**百度贴吧平台的数据挖掘与分析**

10215501412 彭一珅

华东师范大学 数据科学与工程学院

目录

[一、 实验引言 3](#_Toc123593839)

[1.1 选题背景 3](#_Toc123593840)

[1.2 研究背景 3](#_Toc123593841)

[二、 问题描述 3](#_Toc123593842)

[2.1 研究目的 3](#_Toc123593843)

[2.2 研究工具 3](#_Toc123593844)

[2.3 研究难点 3](#_Toc123593845)

[三、 实验方法 4](#_Toc123593846)

[3.1 数据获取 4](#_Toc123593847)

[3.2 数据预处理 6](#_Toc123593848)

[3.3 探索性分析 7](#_Toc123593849)

[3.3.1 不同省份人数占比 7](#_Toc123593850)

[3.3.2 各省用户发言活跃度统计 9](#_Toc123593851)

[3.3.3 各吧删帖率对比 10](#_Toc123593852)

[3.3.4 用户性别比例 12](#_Toc123593853)

[3.3.5 吧龄分布 13](#_Toc123593854)

[3.3.6 楼主发言比例 15](#_Toc123593855)

[3.3.7 等级与发言数关系 16](#_Toc123593856)

[3.3.8 等级分布 18](#_Toc123593857)

[3.3.9 主题帖发帖日期分布 20](#_Toc123593858)

[3.3.10 各吧发言活跃时段 21](#_Toc123593859)

[3.3.11 常用表情统计 23](#_Toc123593860)

[3.3.12 常用词汇统计 24](#_Toc123593861)

[3.3.13 单个帖子发言数随时间变化趋势 25](#_Toc123593862)

[3.3.14 吧之间的关联情况 27](#_Toc123593863)

[3.3.15 用户指标热力图分析 29](#_Toc123593864)

[3.4 数据建模 30](#_Toc123593865)

[四、 实验评价 33](#_Toc123593866)

[五、 后续工作 34](#_Toc123593867)

[六、 实验结论 34](#_Toc123593868)

# 一、 实验引言

1.1 选题背景

百度贴吧是建立于2003年的中文社区，根据不同话题形成了不同分区，它为拥有共同爱好的网友搭建了交流经验、共享资源的平台，受到广大网友的青睐。截至今天，百度贴吧已经建立了超过2000万个不同的吧，最多人关注的李毅吧，关注人数达到了3000万以上。

基于如此巨大的用户量和流量，我想要通过数据科学过程，对百度贴吧几个吧的用户组成、发帖内容、发展趋势等进行初步的研究。

1.2 研究背景

在2018年前，百度贴吧曾是盛极一时的网络论坛，但是根据我的粗略观察，近年来，百度贴吧的用户量、发言量等似乎出现了衰退的趋势，对于百度贴吧数据的采集和分析，有助于我对其发展现状和趋势产生更深刻的理解，从而对百度贴吧及国内相关论坛平台的发展状况做出判断，并提出建议和展望。

# 二、 问题描述

2.1 研究目的

爬取百度贴吧首页展示的几个发帖量与关注量较大的吧的数据，并进行数据挖掘，对用户数据、发帖数据等作可视化分析，并将相关数据汇总，建立情感分析模型，对百度贴吧中的内容进行全面分析。

通过上述实验过程，对百度贴吧的现状作出评价，对该论坛的发展趋势作出展望。

2.2 研究工具

在Windows环境下，使用python 3.10，运用TensorFlow、pandas、selenium等库辅助进行数据收集和探索，运用pyecharts实现数据可视化。

2.3 研究难点

2.3.1 爬取数据

百度贴吧具有严格的反爬机制，要想高效率获得大量数据，必须针对这种反爬机制对数据收集代码做一定处理。

2.3.2 存储数据

百度贴吧由以讨论话题不同的许多社区（吧）组成，其中用户可以同时关注若干个吧，并在若干吧中发言。因此，不同的吧形成了相互关联的图状拓扑结构。用户在吧中可以发主题帖，也可以在主题帖下进行回帖。因此，每个吧下属的每个主题帖，以及主题帖下属的每个回帖（楼），形成了树状分支结构。这要求我们以合适的方式存储这些数据。

# 三、 实验方法

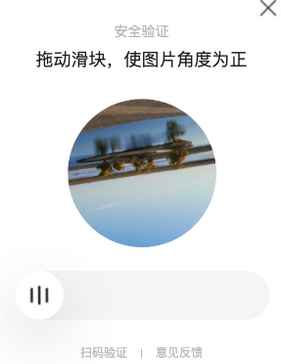
3.1 数据获取

此步骤爬取了四个维度的信息：吧中所有发帖楼层的内容、在吧中发言的用户的信息、根据用户关注的吧记录收集到的吧的信息、吧中所有主题帖的信息。

3.1.1 基于requests库爬取

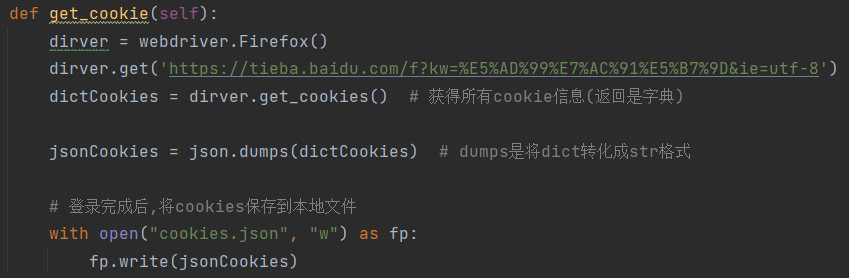
首先，我根据requests爬虫的基本样式，构建了一个访问输入的吧，并从第一页开始，获取页面上每个主题帖的地址，并访问每个主题帖对应的网址，爬取每楼层发言内容的爬虫，class当中包含init、get\_ba\_content、get\_tie\_content、get\_page\_url、run五个部分，根据xpath解析html文件，读取网页文本中的信息。使用这种方法时，爬取到的网页中大部分内容是被注释的状态（存在于符号<!-- -->之间），因此我先通过正则表达式对网页内容进行切割和筛选，选取出字符长度最长的被注释部分即为贴吧页面的正文部分。

但是以这种方式爬取数据，每隔20分钟左右就会爬取到乱码页面，而此时我再手动访问贴吧页面，就会弹出滑动验证码：



而滑动通过验证后，会发现访问网页时的cookies改变了。为此，我设法绕过验证码。

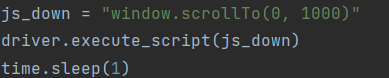
首先我手动修改cookies的内容，但是这样效率太低了，于是我增加了一个获取cookies的函数：



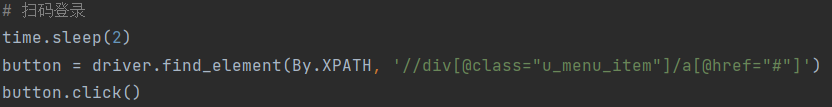
意图实现自动更新cookies，但是在不进行验证时就无法获得新的cookie，因此我换用了基于selenium的爬虫。

3.1.2 基于selenium库爬取

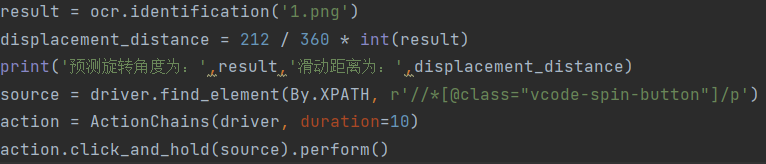
这个爬虫继承了上述基本功能，而且实现了一系列模仿正常用户访问的行为，如下滑页面，sleep等：



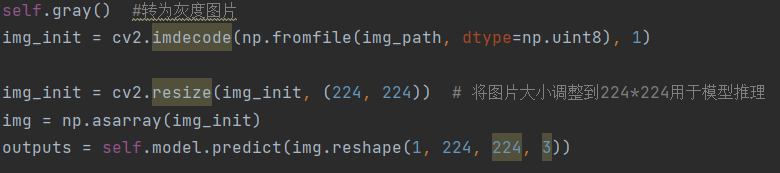
同时实现了扫码登录功能：



但是这样也只能延长验证码出现的时间，无法避免这种情况的发生。因此，利用网络上下载的，已经训练好的基于ocr技术识别旋转验证码的模型，用TensorFlow载入并预测每次显示的验证码所需要拖动的距离，并利用selenium自动拖动验证码：

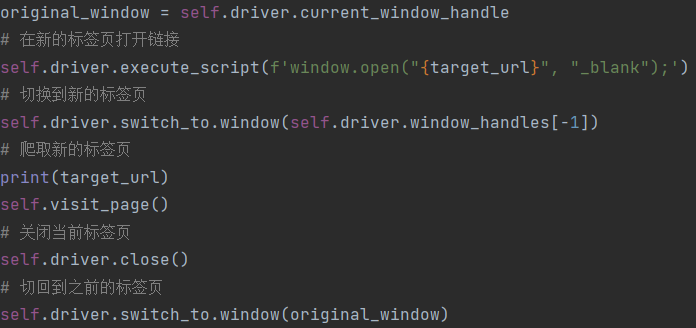


这一部分是旋转验证码模型附带的运行代码：



从而实现自动滑动验证码。

此外，为了避免打开同时太多页面，进行了标签页的管理：

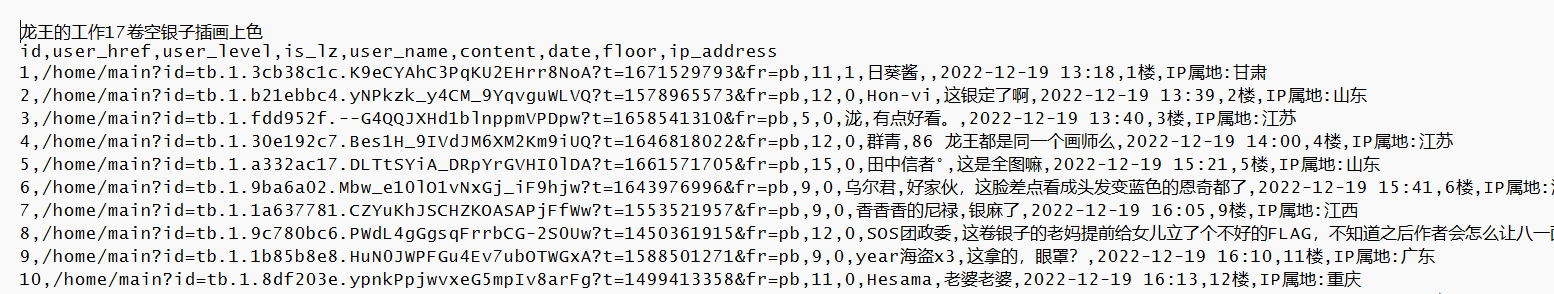


随后增加了user\_spider ba\_spider zhutitie\_spider，来爬取相应的信息。

3.1.3 数据介绍

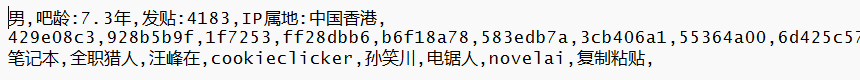
每个主题帖的内容：

id,user\_href,user\_level,is\_lz,user\_name,content,date,floor,ip\_address



用户信息：

sex,age,num,ip\_address,follow\_peo,follow\_ba



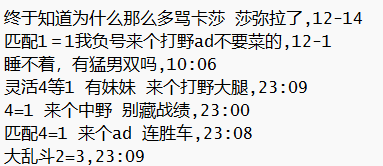
吧的信息：

name,主题帖数,发帖数,关注数



主题帖的信息：

帖子标题,发布时间



3.2 数据预处理

3.2.1 数据清洗

重复性：

首先，由于贴吧中的主题帖的排列顺序处于动态的状态，原先出现在前一页的内容，在爬取完这一页后可能会被挤到下一页去，所以会出现重复爬取到同一个帖子的情况。根据唯一的帖子标题来删除相同的主题帖。

其次，同一层楼的信息也有可能被重复爬取，这种情况下，就需要根据每个主题帖都有顺序编号的楼层序号（1楼、2楼等）来去重。

杂乱性：

统一单位，将后面跟“万”字的数字替换成用10000与前面数字相乘的结果。统一数据类型，用pandas中的astype进行修改。

不完整性：

在爬取的时候，可能会有某些信息不在网页对应的xpath位置中，或者被用户人为地隐藏了，这会造成数据的空缺。根据数据可视化的相应需求，将空行进行删除或用平均值替换。

3.2.2 数据集成

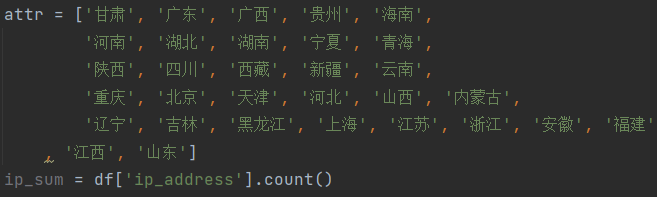
由于每次爬取下来的数据集都会放置到不同文件夹中，并且从0开始编号，因此需要将这些数据集合在一起，并重新编号，便于访问。

3.3 探索性分析

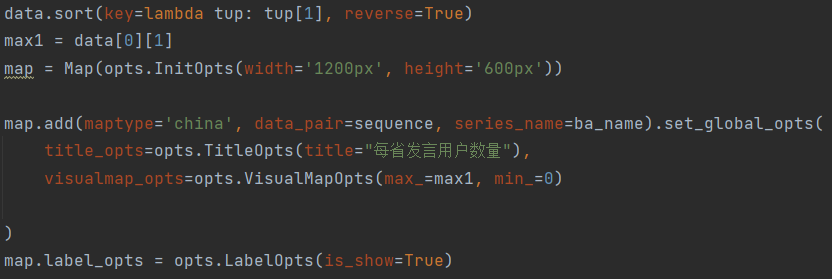
## **3.3.1 不同省份人数占比**

由于近年来互联网政策的更改，在网上各种平台发表的用户言论下方都会附带ip地址所对应的省份，这使得统计各吧省份人数占比成为可能。这一部分使用了爬取的一个吧内主题帖及其内容的数据，把每一层楼的内容作为一项数据，清除相同用户的发言，清除'id', 'user\_level', 'is\_lz', 'date', 'floor', 'content'等列。此时经过清洗的数据只含有数个不同的用户，以及他们的ip属地。

