基于大数据分析的

中小学在线教育平台基本情况的研究

——以“乐乐课堂作文库”为例

范欣妍 2016202206

摘要 本文从大数据的概念及教育大数据分析的用途入手，结合大数据分析的过程，分析了中小学在线教育平台的基本情况。本文以爬取的“乐乐课堂作文库”中的话题作文为基础数据，分析了乐乐课堂用户和收录文章的基本属性，为该网站的作文库建立了基本的画像。同时使用机器学习中一些方法，为网站的功能建设提供了多种方案。并根据分析结果对乐乐课堂及在线教育平台未来的发展提出了一些参考建议，以便其更好地适应大数据时代的发展要求。

关键词 大数据 在线教育平台 作文库 可视化 TF-IDF Doc2Vec

目录

引言

• 研究背景

为深入贯彻落实党的十九大精神，加快教育现代化和教育强国建设，推进新时代教育信息化发展，培育创新驱动发展新引擎，结合国家“互联网+”、大数据、新一代人工智能等重大战略安排，国家推出《国家中长期教育改革和发展规划纲要（2010—2020 年）》《国家教育事业发展“十三五”规划》《教育信息化十年发展规划（2011—2020 年）》《教育信息化“十三五”规划》等文件。“互联网+ 教育”则变成了愈演愈热的在线教育。

在如今的大数据时代背景下，人们开始更加关注与认可在线教育，于是也研发出越来越多的在线学习平台。通过诸多平台能够帮助人们接触更多的数据，对于大量数据的处理也成为人们普遍关注的问题。因此，在大数据时代背景下，通过对在线教育平台大数据的科学、合理运用，可以充分探究在线教育的发展规律与潜在问题，为未来在线教育的发展提供借鉴。

• 研究意义

通过对教育大数据的收集、归纳、整理，以及对繁杂的教育数据的分析，能发现其相关关系，诊断现存问题，预测发展趋势。对于学习者而言，可以更加清楚地了解自己的学习行为、学习方式、对学习内容的偏好、学习的掌握程度，从而制订出更适应自身学习的个性化的学习规划，实现个性化学习，并对未来发展做出一定的诊断，提升教育质量。对于教师而言，通过教育大数据分析，可以发现最优的教师，教师也可以制订个性化学习方案，规划更优的教学路径，设计更适合学习者的学习互动与创设场景，优化教育资源配置，使教育决策更具有科学性。

乐乐课堂是一个基于互联网/移动互联网平台的 K12 模式的中小学个性化学习网站，由来自北大、清华毕业的多年资深教育讲师及前数学奥赛金牌选手和高考状元等发起，于 2014 年正式成立。作为新兴的在线教育网站，乐乐课堂课堂一定有其成功的秘诀。本文将通过大数据分析的方法，对收集的乐乐课堂作文库的数据进行深入挖掘，探究其发展的规律，分析其存在的问题，希望能为网站管理者提供更好的建设方案，为网站使用者提供更佳的学习上的帮助。

- **研究思路**

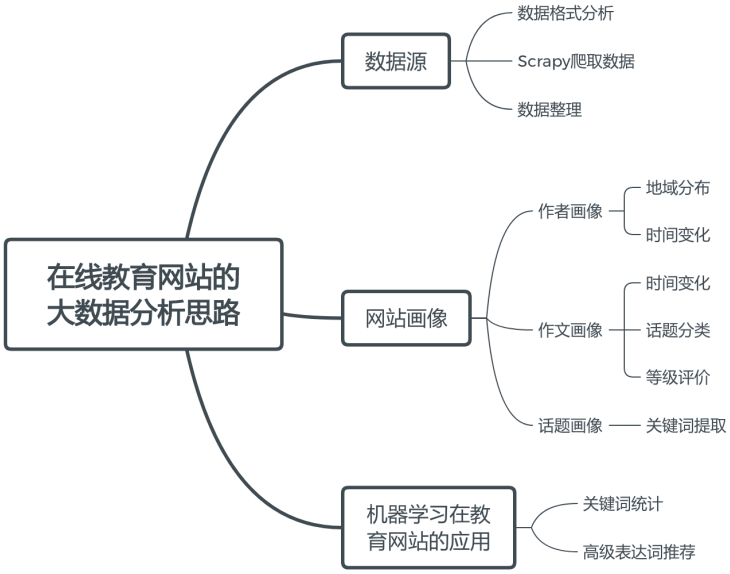

图 0.1 分析思路

1 以“乐乐课堂作文库”为例的大数据分析

1.1 数据预处理

1.1.1 Scrapy 爬取数据

（1）数据源介绍

图 1.1.1 显示了乐乐课堂网站主页，其中包括“天天练”“乐学堂”“乐乐题库”“作文库”“成语大全”几大模块，但有些模块的数据是视频的形式，比较难爬取；有些是不规则的考题的形式，不易获得用户和网站的相关信息，且进行文本分析较困难。综合以上情况，最终选取“作文库”作为数据源。

图 1.1.1 乐乐课堂主页

作文库具有数据量大、数据格式整齐、分类详细等优点（如图 1.1.2）。且每篇作文中包含作者和网站丰富的信息：标题、时间、作者、学校、正文、评论、标签等（如图 1.1.3 ），通过分析这些数据，可以获得网站的使用情况和用户的基本信息，从而进一步对该在线教育平台的基本情况进行分析。

