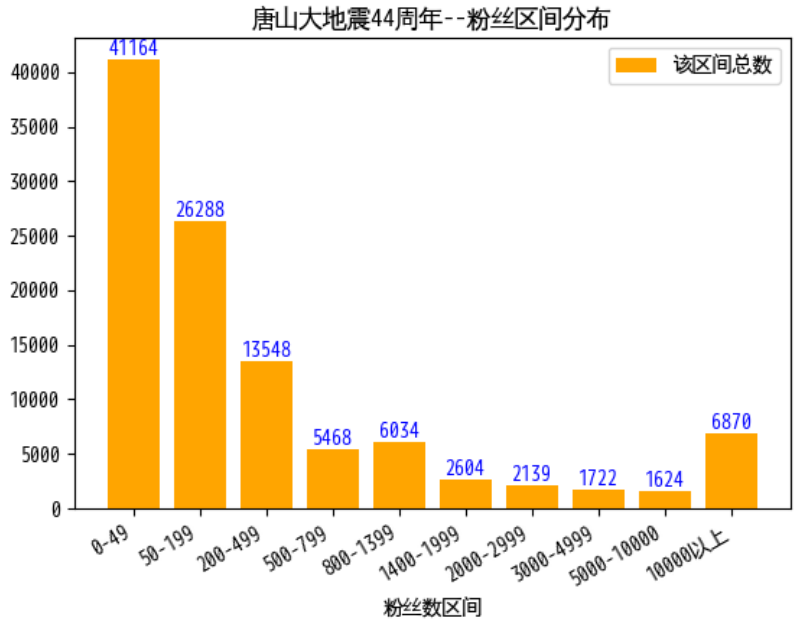
- 认证类型分布_重庆尾号888888手机号法拍85万



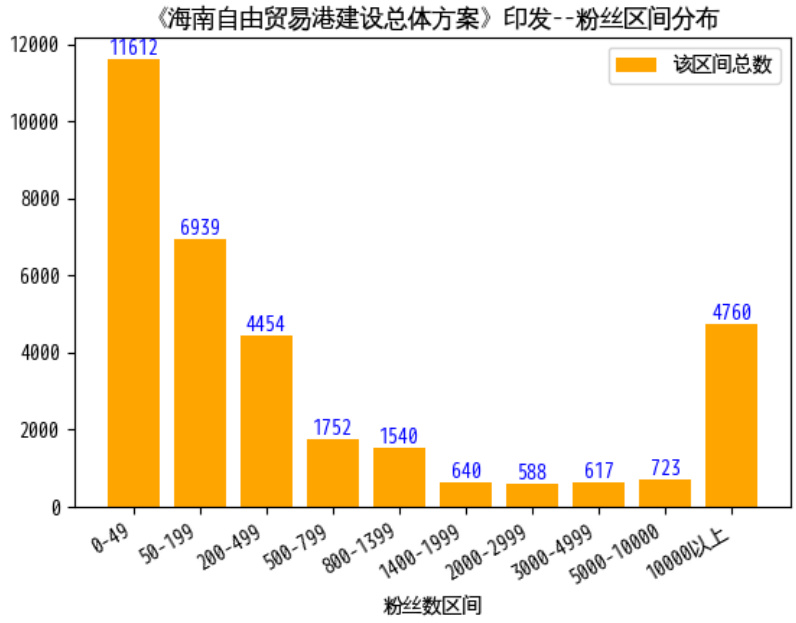
- 分析:
    1. 事件的主要参与者是普通用户, 一般占比60%以上, 其次是橙V,蓝V. 达人和金V相对占比较少.