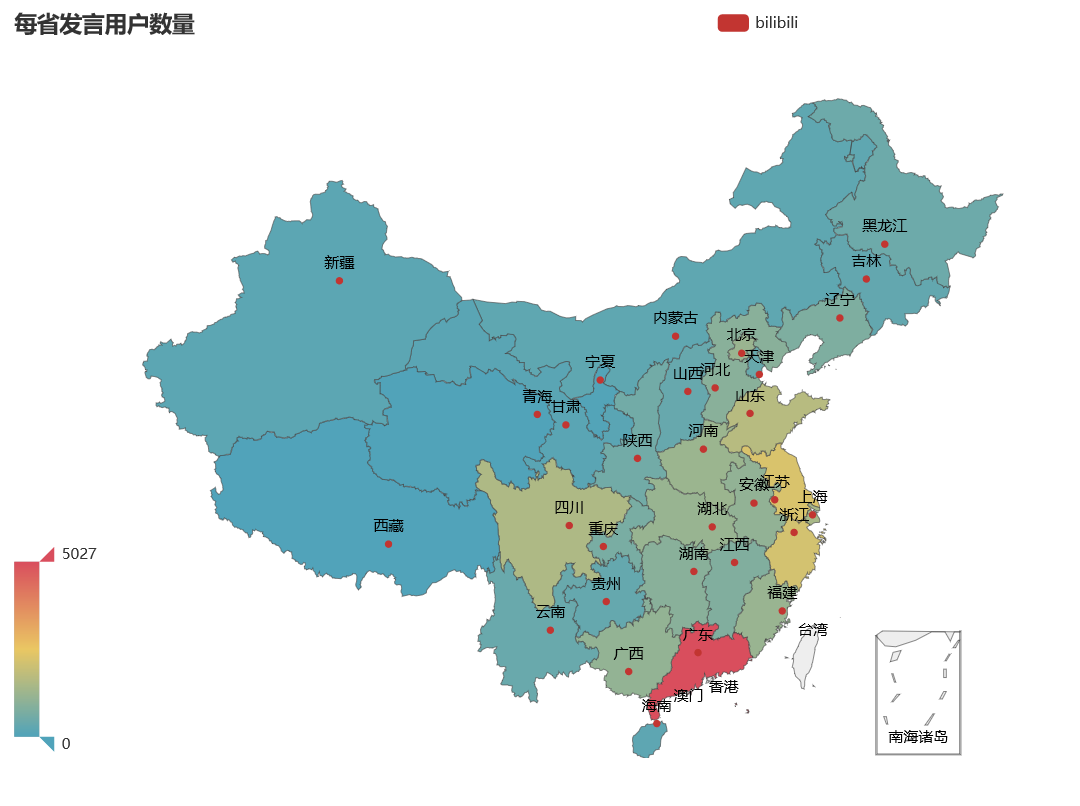
随后根据不同的省份将用户分类计数，填入图中：

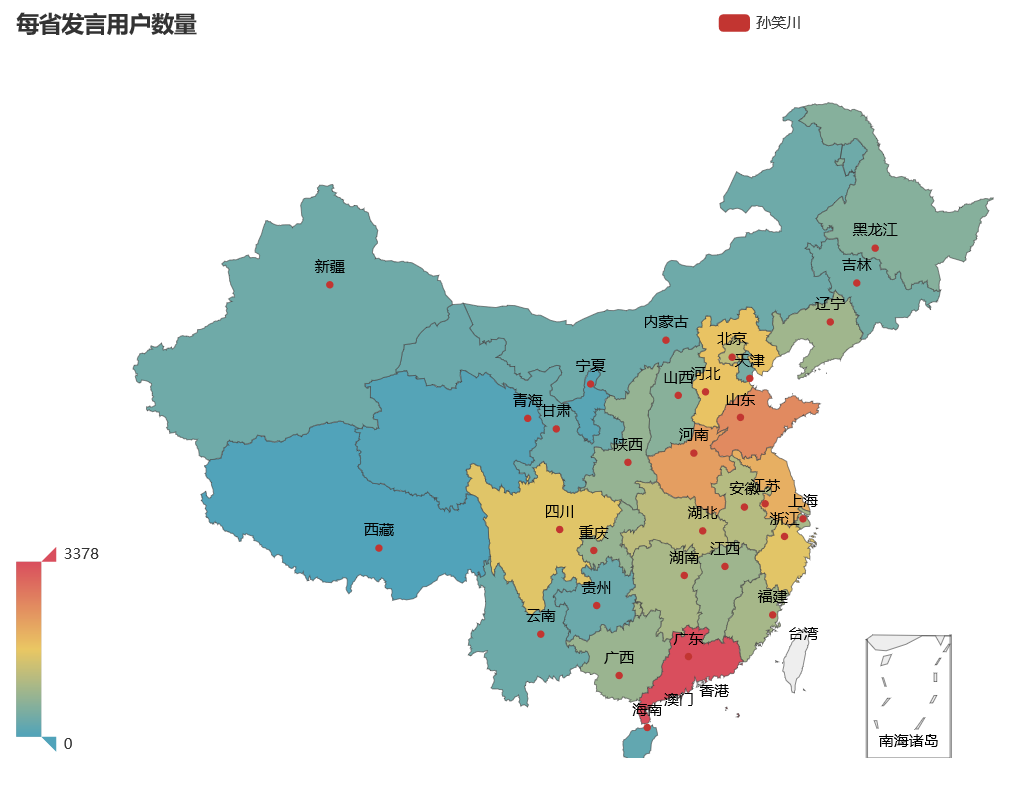


根据人数多少对省份列表排序，并将最多的人数作为画图时的人数上限，0作为下限：



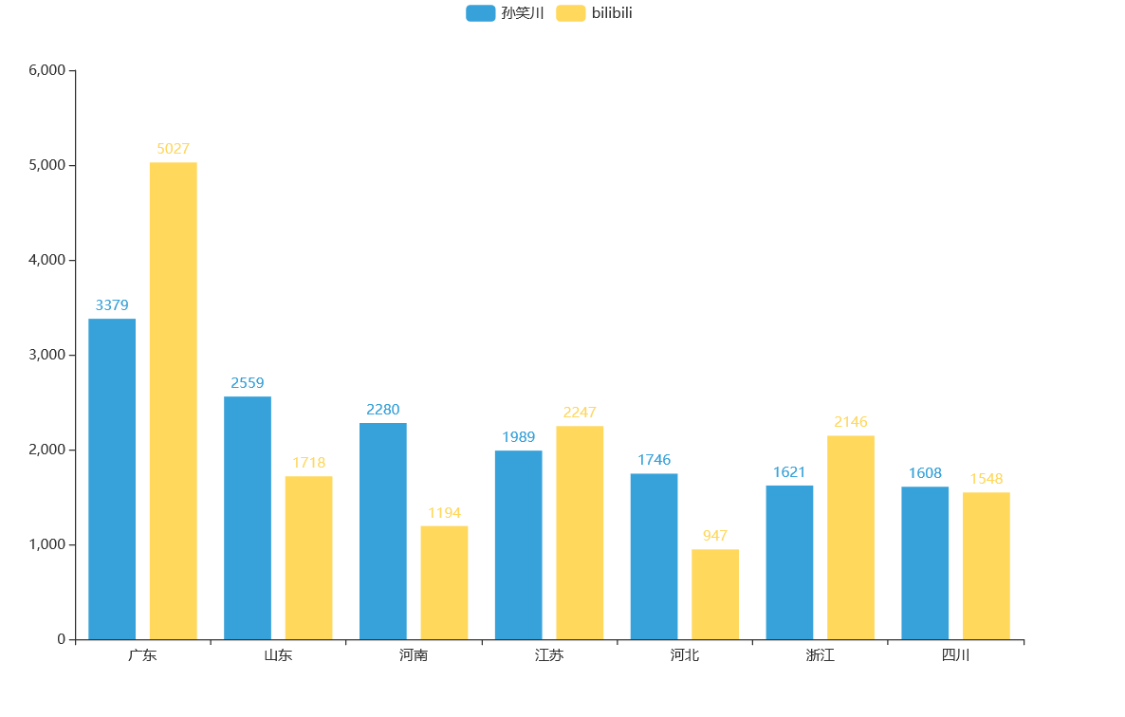
最终生成的地区热力图如下所示：





由此可以看出，总体来说，每个吧的发言用户都较多地聚集在东南部经济发达的沿海地区，尤其是广东省，这与我国的人口分布基本一致。对孙笑川吧和bilibili吧进行横向对比，可以看出在孙笑川吧发言的用户，分布在山东、河北、河南等北方省份的占比相对较大一些，分布也更加均匀；而bilibili吧的用户分布在江苏、浙江的更加多。孙笑川吧的用户多是直播平台的观看用户，而bilibili吧的用户一般是bilibili视频网的观看用户，从图中可以看出两个爱好群体的地域差异。

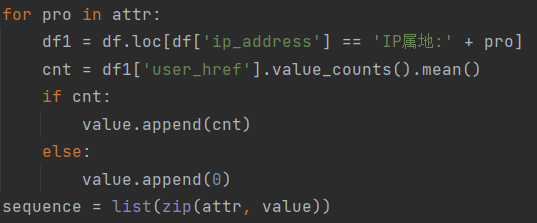
为了更好地说明这种差异，我抽取了每个吧30000个用户与他们的ip地址，制作了如下条状图：



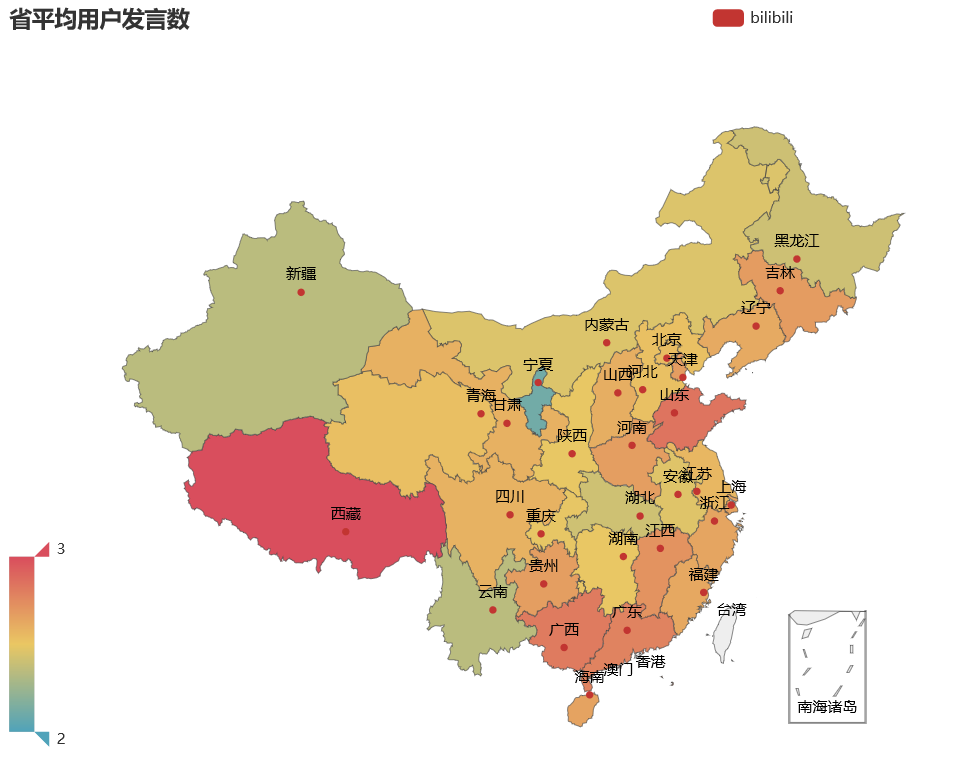
从这张图可以更明显地看出，在样本量一致的情况下，广东省人数在bilibili吧的占比格外大，而靠北方的三省山东、河南、河北，虽然在孙笑川吧具有较大的占比，但是在bilibili吧占比不大。

## **3.3.2 各省用户发言活跃度统计**

在一个吧里，一位用户可能只是发表了一句回帖，也可能在每个主题帖都留下过好几条自己的足迹。为了探究用户的发言活跃度与省份的关系，可以对每个用户的发言数进行计数，然后再根据省份对用户进行分类，最后对这些发言数求取平均值：



由此得到的ip属地与平均发言数的热力图如下：

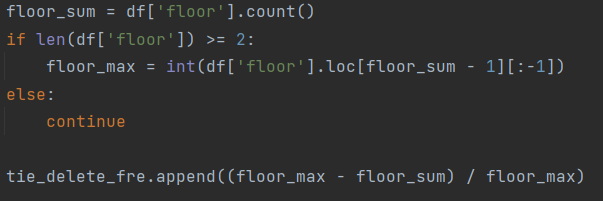


在这张图中，除了本身发言人数较少的西藏、宁夏等地区具有一定偶然性，其他省份主要是沿海地区的用户发言较为积极。这可以说明发言数与地区有一定关系。

## **3.3.3 各吧删帖率对比**

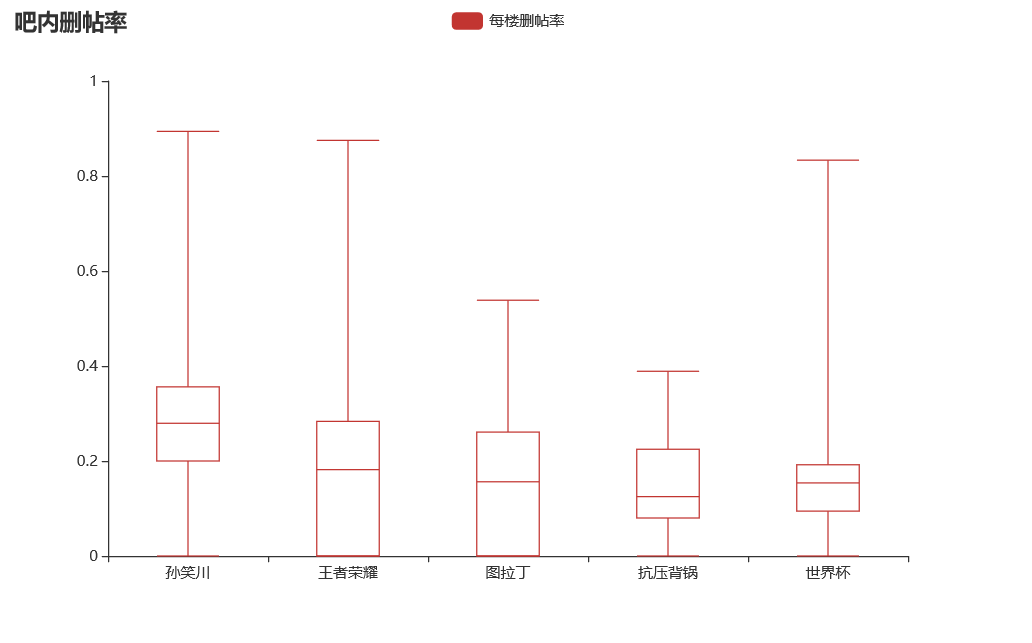
在一个主题帖中，偶尔会有用户的发言由于某些原因被自己或楼主删掉，或者发生被系统“吞楼”的情况。一个主题帖中，楼层的标号是唯一的。而这些情况的发生频率，可以通过楼层标号（1楼、2楼等）的不连续度进行估计。

