林大集市发帖话题类型调查

一、 背景与研究意义

“集市”微信小程序是一款主打校园论坛的平台，自其创立之初便与各大高校合作，利用学校系统构建了一个以“校园、友爱、互助”为核心主题的半匿名校园论坛。这里的“半匿名”并非放任自由，而是需要用户注册时提供学号等学校认可信息。如今，“集市”已成为国内最大的校园论坛之一，为同一高校的用户提供了一个交流分享的平台。在“集市”微信小程序中，帖子仅对同校用户可见，内容大致分为五个类别，涵盖了二手闲置、求助打听、恋爱交友、兼职招聘以及校园招聘等多个方面。

随着时代背景的不断变化，大学生的日常生活也发生了较大的转变。考虑到就业环境的变化，有相当一部分学生开始倾向于选择稳定的就业道路，如考取公职或者选择深造考研。因此，学习成为了他们生活的重心。

与此同时，现代社会提供了更多样化的娱乐途径，也有一部分学生更倾向于在大学时代尽情享受生活。这种转变使得大学生的生活变得多姿多彩，融合了学习与娱乐，使得校园成为了一个充满活力和活跃氛围的社交平台。

对于东北林业大学的老师、辅导员以及其他非学生身份的群体来说，他们的工作与学生息息相关，因此了解学生的生活和需求变化对于他们的工作至关重要。这项调查将帮助他们更好地了解学生们的兴趣爱好、关注的问题以及在校园中所面临的挑战。同时，通过对话题类型的调查，也能探索学生们对于不同话题的关注程度，从而为学校提供更有针对性的支持和指导，促进学生全面发展和健康成长。调查结果将被用于为学校提供更贴近学生需求的服务和支持，促进学生的综合发展和健康成长。

二、 目标与过程

课题目标是对从 2023 年 9 月 1 日起，至 10 月 20 日林大集市所发的所有帖子的话题类型进行研究调查。

受限于技术手段和机器处理算力，帖子的话题类型分布将仅从帖子的标题进行判断。集市针对每个高校会设置多个微信群聊，群聊机器人定时抓取集市上所发布的帖子并推送至群聊内。借助这个机制可降低获取帖子标题的难度。

在研究调查的初期阶段，笔者将对每个帖子的标题进行收集和整理。接下来，笔者运用自然语言处理技术，包括文本分析算法，对标题内容进行过滤，从而识别出每个帖子标题中的核心词汇。利用核心词汇能够快速准确地将帖子归类为特定的话题类型，从而为后续的深入研究奠定基础。

研究过程中，笔者觉得最难的部分在于繁琐的数据收集。从开题至 10 月 20 日的每日都要将当日群聊内所有的聊天消息逐条复制进 Excel 表格中，倘若某日忘记处理当日

的数据，则下一日要处理的数据量则要成倍增加，而尽管单纯的复制与粘贴看起来很简单，但这种重复乏味的工作也非常非常无聊，极大地考验一个人的耐心和持之以恒的毅力。

研究过程中也曾遇见过几次困难。研究初期，笔者打算利用爬虫技术实现集市帖子内容的自动抓取，然而在自学过几天后笔者发现，腾讯对小程序的消息内容保护的非常严密，想要短时间内学会，考虑到笔者本人的其他学业，这几乎不可能实现，研究进度一度被阻塞，差点令笔者想要放弃。后来，笔者打算使用自学的机器学习算法对所获取的文本进行聚类，从而将帖子类型从总体的角度上进行分类。然而机器学习的效果并不好，因为机器学习算法对分析文本的数据长度有要求，大致要求长度范围在几十至几百之间，短了则少了很多信息，长了则多了很多冗余信息。然而帖子标题的长度大概也就十个字以内，远远短于最低要求。导致聚类效果不佳。为处理长度问题导致的聚类失败，笔者考虑过将发布时间相同的帖子标题文本进行合并再用机器学习算法进行聚类，然而冒然地对数据进行合并所带来的问题就在于文本内容的倾向性被泛化，从而导致文本的分类模糊，难以进行聚类分离。这对笔者的打击很大，因为在研究课题创立初期便构建了以自身所学的机器学习算法的课题研究模型大部分都失败了，研究过程一度受阻。