2.5.1 粉丝数量区间分布

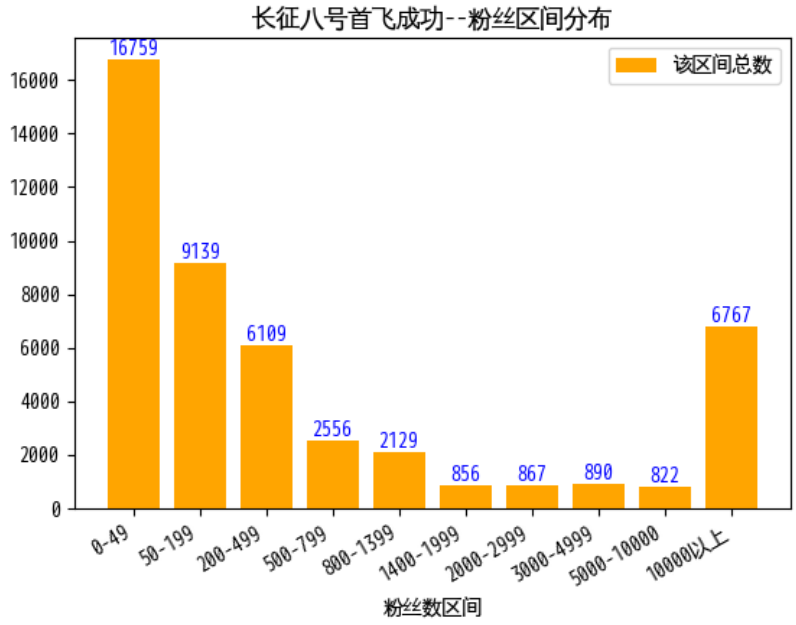
- 思路: 统计`粉丝数`, 做区间划分和统计
- 结果: 三个事件分析
    - 粉丝数量区间分布_唐山大地震44周年



- 粉丝数量区间分布_《海南自由贸易港建设总体方案》印发



- 粉丝数量区间分布_长征八号首飞成功



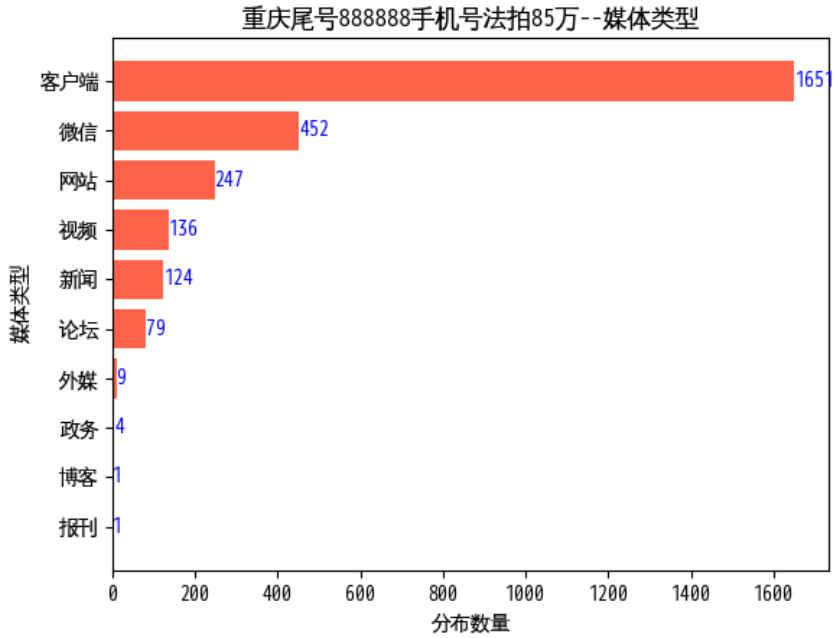
- 分析:
    1. 粉丝分布区间基本较为相似, 多数用户的粉丝数量都在500以内, 50个粉丝以上的占据最多数目.
    2. 很多意见领袖都粉丝数都是百万或者千万级别的. 很多事件走向都是由于这些意见领袖主导的,因此粉丝数1w以上的用户也占有5-10%左右的比例.

2.6 非微博数据简要分析

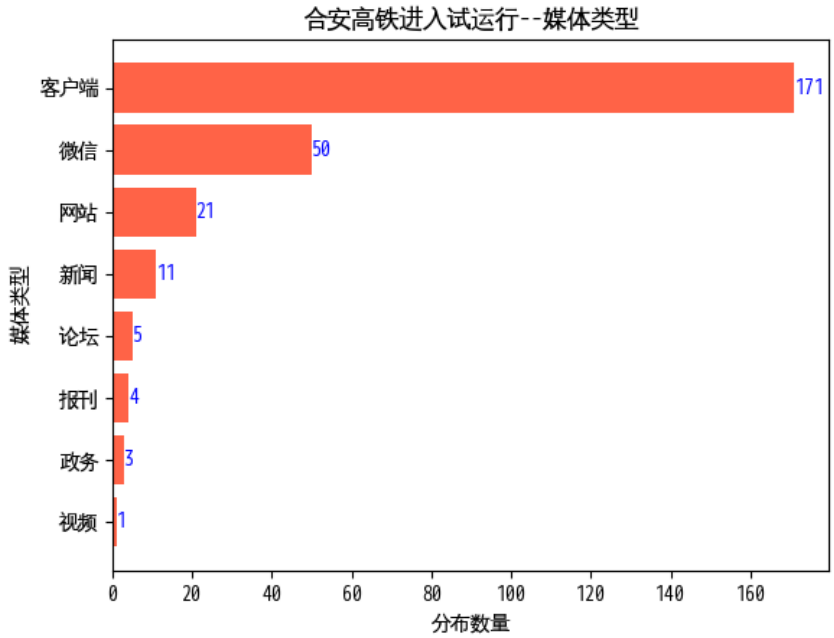
由于非微博数据也包含地域分布, 转发/原创, 发布日期等条目, 因此可以用分析微博数据的代码模块来实现, 在这里仅仅分析一下, 某事件在非微博平台的分布情况以及载体媒介.