乐乐课堂作文库连接：http://www.leleketang.com/zuowen/clist0-0-0-1-1.shtml

图 1.1.2 作文库主页

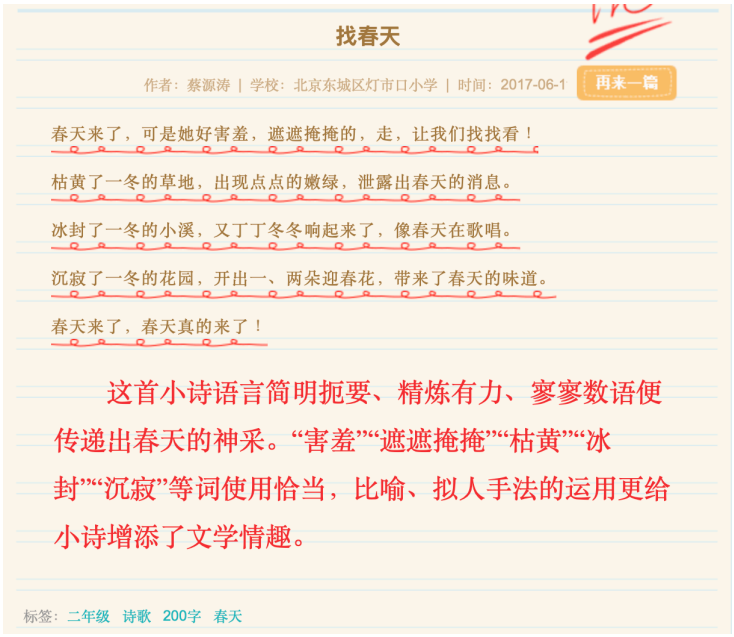

图 1.1.3 每篇作文的页面

（2）核心代码

见代码 1.

代码 1 - 爬虫核心代码

```
#对于一个作文，提取其中的标题，作者，正文，评论等
def parseNews(self, response):
    title = response.css(".cp_htitle::text").extract_first()
    author = response.css(".cp_author ::text").extract_first()
    content = response.css(".cp_content p::text").extract()
    content_ = ''.join(content)
    comment =   response.css(".cp_comment p::text").extract_first()
    yield{"title":title,"author":author,"content":content_,"comment":comment}

def parse(self, response):
    newslist=response.css("div.item div.item_title.clearfix a")#获取作文的列表
    for news in newslist:
        url=news.css("a ::attr('href')").extract_first()#抽取每个作文的链接
        yield response.follow(url, callback=self.parseNews)
    nextpage=response.css(".p_pager_web a.p_next ::attr('href')").extract_first()
    if nextpage is not None:#跳转下一个页面
        yield response.follow(nextpage, callback=self.parse)
```

1.1.2 数据清洗

将爬取后的数据进行清洗、整理，分为三个基本文件“author.csv” “content.csv” “comment.csv”，分别包含作者基本信息、作文标签和正文内容以及作文评语，方便后续的处理和分析。

最终爬取了 15 万余条作文，时间跨度为 2013 年-2018 年。

1.2 基本属性的统计、可视化与分析

pyecharts 是一个用于生成 Echarts 图表的类库。Echarts 是百度开源的一个数据可视化 JS 库。用 Echarts 生成的图可视化效果非常棒，pyecharts 是为了与 Python 进行对接，方便在 Python 中直接使用数据生成图。

本节主要介绍如何在原数据中获取用户和作文的相应信息以及如何使用 pyecharts 对“乐乐课堂作文库”的基本情况进行可视化分析与展示。

1.2.1 用户

（1）地域分析

在原数据集中可以看到，“学校”这一属性中大部分数据都包括省份、市县名或区名的信息。因此可以对该数据进行进一步挖掘，获取文章和作者所在地理位置信息。对于同一个地区，统计该地区作文的数量，在地图上进行热力图展示，直观地分析该网站用户在全国的分布和网站在不同地区的使用情况。

①预处理

利用 jieba 分词中对地名的词性标注，筛选出“学校”这一信息中所有的地名。由于对地名的标注无法区分省份名和城市名，因此预备了中国 34 个省份名的文件和 616 个城市名文件。

由于 jieba 分词得到的地名中不会去掉“省”“市”“县”这几个关键词，从而会得到“山东”和“山东省”这两个不同的词条。因此在预处理中去掉地名中的“省”“市”“县”，得到一致的数据，方便进一步的统计分析。

②城市与省份分布频率统计

I. 频率统计

首先读入省份名称和城市名称并构建“省份-频率”字典和“城市-频率”字典。接着遍历所有学校，切词并找出省份名和城市名，进行词频统计。

II. 核心代码

详见代码 2.

代码 2 - 省份和城市作文数量的统计

```
keyword=jieba.posseg.lcut(line[3])#jieba 切词
    for word in keyword:
        if word.flag == 'ns':#词性标注，筛选地名
            w=word.word
            if w.find('省')!=-1 or w.find('市')!=-1 or w.find('县')!=-1:#去掉最后的省、市、
县，方便判断
                w=w[:-1]
            if w in pro_list:#更新省份的值
                pro_num[w] += 1
            elif w in city_list:#更新城市的值
                city_num[w] += 1
```

③作文数量城市与省份分布可视化

利用 pyecharts 进行可视化展示时需要传入两个主要参数：attr 和 value。其中 attr 表示省份/城市名称的列表，value 为每一个省份/城市对应的作文数量。注意绘制省份热力图需要用到 Map 模块，绘制城市散点热力图需要用到 Geo 模块。

具体代码请看代码 3。

代码 3 - 作文数量的省份分布热力图和作文数量的城市分布散点图

```
#作文数量的省份分布热力图
from pyecharts import Map
map = Map("作文数量的省份分布热力图",
          title_color="#fff",title_pos="center",width=1200,
          height=600,background_color='#404a59')#构建 Map 对象，设置画布大小颜色标题等
map.add("", attr, value,visual_range=[0, 3000],
        maptype='china',is_map_symbol_show=False,symbol_size=10,is_visualmap=True, visual_text_color='#fff')#传入省份和数量的信息，地图类型为中国地图
map.render('province_picture.html')#输出结果，为 html 格式，可用浏览器打开
#作文数量的城市分布散点图
from pyecharts import Geo
geo=Geo("作文数量城市分布散点图","data of
          articles",title_color="#fff",title_pos="center",width=1000,height=600,background_color='#404a59')#构建 Geo 对象，设置画布大小颜色标题等
geo.add("",attr,value,visual_range=[0,1000],maptype='china',visual_text_color="#fff",symbol_size=10,is_visualmap=True)#传入省份和数量的信息，地图类型为中国地图
geo.render("city_picture.html")#输出结果，为 html 格式，可用浏览器打开
```