尽管数据的收集看似枯燥无味，但笔者意识到这项工作并非毫无用处。经过反复思考和对数据的观察分析，笔者发现大部分集市帖子的关键信息往往都隐含在标题中。基于这一发现，并借鉴了之前所学的数据分析知识，笔者提出了一种新的处理思路。即将帖子标题中的关键词作为其所属话题的代表，通过对这些关键词的统计和分析，来揭示帖子的话题核心。

然而，这种处理方法也带来了另一个问题，即一个帖子可能会涉及多个关键词，导致其分类不够明确。笔者深思熟虑后认为，这样的结果反而更加符合实际情况。在实际生活中，一个帖子的分类往往不可能单纯属于一个集合，而是可能涉及多个分类的交集。因此，通过引入多个关键词的概念，可以更准确地描述帖子的内容和特征，使分类结果更加符合实际情况。

整个课题的研究调查过程大致如下：

①从 2023 年 9 月 1 日起，至 10 月 20 日每日晚笔者手动从微信群“校园集市林大站”中，将当日群聊内所发布的所有聊天消息（如图 1）以发布时间为索引复制进 Excel 表格中（如图 2）。共计抓取 4757 条聊天消息。

图 **1**　群聊截图

	A	B	C
1	消息时间	消息内容	
2	2023/9/1 0:01	收仓鼠粮 https://c.zanao.com/l/1SnsEt 溪溪状元笔记不好意思刚刚留错了联系方式 https://c.zanao.com/l/1SnsKe 高数A2补考 https://c.zanao.com/l/1SnteK 求问心理问题 https://c.zanao.com/l/1Snth9 求问求问 https://c.zanao.com/l/1Snthq	
3	2023/9/1 0:14	想收个二手自行车 https://c.zanao.com/l/1SntE4 英**留学费用 https://c.zanao.com/l/1SntL0	
		卡帕床垫 https://c.zanao.com/l/1Snueg 找羽球搭子 https://c.zanao.com/l/1SnufO 6号楼怎么晒被子	

图 **2**　群聊消息抓取截图

②10 月 20 日晚，笔者将抓取的所有信息使用 Python 进行处理，删去其中的错误项

和无效项，再删去每条聊天记录的网址，并将每条聊天记录的多个帖子分离，存进新的 Excel 表格中（如图 3）。共计获取帖子数 38520 条。

	A	B
1	发帖时间	发帖内容
2	2023/9/1 0:01	收仓鼠粮
3	2023/9/1 0:01	溪溪状元笔记不好意思刚刚留错了联系方式
4	2023/9/1 0:01	高数A2补考
5	2023/9/1 0:01	求问心理问题
6	2023/9/1 0:01	求问求问
7	2023/9/1 0:14	想收个二手自行车
8	2023/9/1 0:14	英**留学费用
9	2023/9/1 0:32	卡帕床垫
10	2023/9/1 0:32	找羽球搭子
11	2023/9/1 0:32	6号楼怎么晒被子
12	2023/9/1 0:32	请问八公寓上铺的遮光帘尺寸
13	2023/9/1 0:32	转专业求助
14	2023/9/1 0:32	转专业求助
15	2023/9/1 0:52	22学生公寓限电功率多少
16	2023/9/1 0:52	家人们跟我聊聊天

图 **3**　分离结果

③用 Python 对每个帖子的标题进行分词处理，去除长度为 1 的词汇，并进行词频统计，最终可获得所有词汇的出现频数（如图 4）。共计出现的词汇数为 15448，删去频数低的无效词汇，留取核心词汇 205 个。