2.6.1 媒体类型

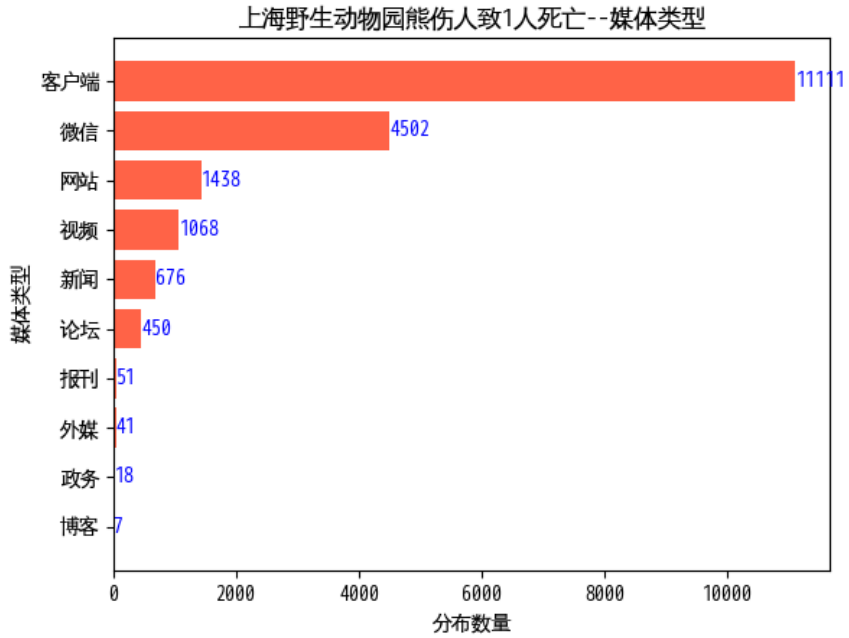
- 思路: 统计媒体类型一列
- 结果: 三个非微博数据事件
    - 媒体名称分布_重庆尾号888888手机号法拍85万