可视化结果如图 1.2.1 和 1.2.2 所示。图 1.2.1 展示的是作文数量在 0-1000 范围内的城市分布散点图，图 1.2.2 展示的是作文数量在 0-3000 范围内的省份分布热力图。在两张图中，均为颜色越偏向暖色系代表作文数量越高，颜色越偏向冷色系代表作文数量越低。

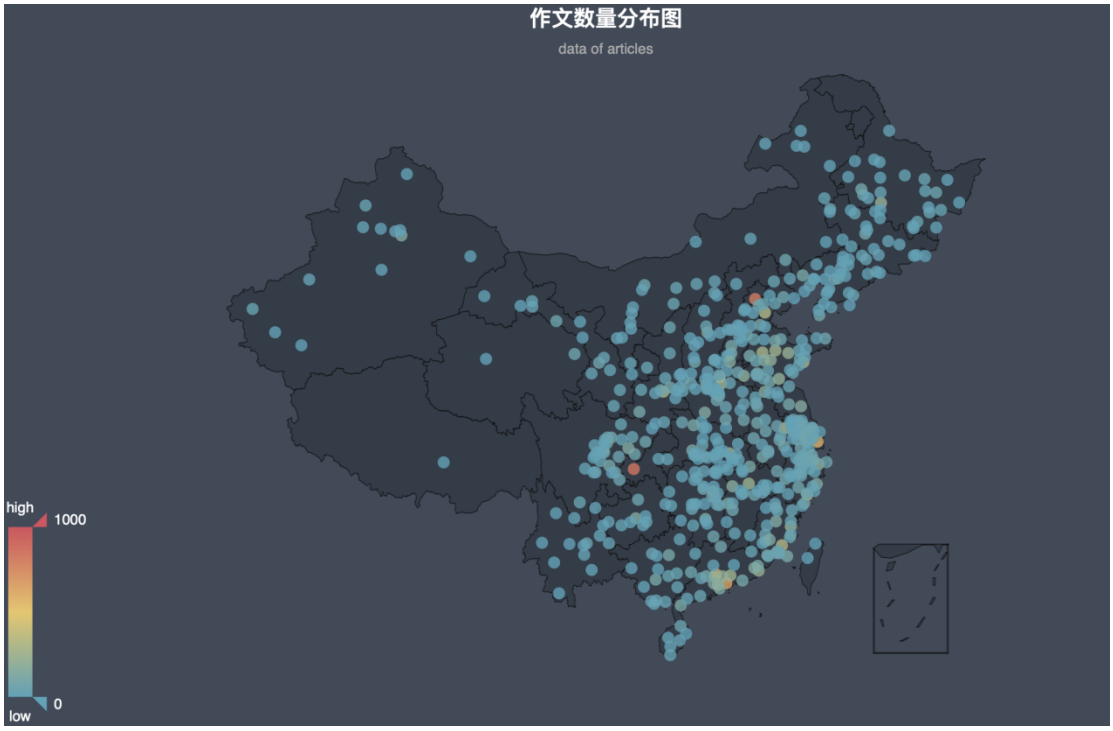

图 1.2.1 作文数量的城市分布散点图

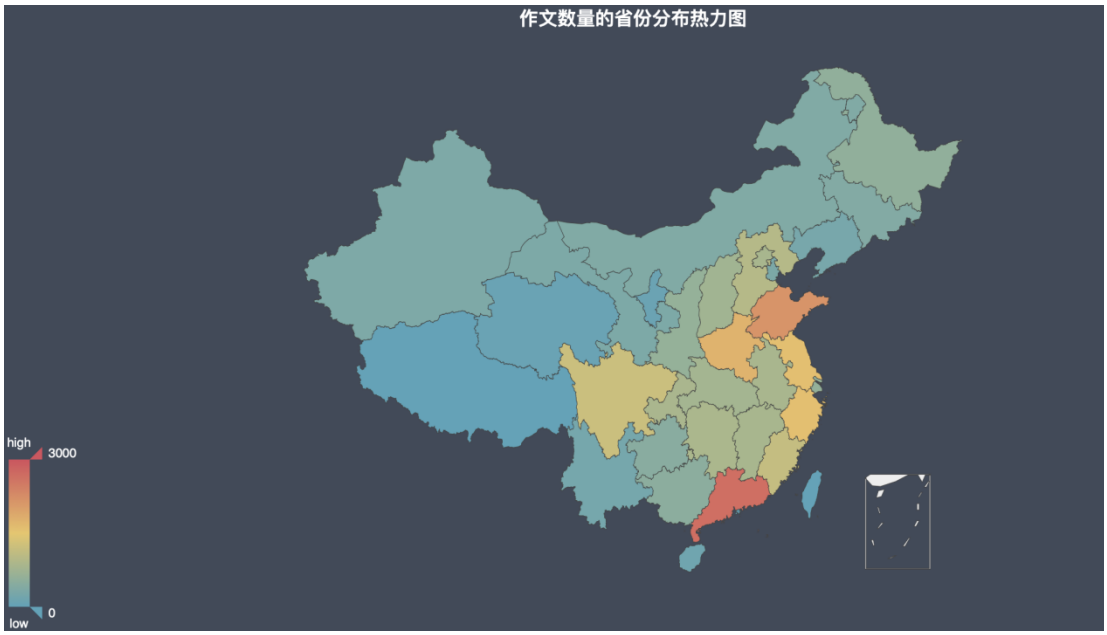

图 1.2.2 作文数量的省份分布热力图

④结果分析

在地域上，受到经济发展的限制，沿海及经济发展较快的省份更关注在线学习，愿意花费时间、金钱在在线教育上，因此该网站使用人数较多的省份有山东、江苏、浙江、福建和广东。相比较而言，西部和经济条件落后的省份如西藏、青海、新疆等，对在线教育的关注