这是用来统计一个主题帖内的删帖率的代码：

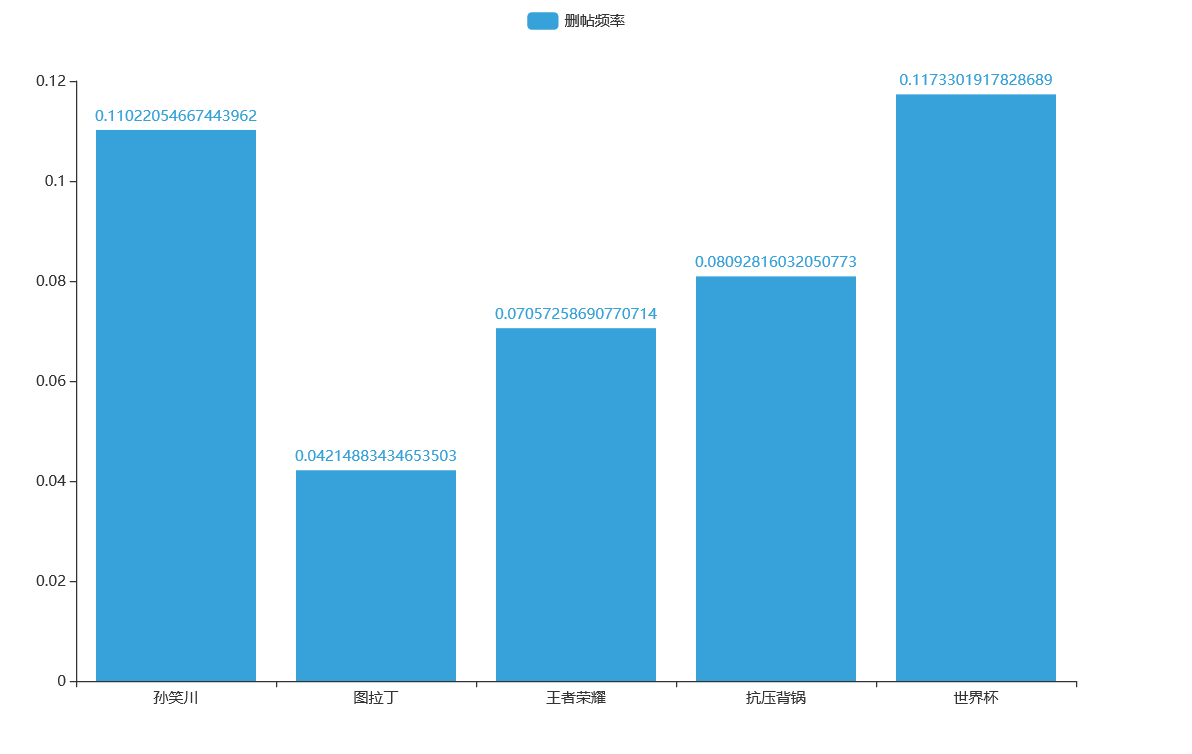


其中，有些主题帖只有几条或十几条发帖，过少的发帖量会导致偶然性的数据，因此把它们去除。

将几个吧内每个主题帖的删帖率进行汇总，得到如下箱线图：



同时，这几个吧的平均删帖率有如下呈现：

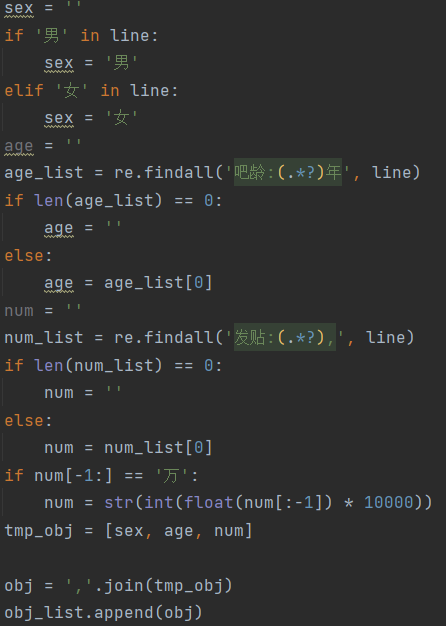


每楼的删帖率可以在0-1之间取任意值，对这个连续随机变量绘制箱线图，可以看出，这些吧的删帖率都呈现右偏态，大多数帖子都保持一定的完整性，删帖率较低，因此可以忽略删帖率较高的异常值。同时，吧的删帖率也受讨论话题、发言情绪激烈程度，吧主与楼主的管理严格程度等影响，因此，我推测，娱乐主题的吧，例如孙笑川、王者荣耀、抗压背锅、世界杯等，可能会发生情绪激烈的冲突，或者违规的发言，以至于被系统或楼主删帖。其中王者荣耀和抗压背锅吧，由于是游戏主题，用户年龄可能较小，违规发言的内容较少。而技术主题的吧，例如图拉丁吧，发言情绪比较冷静，就会有较低的删帖率。

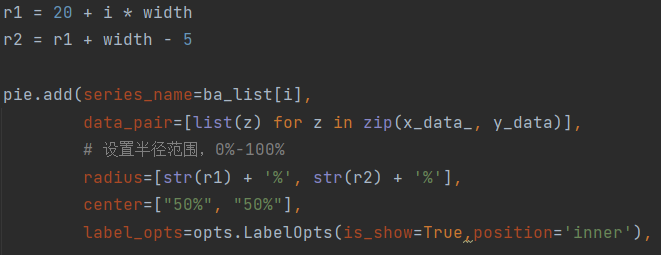
## **3.3.4 用户性别比例**

根据我平日的观察与猜想，百度贴吧的用户多为男性，而男女性的比例也会根据不同吧讨论的话题有所差异。带着这样的猜测，我对几个吧的用户性别比例进行了探究。

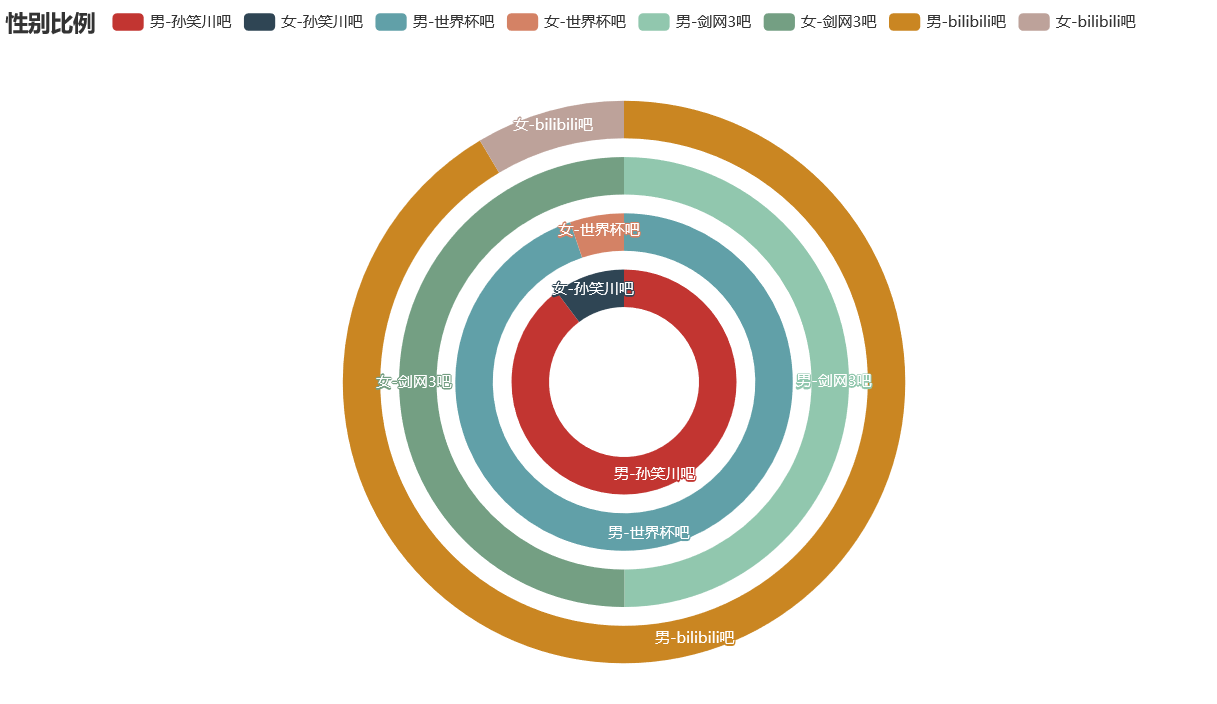
首先对原有的数据文件进行切分，提取出性别以及其他的数据：



然后划分不同半径范围，画出嵌套饼状图：



最后呈现效果如下：

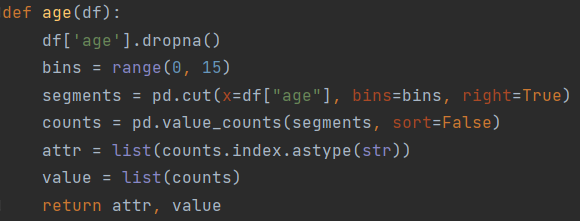


通过这张饼状图，可以发现，孙笑川、bilibili这两个有关视频平台的吧，男女占比比较相似，男性比例都多达90%以上。视频平台逐渐成为如今网民浏览网络时占比相当大的一部分，加上百度贴吧自创建以来男性用户多于女性的特点，便呈现出这种比例。而剑网三这款女性玩家非常多的游戏，以它为话题的剑网三吧男女比例相当。而较少女性关注的世界杯话题，女性比例非常少。

