老师同学们大家好，我研究项目的主题是 百度贴吧的数据挖掘与分析。

按照数据科学过程的划分，我的项目实践内容包含以下五个部分，

首先是项目背景与研究目标。

百度贴吧呢，相信大家都有所了解，是一个有些过气的平台了，但是这张图是百度贴吧首页上比较热门的几个吧，可以看到发帖量和关注量还是比较可观的。我选择这个题目的原因呢，主要是曾经比较喜欢使用这个平台吧，然后想要借助这个机会，从数据的角度对这个平台的内容和用户信息等，有一个宏观的理解。

再来说说如何获取百度贴吧的数据。

首先我是用request库爬取网页信息，但是这种方法有一个问题，每过大概半个小时就会弹出一个乱码的网页，其实是加载不出来的验证码，实际上这个验证码长这样。

一开始是通过更新这一大长串的cookie，但是往后做了一些改进，用selenium无头浏览器访问，在网上下载了一个模型可以识别这个验证码需要旋转的角度，再让浏览器自动操作，就可以破解这个验证码了。

然后介绍一下我爬取的数据，主要包含这四部分。

爬取的吧是从首页上选取的几个热门的娱乐类型的，也是大家比较熟悉的。

数据预处理方面，主要是去除空值、去除重复项，然后提取出需要的表项。这些内容去重过后大概有100兆。

然后，我对处理好的数据进行了以下分析

首先按照ip地址，去除重复发言的用户，画出如下省区分布图，可以看到孙吧的用户分布基本符合我国的人口密度分布，但是bilibili吧明显南方用户占比更高。这个条形图可以更直观反映这一点。

接下来这张图反映每个省用户的发言活跃度，依旧是这两个吧做一下对比，总体来看bilibili吧倾向于一个用户说好几句，同时也是bilibili吧的南方用户更加活跃一点。

这张图统计了每个楼层的删帖率，反映了每个吧的删帖率的不同，可以看到还是有很大区分的，然后这个平均删帖率也可以有一个很直观的发现，娱乐性的吧，会有冲动的发言甚至骂战，然后被屏蔽或者用户事后删掉，而相对学术性的比如图拉丁吧，就会有更少的删帖率。

这张图是统计了四个吧的性别比例，可以看到除了剑网三男女比例大致相同以外，另外三个吧男性占比相当大。

这里对比了孙笑川吧和剑网三吧的吧龄比例，可以看到在孙吧发言的用户更多是最近注册贴吧的，在剑三吧发言的则是一些老用户，这个箱线图也能很直观的表现这一点

然后这个是楼主发言的占比，其中王者荣耀吧、图拉丁吧这种以楼主讲述，吧友评论为主的就会楼主发言占比高一点，另外三个吧呢就是吧友讨论为主的楼主占比就会低一点。

这个是bilibili吧用户的等级与发言数的关系，由于贴吧是发言数一定程度上决定用来升级的经验值，所以发言数多的人等级更高，同时对比下bilibili吧和孙吧的等级比例，可以看到世界杯吧发言的有很多一级用户，而孙笑川吧等级分布更加均匀一点，说明用户粘性较大，用户群体相对固定。

这张图是主题帖的日期分布，由于每个吧都有清理坟贴也就是长时间没人回复帖子的习惯，所以以前发的帖子数据缺失，此外能看出周末的发帖数会更多一些。

通过这个活跃时段的折线图可以看到网友们很少在早上和上午发言，而除了王者荣耀吧和世界杯吧以外，其他三个吧活跃时段都比较一致。

这是一个关于常用表情的统计，而蓝色的部分是最近一日用户使表情的数量，可以看出，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。

然后这是根据表情的使用频率做的一个词云图片。

这边用jieba分词做了，选取发言中的名词做了一个词云图片，可以看出它们很好的反映了每个吧的话题和流行词。

这个是一个相关系数矩阵的可视化，用来反映用户各指标之间的关系，只是这里相关性不高，就没有做进一步研究。

这边选取了两个时间跨度比较大的帖子，分析了一下每15分钟内发言数的变化，帖子的热度变化受很多因素影响，由于回复帖子可以把它顶到用户可以最先看到的位置，所以帖子都符合一个升温快，降温慢的规律。

同时我还做了累计发言数的趋势图

在这部分的最后，我根据共同关注的用户以及用户在吧中的等级来判断两个吧之间的关联程度，做了所有吧之间的力导向关系图。由于节点量太大，这里只保留了几个发帖数超过一千万的吧，节点大小由发帖数决定。在这张图上大家可以很明显看出自己感兴趣的领域的吧属于大众还是小众。

接下来讲一下我基于jieba分词做的很不成熟的一个无监督学习模型，可以对吧内的文本进行情感评分。首先将句子分成词，根据已有的词典对词汇进行打分，句子如果是正向情感分值为正，负面情感为负，然后根据句子的得分将句子中不在词典里的词加入词典并赋相应权值。最后储存训练后的词典。

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。

这是孙笑川吧用户发言的情感评分分布图，可以看到吐槽类的负面情感占大多数。

我的演讲到此结束，谢谢。