较少，花费在在线学习的时间少。

当然，人口基数也是影响网站使用人数的一个重要原因。众所周知，河南、山东、四川等地是有名的人口大省，因此相对于其他省份，这些地方的在线网站使用者数量相对较高。

另外，由于北京、上海、广东、浙江等互联网发达省市对 IT 互联网人才的需求较高，这些地域学习该领域课程的用户远高于其他地区，对互联网类课程的关注都比较高。

（2）时间分析

原数据集中包括“时间（年-月-日）”“作者姓名”这些属性。因此可以利用这两类数据进一步挖掘获取某个时间段作文的数量和作者的数量。考虑到精确性，将时间段间隔设为一个月。对于同一年同一个月，统计该时间段的作文、作者的数量，以时间为变量，数量为因变量绘制“数量-时间”变化折线统计图，直观地分析该网站从成立初到现在的发展情况。

①预处理

构建嵌套字典，形如{‘2013’: {‘01’: [‘A’, ‘B’ ...] } }，表示每年每月的作者列表，用来记录每个月中作者的数量（去重）。

构建 pandas 中的 DataFrame 结构，列索引为'article_number'、'author_number'，表示作文数量和作者数量；行索引利用多重索引结构，一级索引为年，二级索引为月，初始所有数量为 0。具体构建代码见代码 4.

代码 4 - 构建多重行索引 DataFrame 来统计每个月作文和作者数量

```
1  from pandas import Series,DataFrame
2  import pandas as pd
3
4  year=['2013','2014','2015','2016','2017','2018']
5  month=['01','02','03','04','05','06','07','08','09','10','11','12']
6  date_df=DataFrame(np.random.randint(0,1,size=(72,2)),columns=['article_number','author
7              _number'],index=pd.MultiIndex.from_product([year,month]))
```

②作者、作文数量随月份变化

I. 频率统计

遍历 author.csv，获取每一个作文的年和月。对于作文数量，直接在相应的年和月的作文数量值加 1 即可；对于作者数量，需要维护一个每年每月的作者列表，如果是新作者，则作者数量加 1 并加入对应列表中，如果作者已经存在，则跳过。

II. 核心代码

代码见代码 5。

代码 5 - 每年每月作文和作者数量的统计

```
date=line[index_date]
index1 = date[:4]#获取年份
index2 = date[5:7]#获取月份
date_df['article_number'][index1][index2] += 1#作文数量+1
if line[2] not in author_dict[index1][index2]:#判断是否为新作者
    author_dict[index1][index2].append(line[2])#更新作者列表
    date_df['author_number'][index1][index2] += 1#作者数量+1
```

③可视化

折线图主要用到 pyecharts 中的 Line 模块，核心参数为 attr 和 value。其中 attr 表示横坐标的值，即日期；value 表示纵坐标的值，即每年每月的作文/作者数量。绘制多条折线图在一幅图中只需不断向构建好的 Line 对象 add 所需要的线即可。

绘制双折线图的代码见代码 6。

代码 6 - 绘制作文数量和作者数量随月份变化的双折线统计图

```
attr = l_time#日期，格式为“年/月”
v1 = article#作文数量
v2 = authors#作者数量
line = Line("作文数量与作者数量随月份变化折线图",width=1200,height=700)#构建 Line
        对象，设置画布大小
line.add("作文数量", attr, v1, mark_point=["max","average"])#画出作文数量的折线图，显
        示最大值和平均值
line.add("作者数量", attr, v2, is_smooth=True, mark_line=["max", "average"])#画出作者数
        量的折线图，显示最大值和平均值
line.render('two_line.html')#输出结果，为 html 格式，可用浏览器打开
```