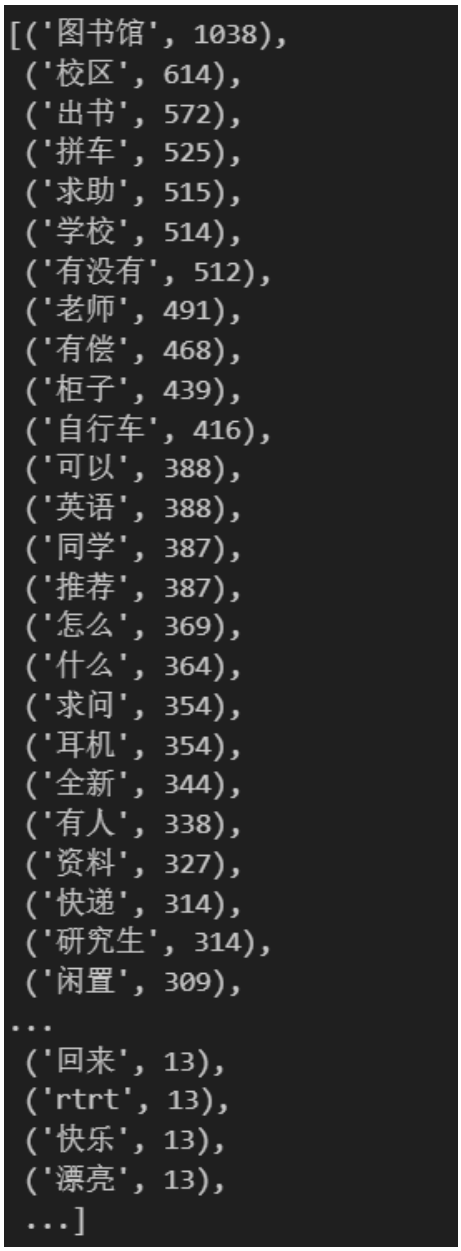

图 **4**　词频

三、 课题研究成果

将所得的 205 个核心词汇及其频数，使用 Python 绘制词云图如图 5：

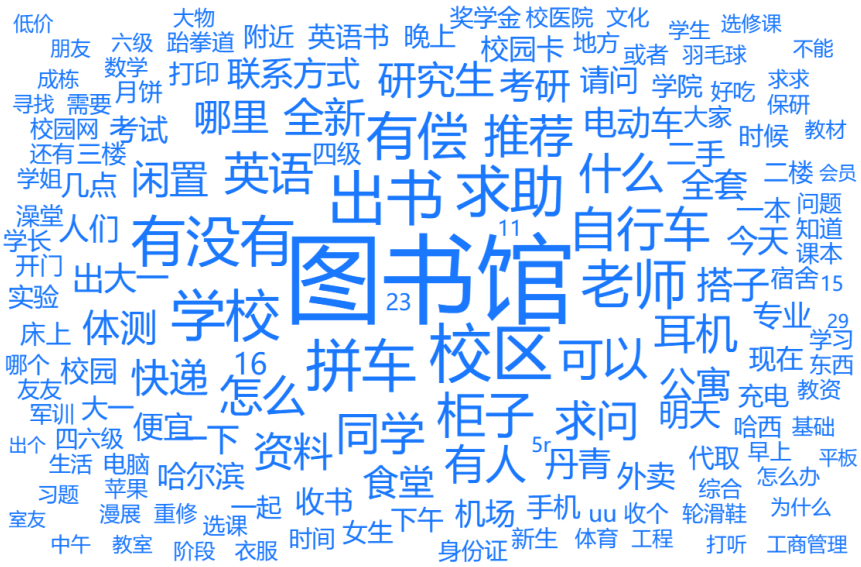

图 **5**　词云图

同时绘制对频数高的部分词汇绘制统计图如图 6：

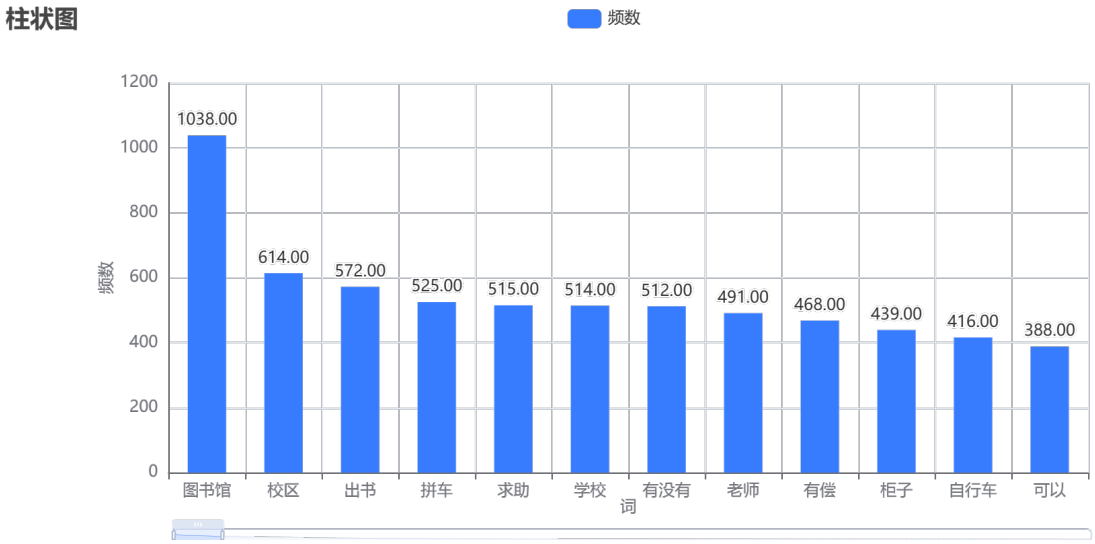

图 **6**　柱状图

容易知道，核心话题“图书馆”出现次数 1038 次，几乎接近排位第二的核心话题“校区”的 614 次的两倍。而“图书馆”这个词则往往与学生的考研、考公、考编等行为挂钩，另外“考研”话题的出现频数也不低，共出现 270 次，排位 31。可见，目前的大学生生活中，考研一事占有较大比重。再考虑到会讨论考研的学生多为大三大四，且目前距离考研时间已近，笔者认为东北林业大学的大三生大四生中考研这一话题以及相关话题热度会更高。另外，“柜子”话题出现的频数也不低，出现 439 次，排位 10。观察

林大集市的相关话题帖可知，部分学生对图书馆柜子被长期占用的情况持有不满意见。

笔者建议学校在面对考研等事宜上可以为大学生提供相关的帮助，例如：为图书馆设置更多的柜子和座位、在图书馆附近设置一流动食堂、以及在图书馆中提供相关考研书籍等，从而优化学生的考研体验。

另外，还存在部分话题仅在一段时间内大量流行，如“拼车”、“体测”和“军训”，频数分别为 525、298 和 106。“拼车”话题在开学初被大量讨论，而本次课题收集数据的开始日期是 9 月 1 日，这个时间基本上部分专业已经是开学后，可见倘若把收集数据的时间点再往前十日左右，“拼车”话题的频数将有可能一骑绝尘。“体测”话题的频数高是由于十月初的体测，“军训”话题的频数则是大一新生开学的集体军训时的讨论。值得一提的是，在撰写课题的这个时间点，搜索“军训”关键词，帖子数目仅有两条，而搜索“体测”关键词，帖子数目仍然很多（大部分是在焦虑体测不合格会不会影响毕业和保研），话题的时效性可见一斑。