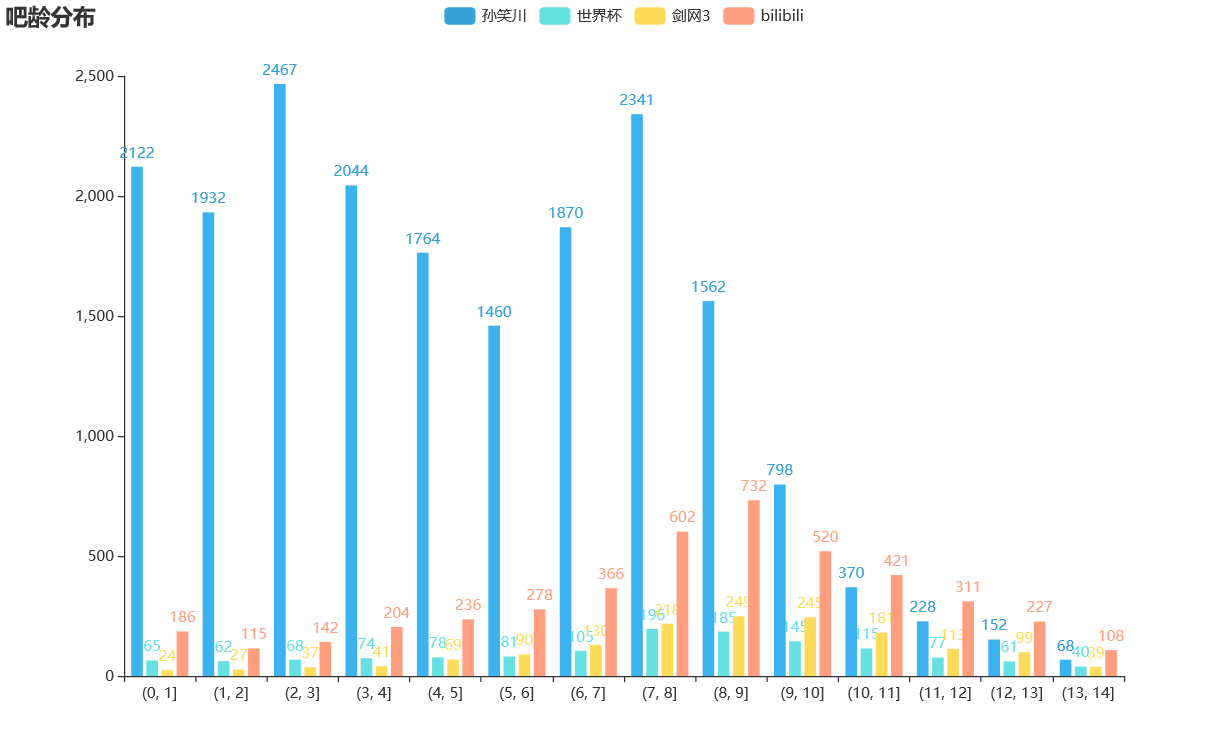
## **3.3.5 吧龄分布**

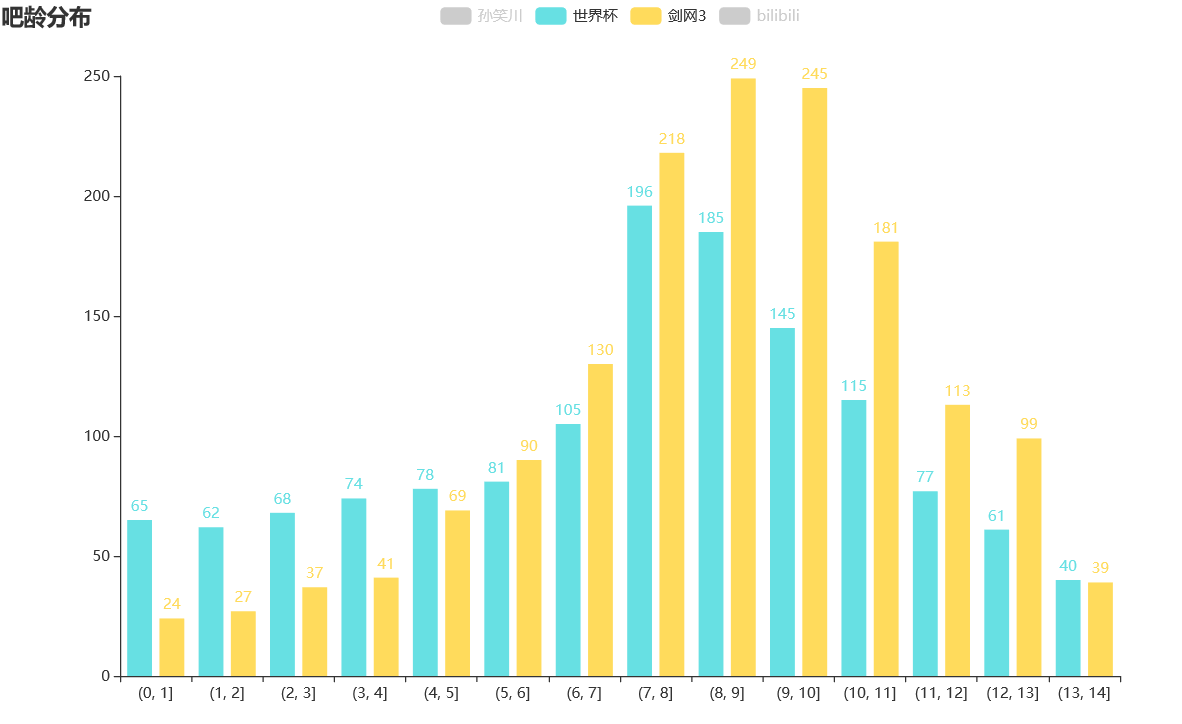
吧龄是指一个用户从注册百度贴吧至今所经历的年数。统计一个吧的用户的吧龄分布，可以看出吧中的新老用户比例，吧兴起的时间，以及一定程度上能反映该话题的用户粘性。

将连续分布的吧龄按照整数值切割：



随后将落在每个区间的人数画成柱状图：

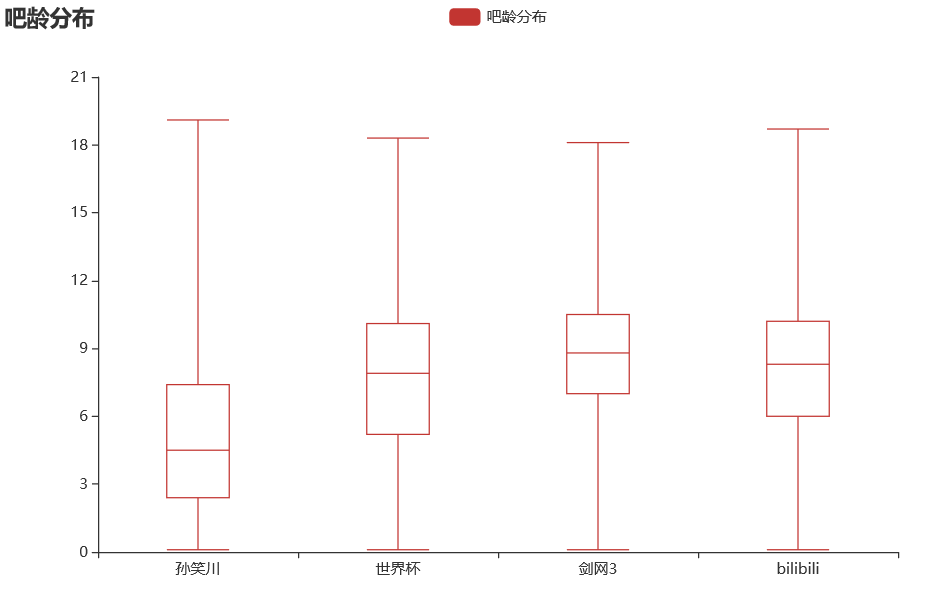




通过上图可以看出，除了孙笑川吧用户的吧龄在9年以下都有较多分布，bilibili吧、剑网3吧和世界杯吧的用户吧龄都呈现一定的峰值，其中世界杯吧的峰值较小，是7-8年，另外两吧则在8-9年。

可以作出猜测：孙笑川吧的用户吧龄较小，一定程度上是由于直播平台是新兴起的网络平台，既反映了孙笑川吧的用户使用互联网的时间总体就较短，可能年龄较小，又反映了此吧目前仍然在吸引大量用户加入百度贴吧，不断涌现出新用户。而另外三个吧的吧龄比例，理论上也会是百度贴吧用户普遍的吧龄比例，而相比于世界杯和bilibili吧，剑网3吧的用户吧龄小的较少，大多用户都是老资历的玩家，反映了较大的用户粘性。

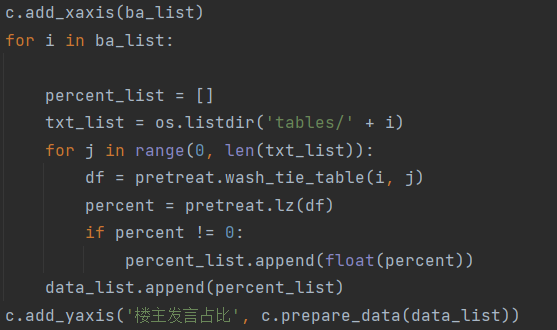
以下箱线图也能很好的反映上述论点：



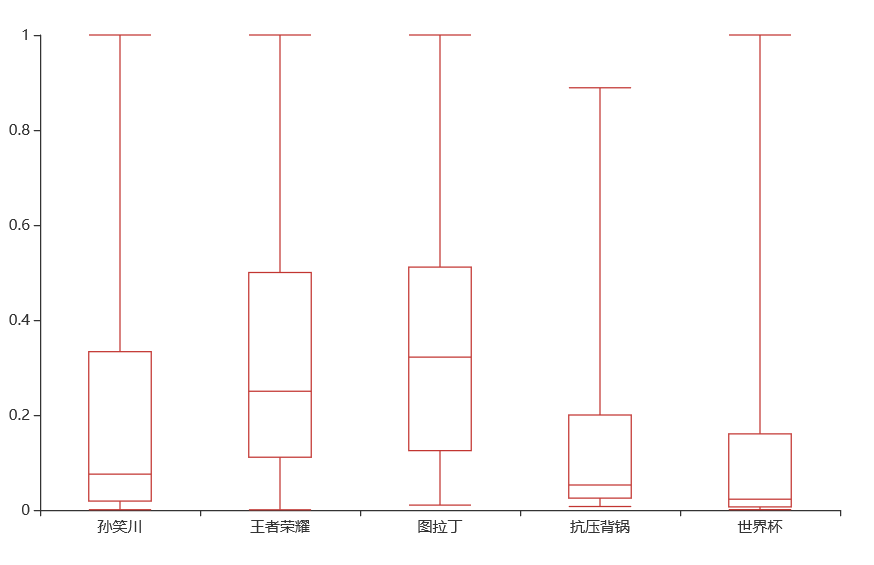
## **3.3.6 楼主发言比例**

在百度贴吧中，每个主题帖的发起者会被成为“楼主”，而楼主在自己帖子内的发言，则会带上楼主标识。由此，我们可以通过数据可视化，来观察不同吧的楼主发言比例有着怎样的差别。

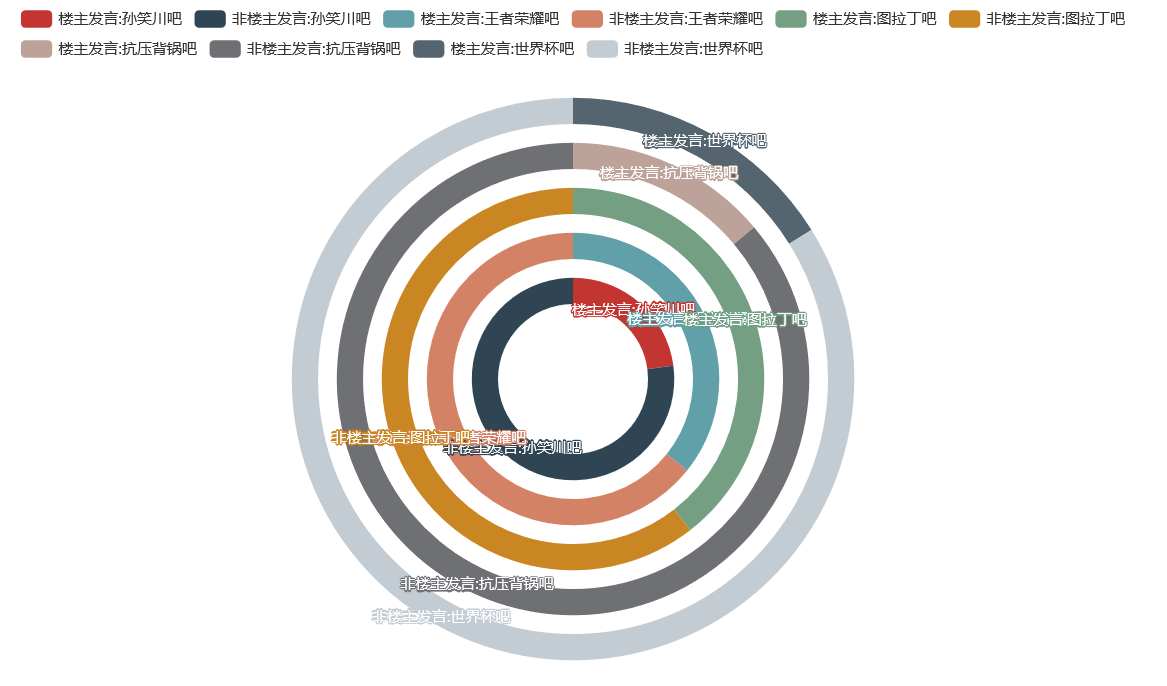
在爬取的数据中有着is\_lz这一栏的信息，是否是楼主用数字1、0来表示，因此直接提取出这一列，求出在所有发言之中的占比，并进行箱线图的绘制即可。



可以画出如下图：



同时，不求比例，而是把一个吧内所有发言汇总，可以画出如下饼图：



可以看到，除了极个别特殊值帖子中的楼主发言占比很大以外，帖子内的楼主发言占比与话题的讨论形式相关性很大。比如图拉丁吧对于相关技术的讨论，就会以楼主讲述为主，吧友讨论为辅；而孙笑川、世界杯这样的吧，通常是根据一个话题展开，吧友各自表达自己看法，相互讨论的形式，就会有较小的楼主占比。

## **3.3.7 等级与发言数关系**

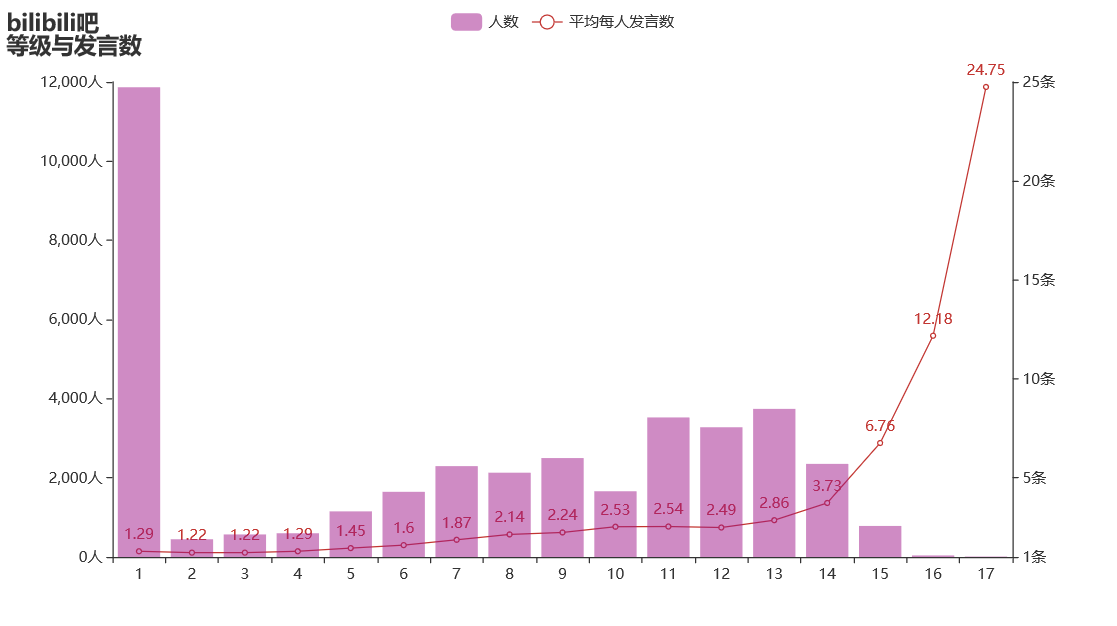
在百度贴吧中，每个用户在每个吧内都有一个独立的等级体系，等级一般由用户在吧内的发言数、签到次数决定，反映了一个用户在吧内的活跃度，这个数值是整数类型。