可视化结果如图 1.2.3 所示。其中红色的线表示作文数量随时间的变化，蓝色的线表示作者数量随时间的变化。

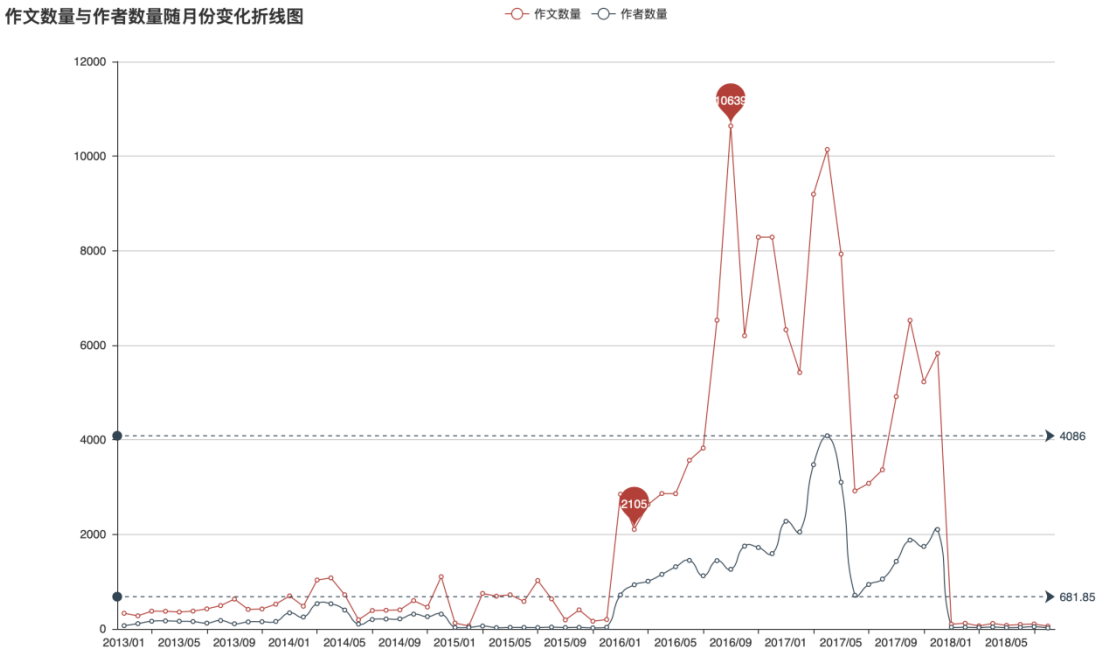

图 **1.2.3** 作文数量和作者数量随月份变化折线图

④结果分析

I. 基本信息

根据图 1.2.3，可以得到乐乐课堂作文库的作文和作者数量的一些基本情况，如下表 1.2.1。

表 **1.2.1** 作文和作者数量统计基本情况

	作文情况	作者情况
单月最大数量	10638	4086
最大数量的月份	2016-11	2017-06
月平均数量	2105	861.85

II. 变化趋势的分析

从图 1.2.3 中可以看出，在 2016 年之前，乐乐课堂作文库中收录的文章数量和用户数量平稳发展，没有太大变化。在 2016 年之后，作文数量迅猛增长，在 2016 年 11 月达到峰值。2016 年 6 月至 2017 年 6 月之间，网站里的单月作文数量维持在 8000 篇左右，处于较高的水平。但到 2017 年后半年作文数量开始呈现下降趋势。

从以上结果分析得到，由于在 2015 年乐乐课堂作文库刚推出不久，2013 年和 2014 年

的作文多为网站中别的模块收集的文章，因此数量较少；2015 年该模块推出后，网站用户数量逐渐增长，且开始收录更多的文章；在 2015 年末完成融资后，吸引了更多的用户，所以 2016 年以后网站作者大量增加，作文数量也有大幅度的增长；2018 年文章的数量之所以这么少，一方面可能由于在线教育平台越来越多，导致网站使用者在不断减少，另一方面可能由于网站更新滞后导致未能及时获得 2018 年的数据。

III. 作者和作文的相关性

从图中我们大致可以看出，作文数量多的时间作者数量也多，作文数量少的时间作者数量也少；且作文数量增减变化趋势与作者数量几乎一致，因此可以进一步探究二者之间的关系。

利用 Python 中的 numpy，计算每月作者数量和作文数量这两个序列之间的相关性。计算相关系数为 0.916548，二者之间有很强的正相关性。利用 matplotlib 包绘制出二者关系的散点图，如图 1.2.4 所示。可以看到作文数量和作者数量几乎呈线性增长的关系，由此可以得出结论：作者数量是影响作文数量变化的一个重要因素。

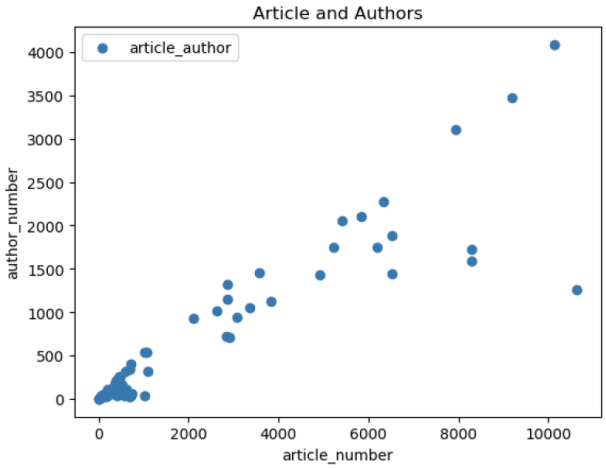

图 1.2.4 作文数量和作者数量散点图

1.2.2 作文

（1）作文按话题的分类统计

在原数据集中，每一篇作文有一个标签，表示所属的话题。通过分析每一类话题下的作文数量，可以得到该网站用户偏好于那一类文章。在一定程度上，这个结果也能反映目前中小学教育中流行的作文话题，为用户在未来的作文学习上提供借鉴。

本节利用 pyecharts 中的 Funnel 模块，绘制话题作文“漏斗图”，可以清晰直观地看出每个话题下作文的数量以及不同话题之间作文数量的对比关系。

①统计每类话题下作文的数量

构建“话题-作文数量”字典，遍历每一个话题，统计该话题的作文数量。核心代码见代码 7。

代码 7 - 每一个话题下作文数量的统计

```
tag=line[1][3:5]#得到话题标签
if tag not in tag_list:#判断是否为新的话题
    tag_list.append(tag)
    tag_dict[tag]=1
else:
    tag_dict[tag]+=1
```

②可视化

漏斗图主要用到 pyecharts 中的 Funnel 模块，核心参数为 attr 和 value。其中 attr 表示话题；value 表示每个话题下作文数量的值。

绘制漏斗图的代码见代码 8。

代码 8 - 绘制话题-数量漏斗图

```
from pyecharts import Funnel
attr = list(tag_dict.keys())
value = list(tag_dict.values())
funnel = Funnel("话题数量对比",width=1200,height=700)#构建 Funnel 对象，设置画布大小
funnel.add("作文话题", attr, value, is_label_show=True, label_pos="inside",
           label_text_color="#fff")#绘制漏斗图
funnel.render('tag_number.html')#输出结果，为 html 格式，可用浏览器打开
```

可视化结果如图 1.2.5 所示。用浏览器打开 html 文件后，鼠标移动到的位置可以显示该类话题的数量，鼠标点击最上边的小方块可以选择显示或不显示该类话题作文的数量。