学校的一些地点也经常作为话题出现在集市中，例如：“公寓”278 次、“食堂”274 次、“丹青”240 次、“宿舍”130 次、“澡堂”124 次、“校医院”108 次、“成栋”85 次。这些地点与大学生的日常生活息息相关，在“丹青”的话题中，帖子往往是求助类型，比如“丹青 6 楼中间男厕第一间求纸”、“丹青找卡紫色库洛米卡套”。“食堂”话题中，帖子往往是询问有没有好吃的餐饮，比如“新食堂有没有什么好吃的面条米粉米线啊”。“澡堂”话题中，帖子往往是询问澡堂的开门时间和洗澡的人多不多。

四、 总结与展望

课题的创立之初的目的，在于通过调查统计林大集市发帖的话题类型，从而帮助非林大学生的人了解林大学生们在大学生活中更关心更在意的事物。笔者在这次课题研究中，最大的收获便是也帮助自己了解同龄人的想法。虽然笔者也是大学生中的一员，但正所谓“不识庐山真面目，只缘身在此山中”，即使笔者是大学生，也仅仅了解一些与笔者类似的学生的想法，然而大学的自由给予了大学生多样性，同学们在进入大学后往往朝着不同方向发展，笔者仿佛脱离了主流很久，在这次研究中也算是短暂的回归了“社会”了。

笔者希望，倘若再给我做一次这次课题的机会，我将会优化数据的收集类型，不再只收集发帖的标题，更包括发帖的内容。可能会很累，所以可以适当的把收集时间从五十天减少到十天。这样一来，拥有了更多的文本数据，更长的数据长度，从而可以使用让课题研究与自身专业技能结合，使用机器学习算法中的文本聚类算法对帖子进行自动分类。

本次课题研究的所有相关资料可以在以下网址中找到：

https://github.com/KomorebiCN/Nefu-Research