因此可以探究各吧用户的等级和他们的发言数的关系，这里以bilibili吧为例。

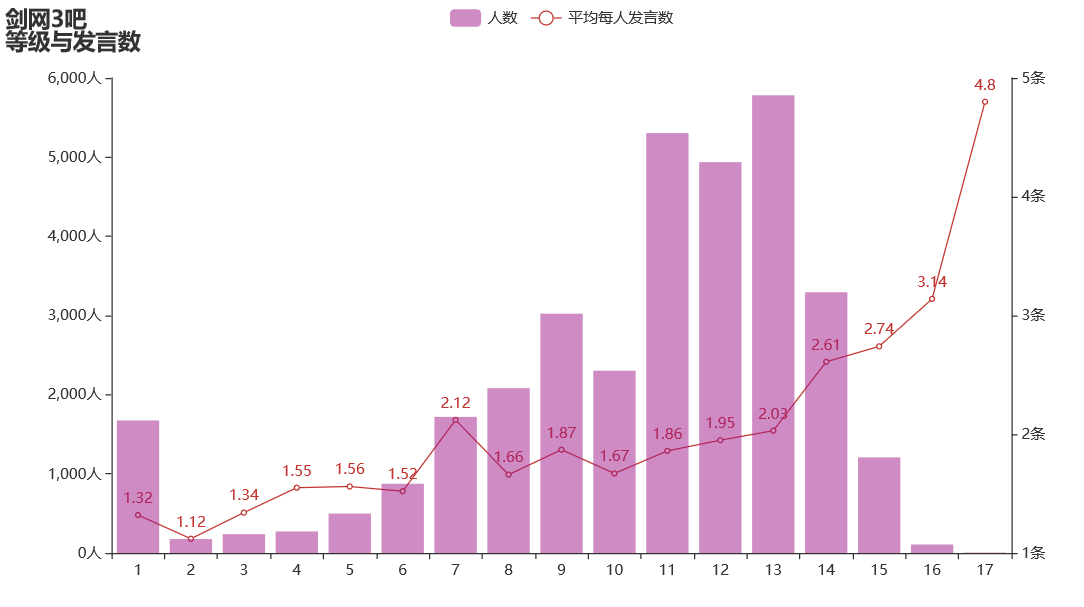
这里在柱状图上再添加一个纵轴，并叠加折线图。为了让柱状图不要遮挡折线图，对柱状图进行了适当的透明处理：



最终画出如下图表：



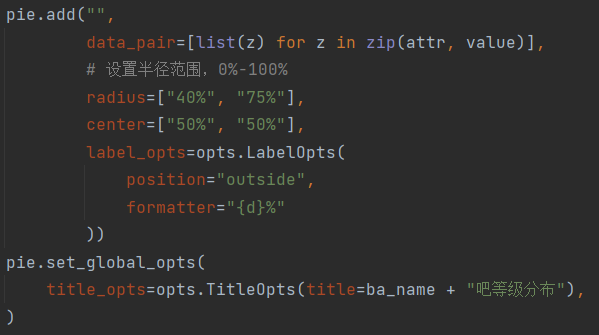
该图可以看出，bilibili吧的用户中，有很多1级用户，其他等级用户在7级附近和12级附近大致有两个峰值，这与前面问题中的吧龄比较类似，也可以反映出用户普遍入吧的时间。1级用户较多，可以说明该吧的用户还在不停增加，但是用户粘性不是很强，这些用户不倾向于进行更多发言。同时，折线图符合等级与发言数的正相关关系，而且这个关系更加接近于指数关系，这说明每升一级所需要的发言数是逐渐增加的。



而剑网3吧的图则有所不同。折线图的波动可能是由于样本量的不足，而1级人数明显相对较少，说明该游戏不如bilibili平台更吸引新用户。

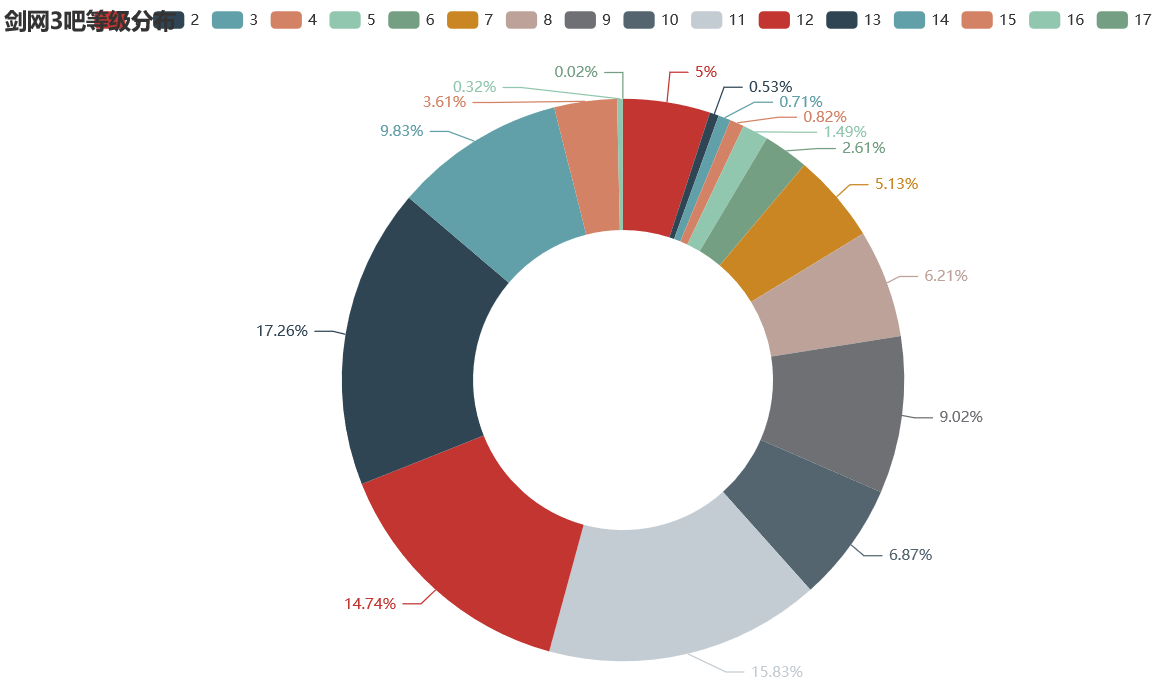
## **3.3.8 等级分布**

为了将3.3.7中等级的比例更加直观地进行呈现，我又绘制了饼图：

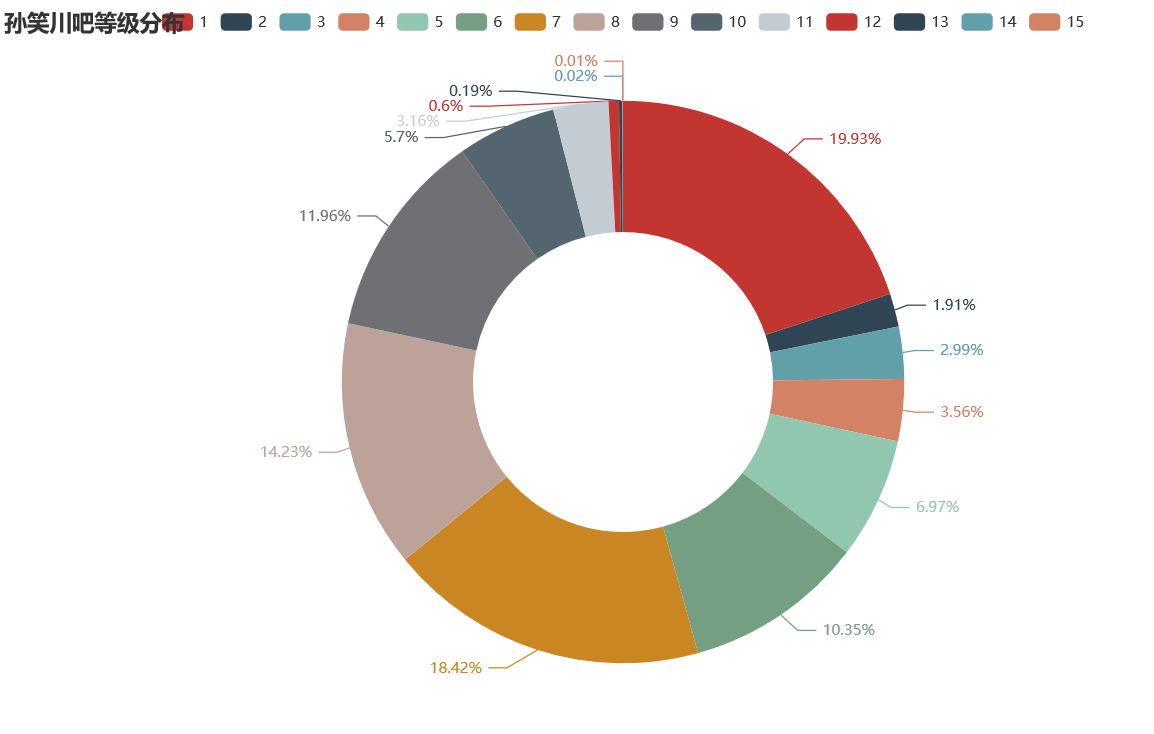


而不同吧的等级分布，最明显的便是1级用户的比例，和其他等级中最多人数等级的高低对比。

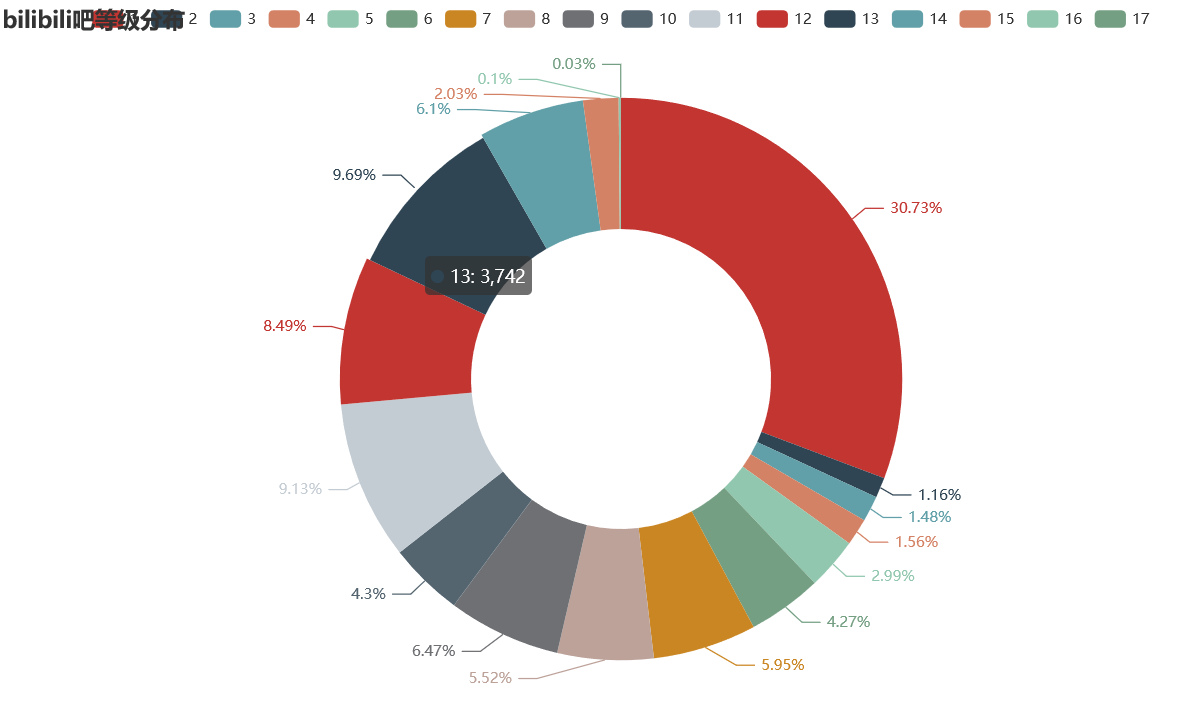
剑网3吧：最少的1级用户比例，说明该话题如今的活跃度降低，而较多的高级（13级）用户，说明该话题资历较老，兴起较早。



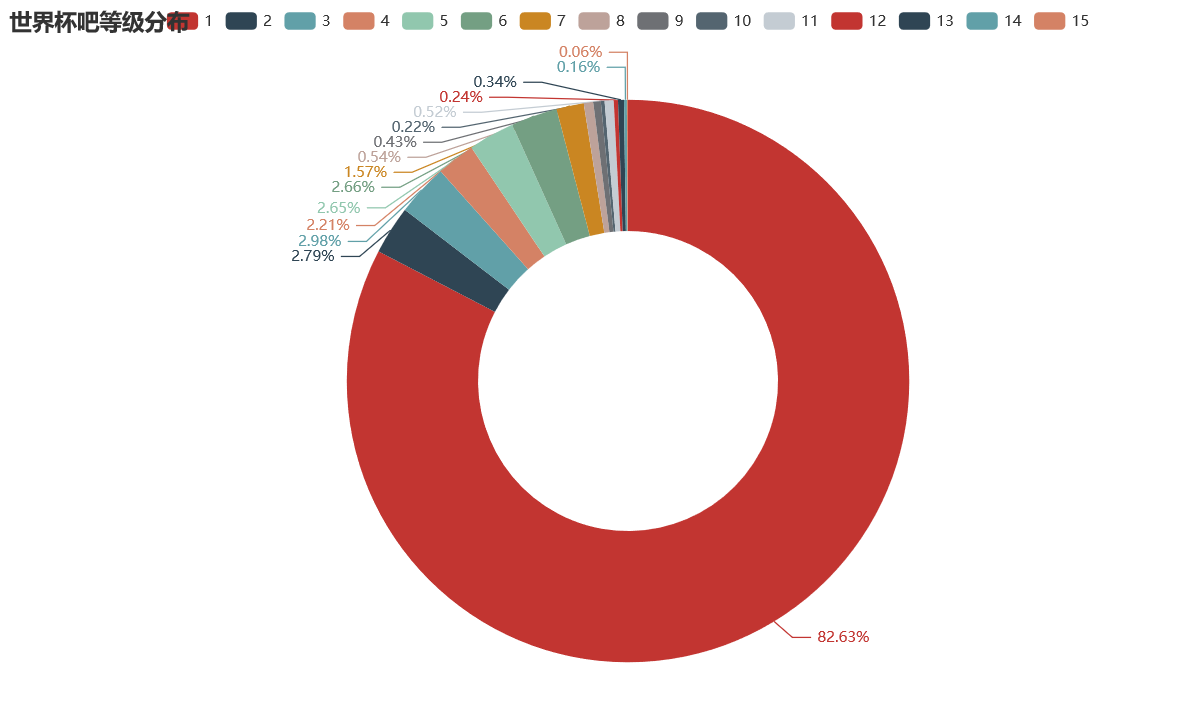
孙笑川吧：稍多1级用户比例，活跃度一般，大量用户聚集在7级左右，说明话题兴起较晚。