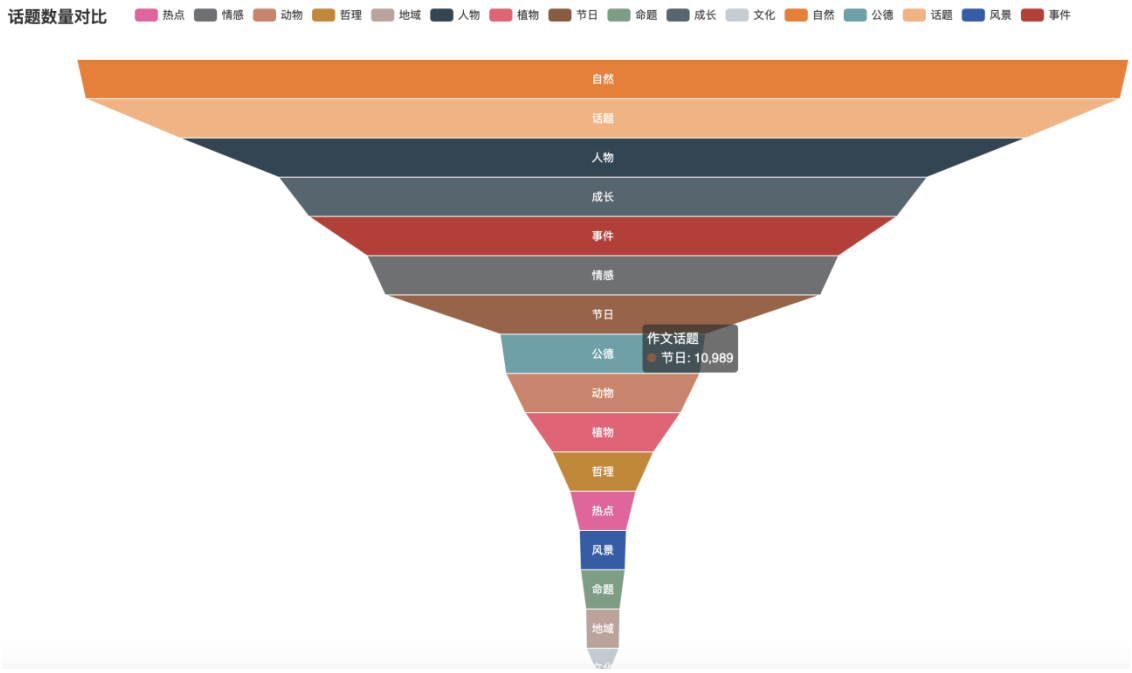

图 1.2.5 每个话题下作文数量漏斗图

③结果分析

根据图 1.2.5，首先可以直观看到作文数量最多的为“自然类”作文，“话题类”作文数量与之相当。其次，“人物类”“成长类”“事件类”“情感类”话题的作文数量位于其后，它们之间的数量也相差不多；这可能与话题本事比较相似有关系。数量最少的几个话题分别是“热点类”“风景类”“命题类”“地域类”“文化类”，它们的数量与最靠前的几个话题的数量有很明显的差距。

结合自身学习经历，可以推断，“自然”和“话题”应该属于不变的经典话题，尤其是“自然类”，对于初学作文的小学生来说是很好入手的话题，教师在教授时基本会选择以“自然”话题为基本出发点教育学生，因此这类话题的作文数量会很高。“话题”类作文在初中教育中是很常见的作文类型，无论在平时训练还是在考试中，都属于老师青睐的对象。话题可以类型有很多，但几乎都是和学生的学习和生活息息相关的，因此它也可以算作作文的经典。紧随其后的“人物”“成长”“事件”“情感”话题则与学生自身的成长有关，学生容易从这些话题入手，描绘一个人物，叙述一件事情，或者抒发自己的情感，都是写作文的经典角度。

（2）作文等级设定与统计

本节主要通过对每一篇作文给出一个量化评分，进而刻画该教育网站所收录的作文的质量。

在原数据中，没有对每一篇作文的评分数据，只有网站的老师对作文的评语，缺少一个对文章的量化评价标准，这样就很难统计作文库里作文的质量情况。但是每一条评语可以在一定程度上代表文章的质量，比如从评语“文章可圈可点的佳句不少，给文章增添了些文学情趣，是一篇成功的习作。”大致可以看出这是一片较好的文章，应该具有比较高的“分数”；从评语“用词不准确，语言平淡，表达不准确”中大致可以知道这篇文章质量一般，应该具有较低的“分数”。因此，可以对每一篇作文的评语进行“情感分析”，得到一个量化分数，作为刻画这篇作文质量的得分。

从评语到得分的路径转换图如图 1.2.6 所示。

图 1.2.6 评语到得分的路径转换

①预处理

采用 python 中的 snownlp 对评论的情感分析。传入参数为评语的文本，结果得到对评语的情感得分，范围在 0-1 之间，分数越接近 1 表示情感越偏向正面，越接近 0 表示情感越偏向负面。

情感分析的核心代码如代码 9.

代码 9 - 对作文评语的情感分析

```
1  s=SnowNLP(comment)
2  score=s.sentiments#获得评语的得分
```

②不同角度的统计、可视化与分析

I. 四个等级的作文数量

根据所得的评分，设定不同等级作文的分数区间，通过统计每个等级下的作文数量，进而描绘出不同等级作文占比情况的“玫瑰图”，从而得到对该网站作文质量的直观结果。

本次可视化采用的是 pyecharts 中的 Pie 模块，可以进行基本饼状图和升级版饼状图的绘制。

a. 分级标准

经过对结果分数和对应的评语进行分析，将作文分为“优、良、中、差”四个等级，每个等级对应的分数和数量统计结果如表 1.2.2 所示。

表 1.2.2 作文等级及数量

作文等级	得分区间	作文数量
优	[0.95,1]	58064
良	[0.75,0.95]	25516
中	[0.5,0.75]	41614
差	[0,0.5]	25188

b. 可视化展示

玫瑰图使用 Pie 模块进行绘制，核心参数有 attr 和 value。其中 attr 表示作文等级，value 表示每个等级下的作文数量。

“玫瑰图”可视化代码见代码 10。

代码 10 - 四种等级下作文数量玫瑰图

```
1  from pyecharts import Pie
2  attr = ['优','良','中','差']
3  value = [58064, 25516, 41614, 25188]
4  pie = Pie("作文等级比例图", title_pos='center', width=900)#构建 Pie 对象
5  pie.add("", attr, value, center=[50, 50], is_random=False, radius=[30, 75], rosetype='radius',
6              is_legend_show=False, is_label_show=True)
7  pie.render('article_class_pie.html')#输出结果，为 html 格式，可用浏览器打开
```