- 媒体名称分布_合安高铁进入试运行



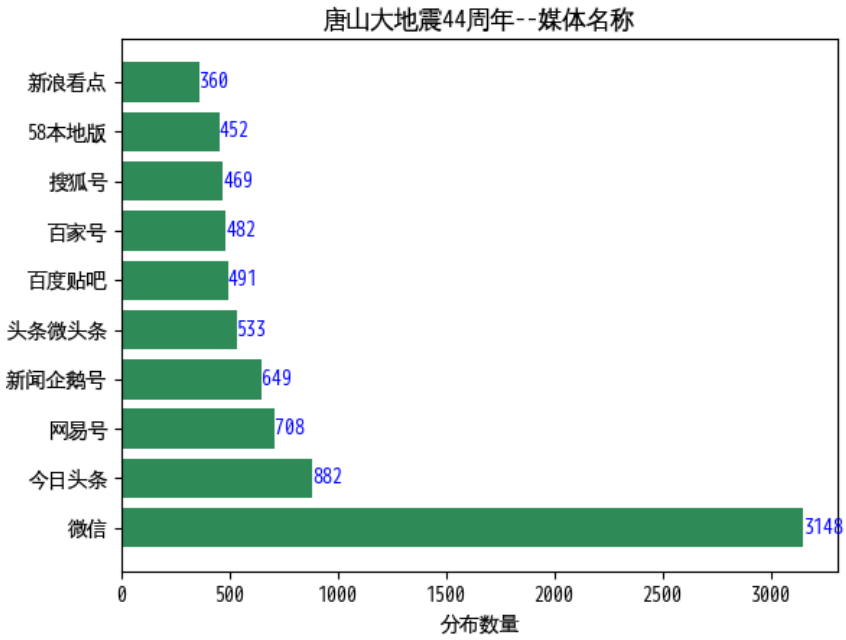
- 媒体名称分布_上海野生动物园熊伤人致1人死亡



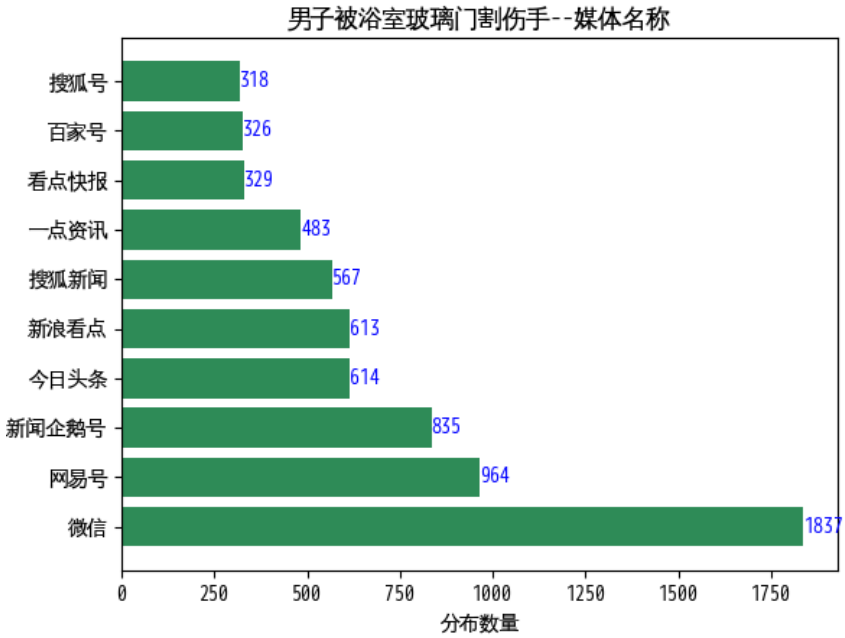
- 分析:
    1. 除微博外的主流媒体平台是客户端, 微信, 网站, 视频
    2. 其他类型偏少.

2.6.2 媒体名称

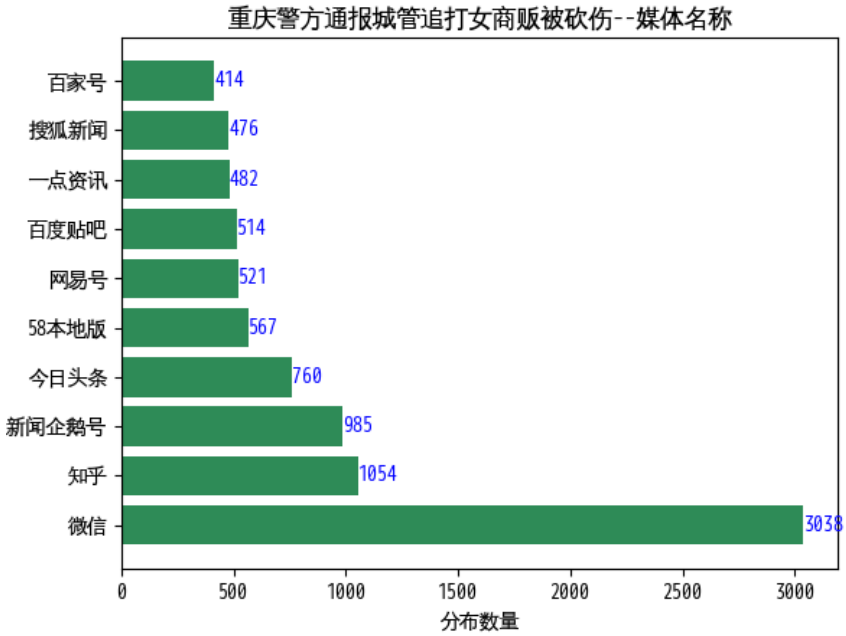
- 思路: 同上统计具体的媒体名称
- 结果:
    - 媒体名称分布_唐山大地震44周年



    - 媒体名称分布_男子被浴室玻璃门割伤手



- 媒体名称分布_重庆警方通报城管追打女商贩被砍伤



- 分析:
  1. 微信作为即时通信软件,很好的成为了非微博平台事件传播的最大载体.
  2. 此外, 一些常见的论坛, 如知乎,今日头条, 新闻企鹅号, 百度贴吧等都成为非微博平台事件传播的主要载体.