Bilibili吧： 1级用户比例较高，活跃度较高，而较多的高级（13级）用户，说明该话题资历较老，兴起较早。



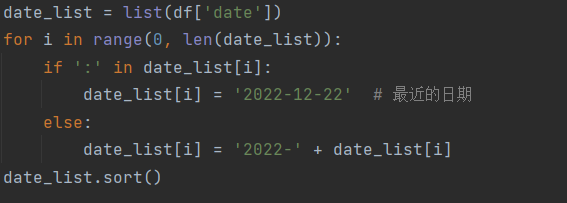
世界杯吧：1级用户比例非常高，活跃度很高，但是用户粘性低，几乎没有高等级用户。



## **3.3.9 主题帖发帖日期分布**

此问题是运用爬取的主题帖及其发帖时间数据进行分析的，试图通过可视化分析，对每个月、节假日、周末的发帖数变化有一个直观的认识。

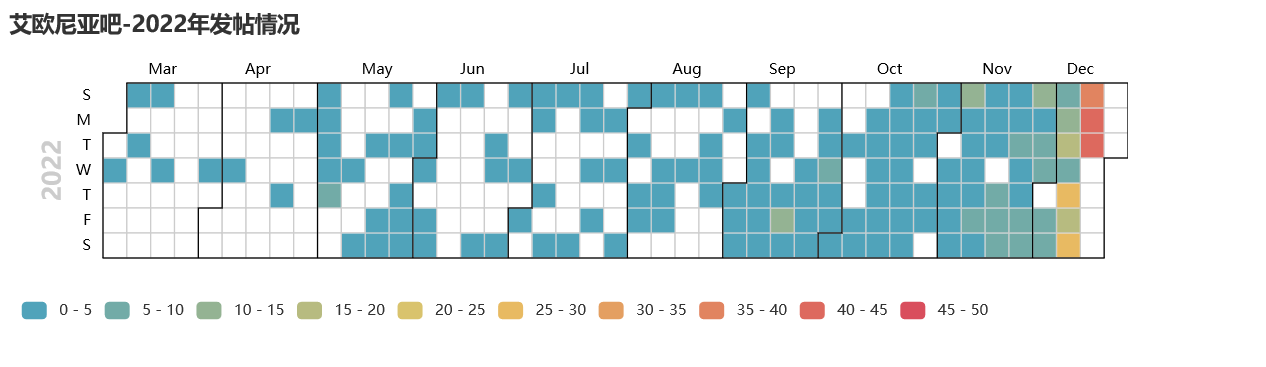
首先爬取到的当日的帖子，所显示的时间是没有日期，只有时分秒的。因此要将它们转换为日期。爬取这些数据是12月22日，故做如下处理：



然后，挑选出最早与最晚日期，在日历上画出热力图：



最终呈现效果如下：



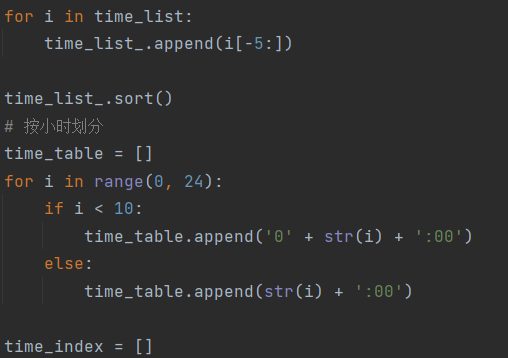
经过分析，呈现效果不理想可能的原因是百度贴吧会定时清理长期没有人回复的“坟贴”，这导致发帖时间很早，而且不再有人回复的帖子数据缺失。同时，由于我的爬虫是从吧的第一页开始爬取，而越新被回复的帖子排列得越靠前，所以爬取到的数据更多都是最近被回复的帖子，就导致越近的日期，爬取到的主题帖数量越多，在图中就会有越高的活跃度。

而且，某种程度上也能看出，周末时段比工作日具有更多的发帖活跃度。

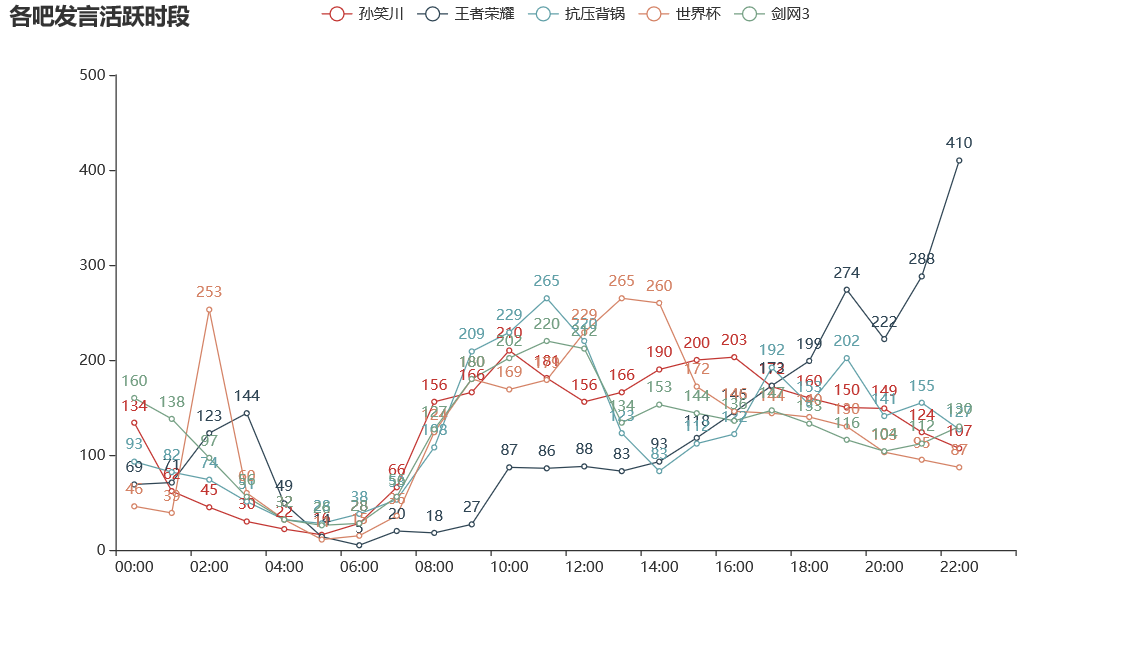
## **3.3.10 各吧发言活跃时段**

在贴吧里，每条发言都会附带精确到分钟的发送时间，由此可以分析每个吧发言数根据时间的变化规律。

将时间按照小时划分区间，计算每个区间的发言数：

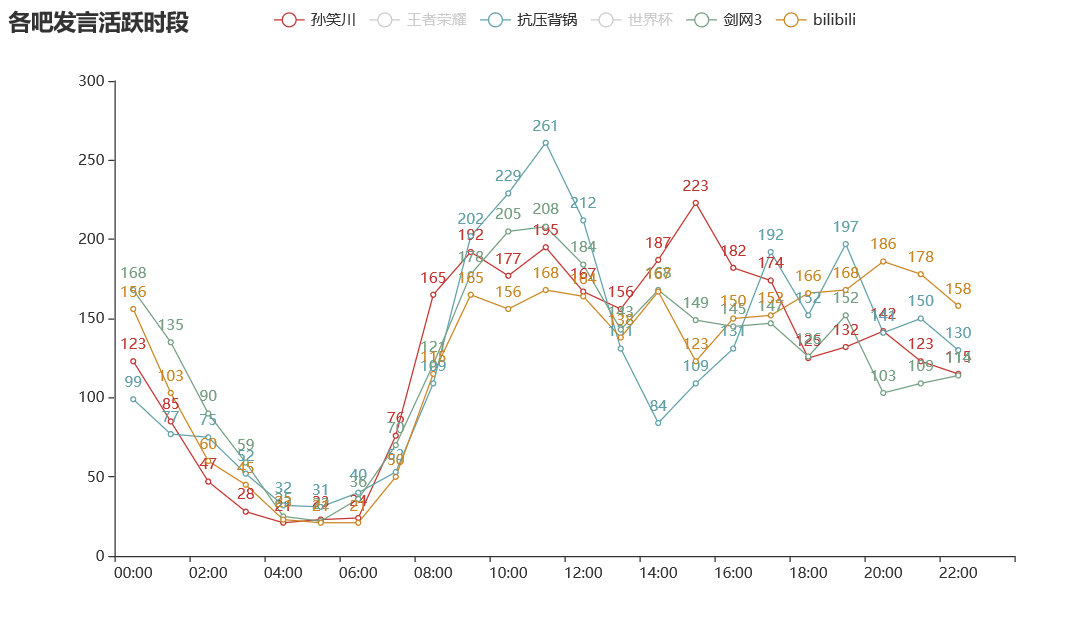


由此得到下面折线图：



可以看到，吧的活跃时段有共同点，同时也受吧所讨论的话题影响。除了和其他吧基本一致的下午活跃时间，世界杯吧的活跃时间集中在凌晨2-3点，这可能是由世界杯的比赛时间决定的。而王者荣耀吧在夜晚比较活跃。

隐藏世界杯吧、王者荣耀吧两个较为特殊的吧后：



可以看到，大多数吧在早上2时到上午8时发言数惊人地一致，均是一个发言的低谷期，而其他时段较为活跃。这符合城市居民的生活作息。

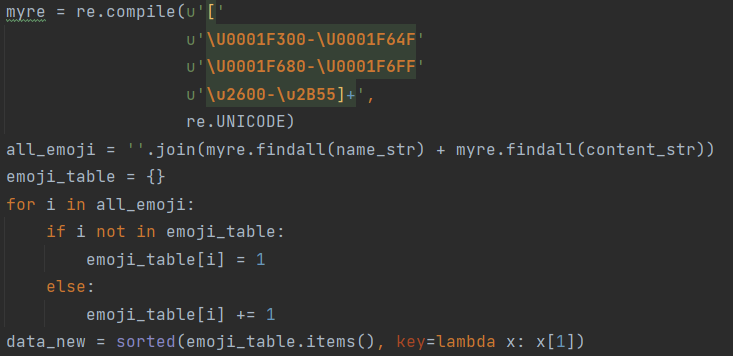
## **3.3.11 常用表情统计**

如今，emoji表情被Unicode编码采纳，普遍应用于当今网民的发表的文字中。为此，我对贴吧用户的emoji使用情况进行了探究。

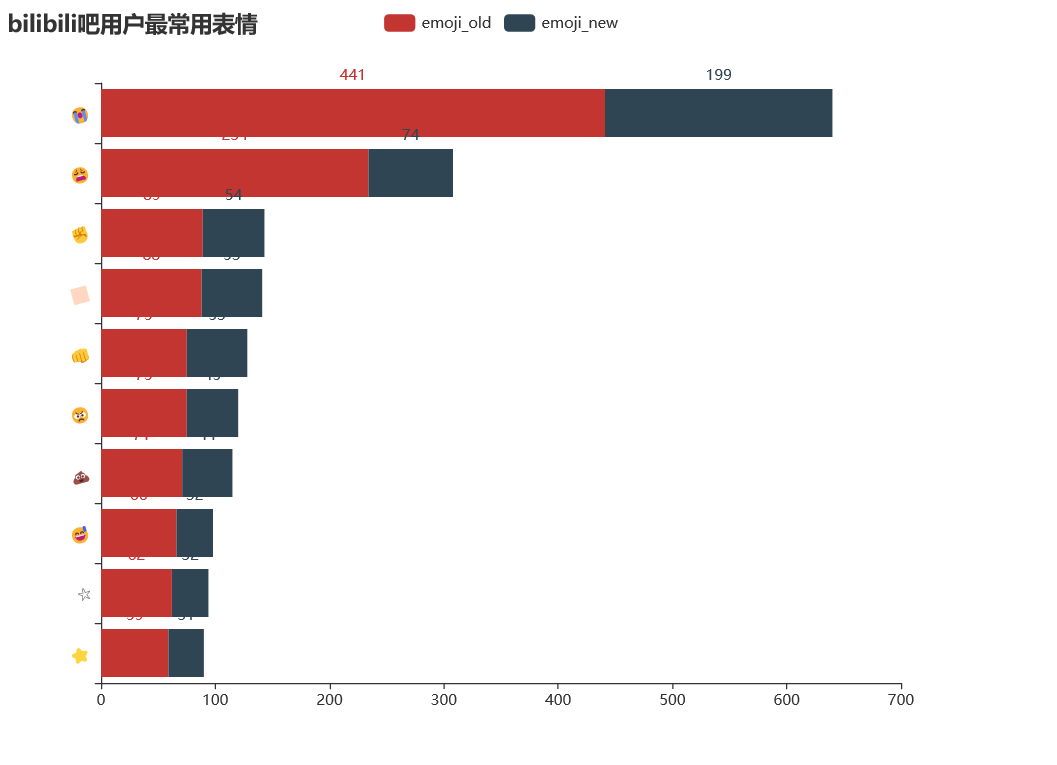
为了研究emoji的使用趋势变化，可以根据时间范围将数据切分成两部分：



同时，提取并对文字中的emoji进行计数：



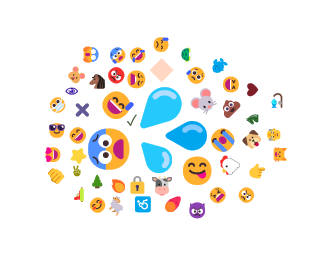
最后，画出emoji使用数目的条形图：



通过上图可以看出，“流泪”表情在bilibili吧使用量遥遥领先。虽然目前第二个和第三个表情差距悬殊，但是用户流行的表情趋势将会拉进他们的距离。

随后，我还制作了emoji表情的词云图片，用来更直观地显示每个吧使用表情的情况：

Bilibili吧： 孙笑川吧：

通过对两张图片的对比，可以发现，在两个吧内，有很多网络流行表情都是一致的，但是使用量呈现出较大的区别，这与两个吧讨论的不同话题有关。

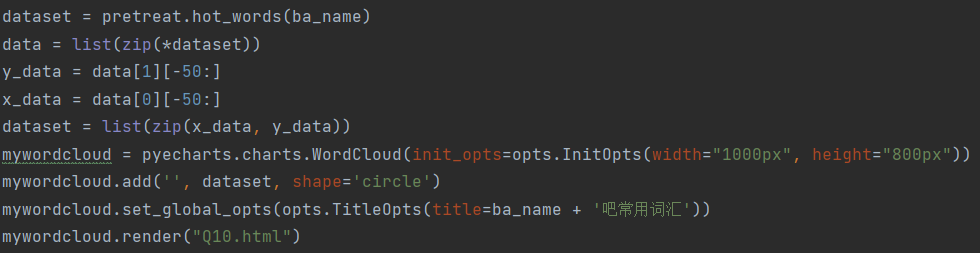
## **3.3.12 常用词汇统计**

为了更好地了解贴吧发言具体的主要内容，我使用jieba库对贴吧用户发言的内容进行切割，提取出关键词制作词云图片。

分词的时候经常会出现辨识度不高，特征不明显的词汇多次出现在每个吧中的情况，为了避免这种情况，我首先限定词性为名词，然后删去了一些在贴吧发言中比较普遍，没有特征性的词汇。