可视化结果如图 1.2.7 所示。在玫瑰图中，不仅圆心角可以体现变量占比大小，环形的面积也可以体现这一点，给人更清晰直观的感觉。

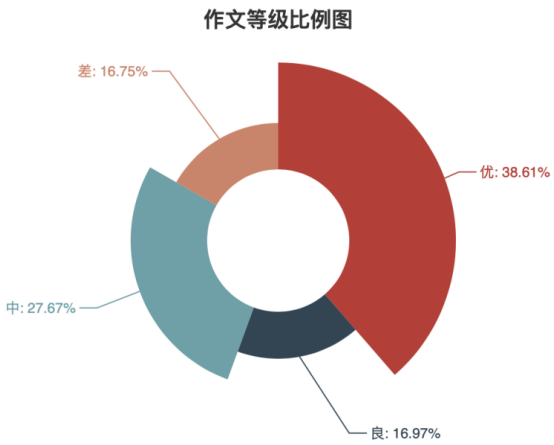

图 1.2.7 四种等级作文占比玫瑰图

c. 结果分析

从图 1.2.7 中我们可以看到，乐乐课堂作文库中的优秀作文占很大比例，很差的作文占比不超过 17%，绝大多数作文还是不错的。

II. 每类话题下“好作文”的比例

本节将通过统计每类话题中优秀作文的占比来绘制不同话题下好作文比例的环形饼状图。这类可视化在很多网站是很常见的。比如在“豆瓣电影”网站会展示出不同类型电影中得分高于某分的电影比例。同样利用 pyecharts 中的 Pie 模块，将在一张图中展示所有话题下优秀作文的比例，直观展示出每一类话题作文的情况。

a. 统计结果

这里统计的“好作文”为等级是“优”和“良”的作文。具体统计结果如表 1.2.3 所示。

表 1.2.3 不同话题“好作文”的数量和作文总数

话题	“好作文”数量/作文总数	话题	“好作文”数量/作文总数	话题	“好作文”数量/作文总数	话题	“好作文”数量/作文总数
事件	3936/14842	人物	6939/21362	公德	1698/5189	动物	1866/4898
命题	417/1125	哲理	1090/2561	地域	378/861	情感	4919/11904
成长	5862/16372	文化	342/828	植物	2117/3934	热点	789/1669
自然	12813/26551	节日	3934/10989	话题	10254/26110	风景	710/1187

b. 可视化

可使用 Pie 模块对多个环形饼图进行绘制，核心参数 attr 和 value。其中 attr 表示一个列表['A','B']，A 为“好作文”，B 为“差作文”，这里为展示方便，将 A 定义为话题名称，B 不做定义；value 也表示一个列表['A','B']，分别为 attr 中两个属性对应的数量。函数会自动算出二者所占比例。可视化代码见代码 11。由于话题数量较多，这里只列出刻画“事件”话题的好作文比例情况。

代码 11 - 不同话题“好作文”所占比例图

```
1  from pyecharts import Pie
2  pie = Pie('各类话题下"好作文"所占的比例', title_pos=400,width=1200,height=450)
3  pie.add("", ["事件", ""], [3936,10906], center=[10, 30], radius=[18, 24],
4              label_pos='center', is_label_show=True, label_text_color=None, )
```

可视化结果见图 1.2.8。

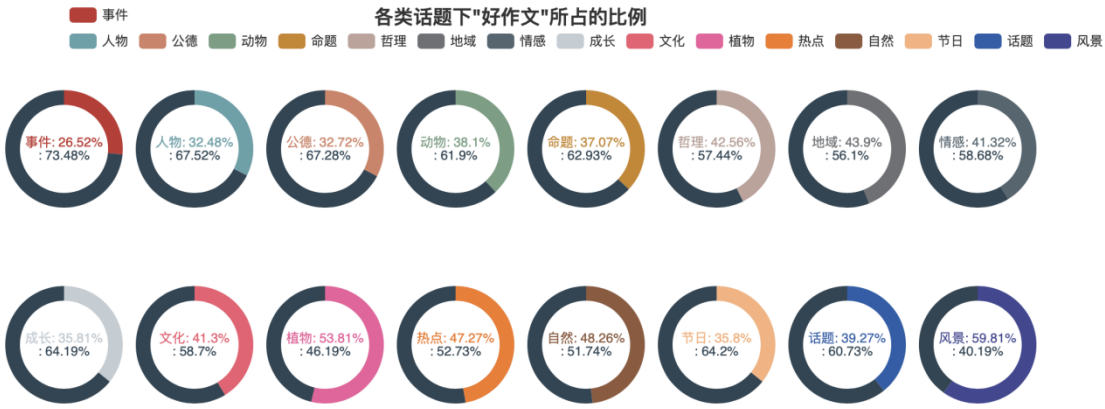

图 1.2.8 不同话题“好作文”所占比例图

c. 结果分析

从图 1.2.8 中可以直观地看出，好作文占比最多的话题是“风景”类话题，占比最少的是“事件”类话题。大多数话题的好作文占比集中在 30%-45%，且超过 50%的话题仅有两个。由此可以推断，学生们擅长于“风景”“植物”类话题作文的撰写，较擅长于“自然”“地域”“哲理”类话题，不太擅长“事件”“人物”类话题。平均每类话题的擅长者约占 2/5。