然后，对分出来的关键词汇取前50个制作词云图片：



最终效果呈现如下：

图拉丁吧： bilibili吧：

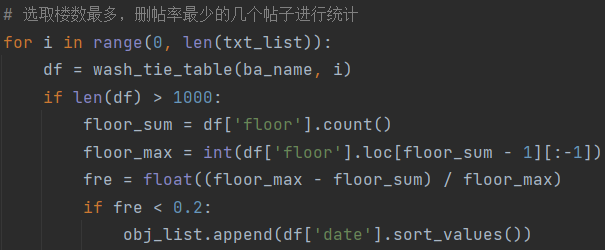
 

可以看到，以电脑硬件技术为讨论主题的图拉丁吧，和以动画剧情为讨论主题的bilibili吧具有截然不同的关键词组成。

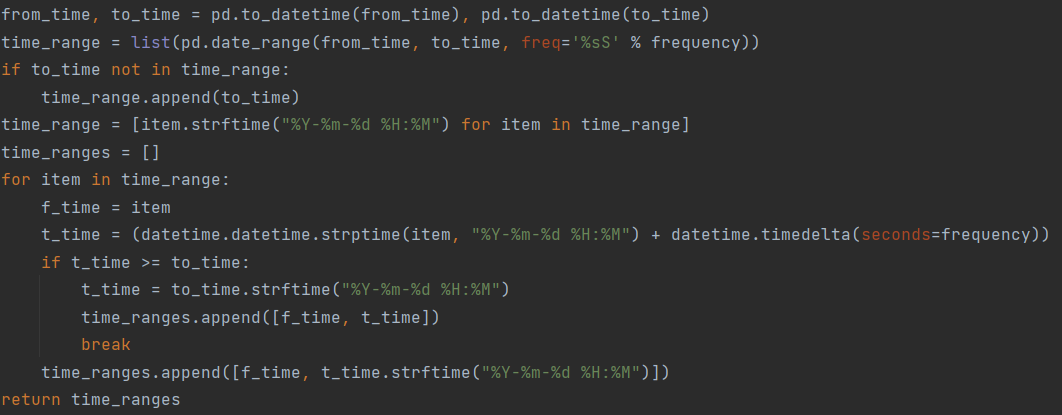
## **3.3.13 单个帖子发言数随时间变化趋势**

百度贴吧中存在一些发起时间很早，但是一直会有人回帖的帖子。为了查明这些帖子“经久不衰”的原因，我对其中发言的时间进行了可视化分析。

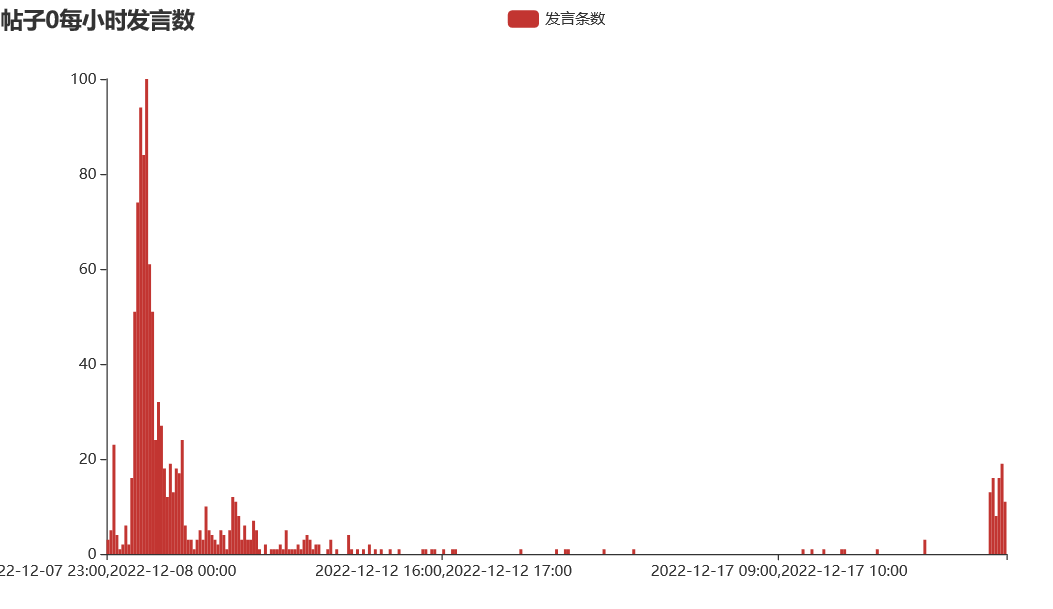
首先提取出楼数最多，删帖率最少的几个帖子进行统计：

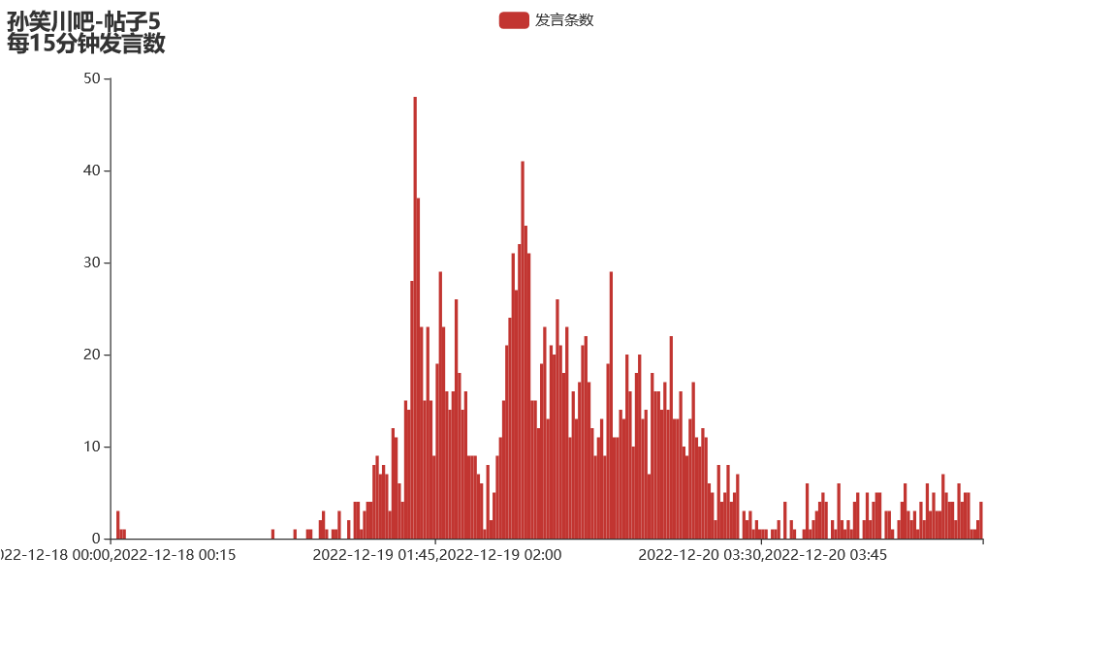


然后分割时间段，一开始设置的每小时不足以反映趋势，则切割成了15分钟：



然后对每个时间段发言的条数进行统计，这是随机选取的两个帖子的情况：



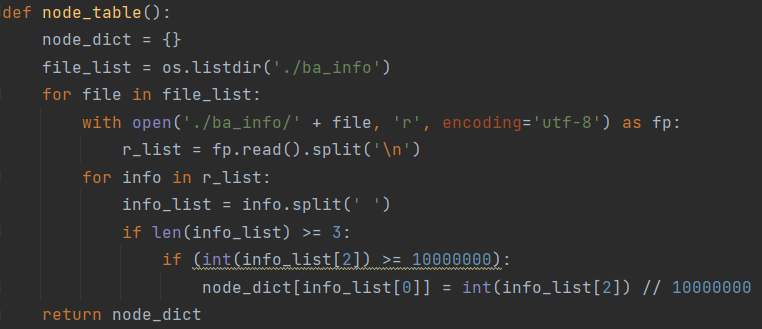


通过上图可以发现，帖子的活跃高峰一般左侧陡峭，右侧平缓，帖子呈现“升温快、降温慢”的规律，这可能是由于被回帖可以使帖子被顶到用户浏览的最上方，而这更容易引起用户的兴趣，收获更多回帖，从而使热度较为经久的原因。

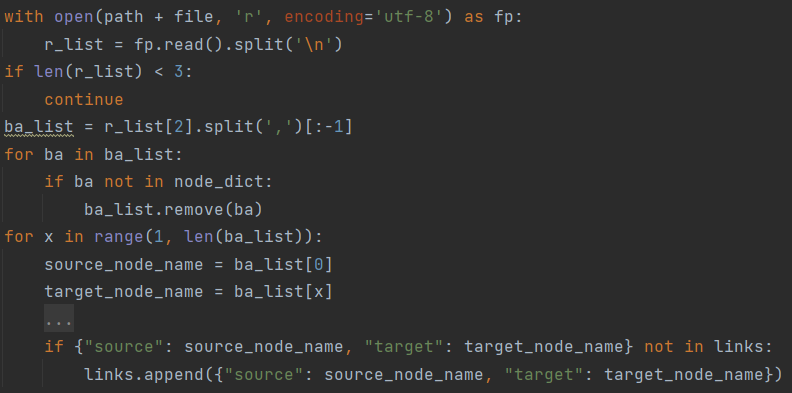
## **3.3.14 吧之间的关联情况**

百度贴吧中存在许多吧，这些吧由用户建立，一般由动漫、电影、明星、网红等的名称命名，可以从吧的名字直接识别出吧讨论的主题。这些吧的建立没有数量限制，而且也没有严格的分类，所以，不能找到一个合适的目录来访问所有吧。这种情况下，为了找寻吧之间存在的内在联系，我通过用户对吧的关注情况，进行了可视化分析。

首先，对吧的信息进行提取，形成节点集。此时对于节点的大小有三个选项：发帖数、关注数、主题帖数。由于使用另外两者会导致节点大小差异过大，于是使用关注人数来决定节点大小，同时去除掉关注人数过少的吧：



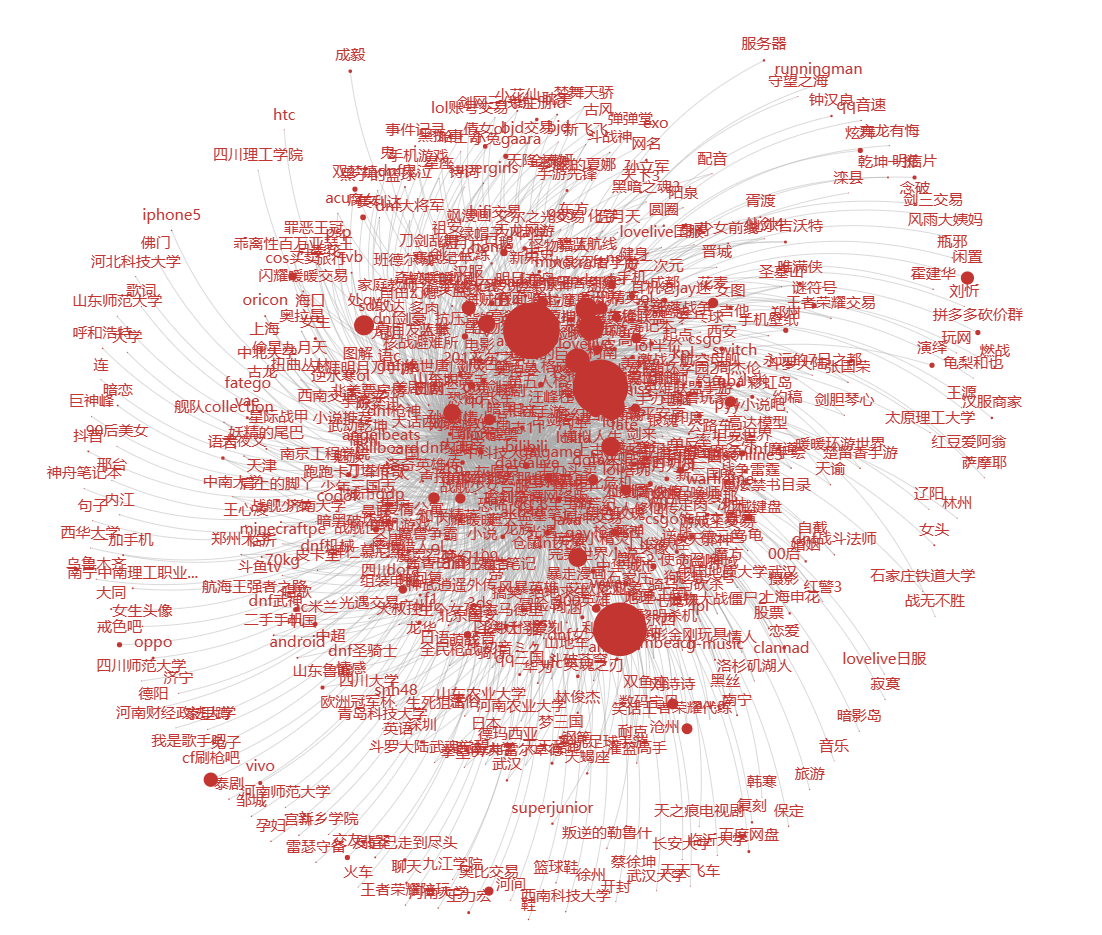
随后根据收集到的用户数据，对用户的关注吧信息进行提取，把被同一个用户关注的几个吧互相连接上边，加入边集：

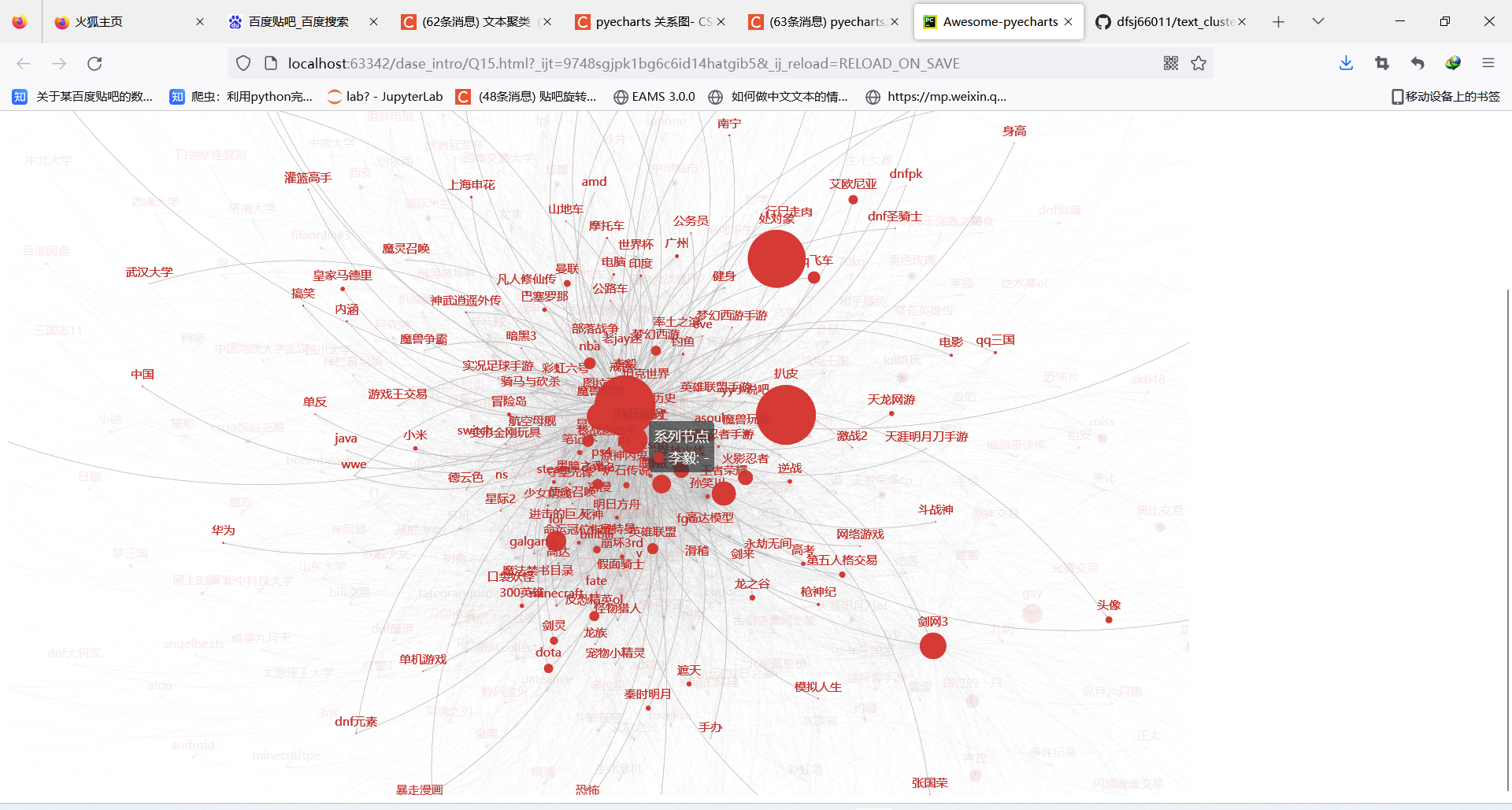


最后，绘制出关系图：



最终呈现效果如下：





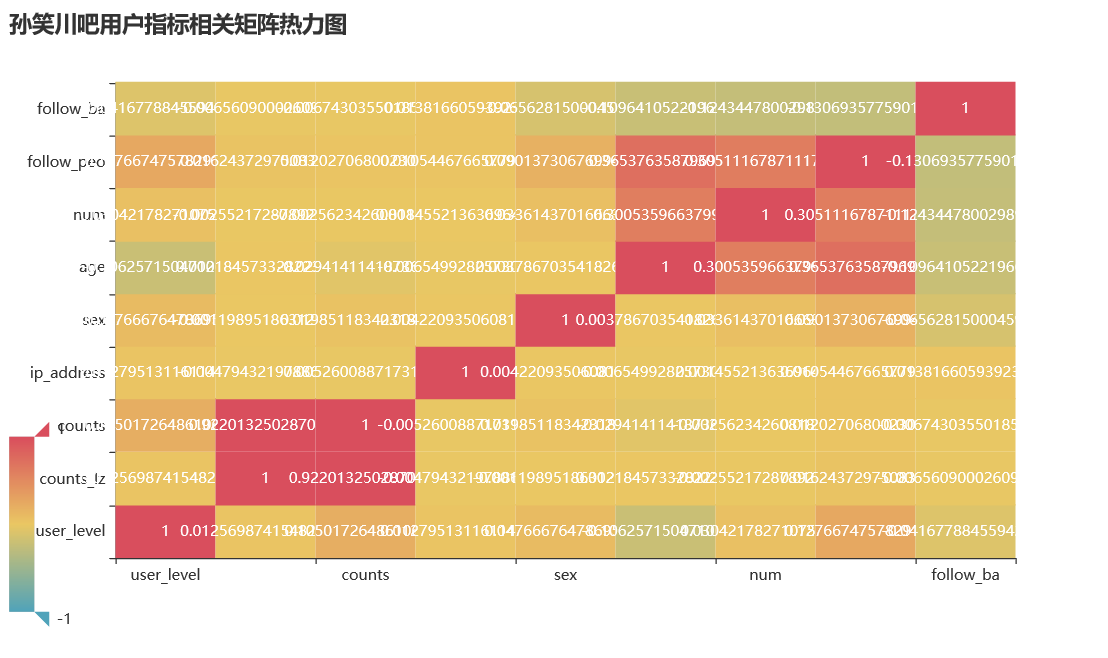
这张图很明显地存在三个较大的节点，是李毅吧、扒皮吧和处对象吧，这三个吧由于资历很老，关注人数很多。此外，还能明显看到四个圈层，这反映了吧的大众性，边缘的吧属于小众爱好，关注这个吧的人可能并不关注别的吧；而接近中心的吧的用户，一般也会关注其他的吧。

## **3.3.15 用户指标热力图分析**

为了对用户发表的内容作出预测，对用户的各项指标进行了相关性分析，其中对ip属地等无法用数字表示的指标进行编号：



但是结果不尽人意，指标之间的相关性不高。因此没有进行后续工作。



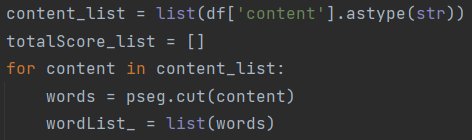
3.4 数据建模

本实验意在通过数据分析，对百度贴吧的发言内容有较为全面的了解。因此，用户发言的情感分析是不可或缺的。我学习了情感分析的原理和基本方法，构建了模型用于更好地掌握网友发言的情感态度倾向。

首先下载如下情感词典，作为初始数据集使用：



然后使用jieba分词，对用户发表的内容进行分词：



最后，对这些词汇进行情感判断，以下是对积极情感词汇进行的赋分规则：



情感词典预先根据词汇的情感倾向，对一些词汇进行了赋值，正向情感是正数，负面情感是负数。而赋分的规则从头到尾扫描以逗号分割的句子，用程度副词与情感词的权值相乘，如果一句里面有多个情感词则将它们的分值相加，最终得到句子本身的权值。

而根据无监督学习的基本思想，可以在每次判别分数之后对情感词典进行更新，将分数高的句子中，不存在于情感词典里的词汇添加进去，并赋相应权值，这样得到的新词典，对于贴吧网友的发言的情感判断就会更加准确。赋权值的方法很多，我最终使用了句子中情感词的平均值来作为新加入词的权值。

然后对模型训练前后对孙笑川吧吧友的发言情感分析准确度进行了分析：

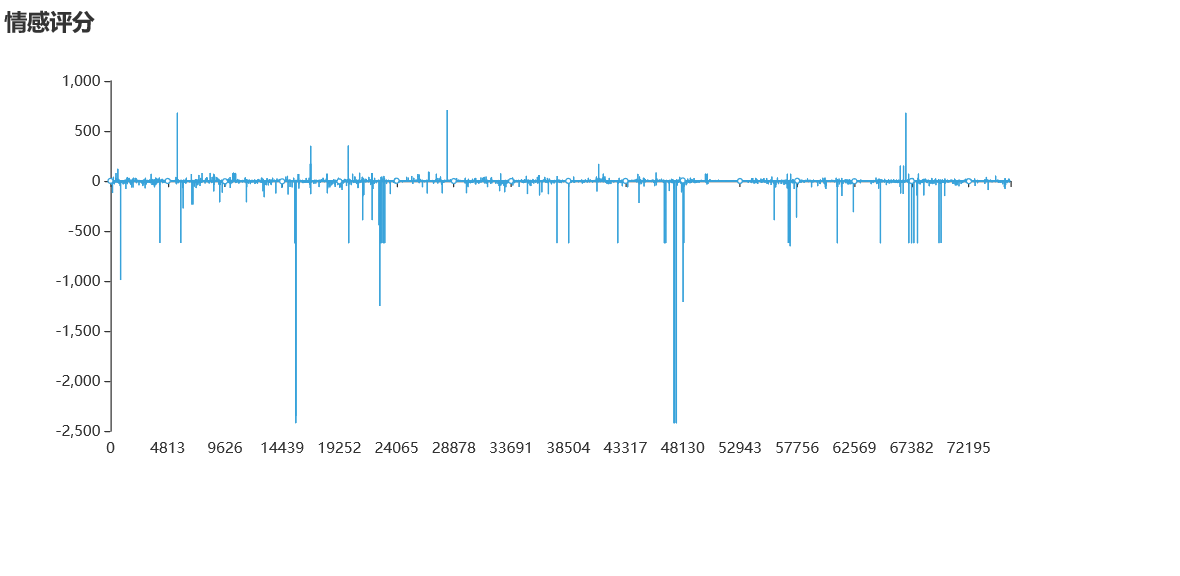




可以看到这是训练前后的对比，有一项的评分错误被纠正了，同时也让分值更为准确了。

同时，由于孙笑川吧发帖常以“如何评价”开头，我对于孙笑川吧的发言进行情感分析，以确定这些发言的情感倾向。

将这些语句的情感评分画出柱状图，可以很明显看出，孙笑川吧的用户发言更倾向于负面情绪的发泄或吐槽类。



# 四、 实验评价

随着互联网产业的兴起，无数用户在网络上发表海量的文字信息，并且这些内容更新迭代快，密度大，要想从这些数据中提取有用的信息，对于用户在论坛上的发言分析，是必不可少的环节。在本实验中，我对百度贴吧的各项用户指标、吧的信息、帖子信息等进行了较为全面的分析，用数据呈现出百度贴吧这个互联网论坛的基本概况。与此同时，我还建立了情感分析模型对用户发言内容进行分析。

在此过程中，我共编写代码2000行以上，制作几十张图，用于分析的数据集达到几百兆。这种工作量对我个人而言是很大的挑战，同时也是一次大胆的尝试。

本次实验同样存在许多不足，有很多想法因时间所限未能实现，比如对吧内主题帖话题的自动分类，和对于emoji指代词汇的推测分析等。而更加深入的模型分析计划，比如对于发言的LDA文本主题分类，也由于时间界限而取消了。同时，对各种指标的可视化分析要点太散，不够集中，应当将以上的各种指标综合起来，试图与用户的发言内容相关联，对百度贴吧中的内容作出更加有价值的分析。

# 五、 后续工作

在本课题的研究和实验过程中，我阅读了大量的文稿和资料，复习了pandas分析数据的方法和构建基本模型的方法，巩固了对于数据科学研究过程以及各项要点的掌握，学习了pyecharts这个高级的可视化工具的应用，进行了一次简单的数据科学过程的实践。在此过程中，我深刻发现知识掌握得不够牢固，应当在后续的学习中，进一步在实践中熟练掌握上述技能，并钻研各种机器学习模型的原理和实现方法，以实现更好的数据分析成果。

同时我将实现代码进行简单的整理，方便进行阅读。

# 六、 实验结论

通过对百度贴吧的数据进行收集、探索性分析以及数据建模等一系列数据科学过程，我们可以得出下述结论：

1. 百度贴吧的活跃用户多集中在东南部沿海地区，其中因话题的不同而具有较为明显的南北差异。
2. 在删帖率方面，技术类的帖子更容易保留其大多数内容，而娱乐类帖子删帖频繁，信息缺损大。
3. 性别比例方面，百度贴吧的用户约90%为男性，而女性感兴趣的话题中，性别比例会大有不同。
4. 考虑到百度贴吧曾在2013-2015年期间格外兴盛，建立较早的贴吧用户吧龄集中在8-9年左右。
5. 楼主发言比例方面，讲述类帖子比讨论类具有更大的楼主发言占比。
6. 用户在吧中的等级与其发言数大致呈指数关系，等级分布比例与吧的内容和话题兴起时间有关。
7. 百度贴吧的发言活跃时间大致与城市居民作息时间一致，在深夜和清晨有较少的发言量。
8. 每个贴吧的关键词汇根据话题的不同具有很大的差异性，而使用emoji表情种类则十分一致，只在数量上有较大区别。
9. 百度贴吧的帖子热度大致呈现“升温快、降温慢”的趋势，并且长时间未被回复的帖子会被清除。
10. 在百度贴吧中，接近大众爱好圈子的话题通常是游戏类、吐槽类，而贴吧用户的其他关注通常被归为小众爱好。
11. 百度贴吧中的发言大多数以负面情绪的表达为主，是网友们发泄负面情绪的平台。