因此该网站的教育者们在未来可以加大对用户在“事件”类等好作文占比低的话题的训练，以普遍提升学生在薄弱项的作文能力。

III. 每年四个等级的作文数量柱状图

通过统计每年每个等级的作文数量并绘制柱状图，可以直观、准确看出各个等级作文数量随时间的变化。本节主要利用 pyecharts 中的 Bar 模块，以时间为横坐标，每一年一个等级的作文数量为纵坐标，绘制四个等级作文数量的组合柱状图。

a.可视化

可使用 Bar 模块绘制柱状图，核心参数为 attr 和 value。其中 attr 为该年份，value 为该年某一等级下的作文数量。并设置显示出每一等级下作文的最大值、最小值和平均值。可视化代码详见代码 12。

代码 12 - 四个等级的作文数量随年份变化的柱状图

```
1  from pyecharts import Bar
2  attr = ['2013','2014','2015','2016','2017','2018']
3  v1=list(year_date['good'])
4  v2=list(year_date['second'])
5  v3=list(year_date['median'])
6  v4=list(year_date['bad'])
7  bar = Bar("2013-2018 年四个等级的作文数量统计图",width=1200,height=700)
8  bar.add("Good", attr, v1, mark_line=["average"], mark_point=["max", "min"])
9  bar.add("Second", attr, v2, mark_line=["average"], mark_point=["max", "min"])
10 bar.add("Median", attr, v3, mark_line=["average"], mark_point=["max", "min"])
11 bar.add("Bad", attr, v4, mark_line=["average"], mark_point=["max", "min"])
12 bar.render('four_bar_year.html')
```

可视化结果如图 1.2.9 所示。

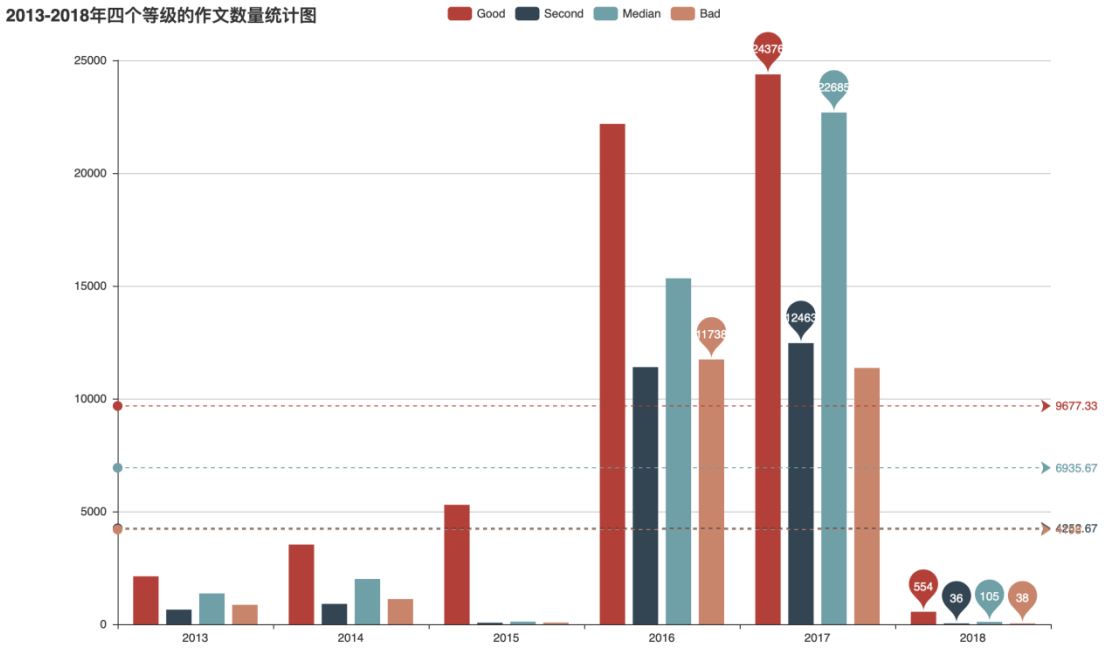

图 1.2.9　2013-2018 年四个等级作文数量变化柱状图

b.结果分析

从图 1.2.9 可以看到，由于 2017 年作文数量整体较高，因此对应的“优”、“良”、“中”等级的作文数量也最多；2016 年作文整体数量仅低于 2017 年，但其“差”作文却占这六年中最大值；2018 年中四种作文等级的数量均占最小值；2015 年“优”作文的占比要远远大于其他三类等级的作文数量。

从以上结果可以分析得到，由于在 2015 年乐乐课堂作文库刚推出不久，2013 年和 2014 年的作文多为网站中别的模块收集的文章，因此数量较少；2015 年该模块推出后，网站为了树立品牌，吸引更多消费者，因此 2015 年的作文主要收录优秀作文；在 2015 年末完成融资后，网站用户大量增加，作文数量也较之前有大幅度的增长；但是，一方面由于用户增加用户水平变得良莠不齐，另一方面由于网站经营者没有刻意收录优秀作文，这导致了 2016 年-2017 年网站中较差的文章占比也大幅增加，优秀作文随一直占据最大比例但是领先的优势显著消退；2018 年文章的数量之所以这么少，一方面可能由于网站使用者在不断减少，另一方面可能由于爬取的 2018 年的数据不充足，或者有可能由于网站更新滞后导致未能及时获得 2018 年的数据。

2 作文内容挖掘—基于 TF-IDF 方法的话题关键词提取 [1]

在文本处理中，提取文本中重要的关键词，是非常有意义的。它可以用来代表文本的重要观点、中心思想，进一步帮助网站的用户快速了解某一话题的作文的整体思路和关键信息，在一定程度上可以帮助用户拓宽写作思路。

本节介绍如何利用 TF-IDF 算法提取一个作文话题下的所有作文集的前 500 个关键词，并将每个话题下最重要的 50 个关键词用词云图的方式展示出来。

2.1 算法介绍

TF-IDF（Term Frequency-Inverse Document Frequency）又称词频-逆文件频率，是一种用于资讯检索与资讯探勘的常用加权技术。TF-IDF 是一种统计方法，用以评估一字词对于一个文件集或一个语料库中的其中一份文件的重要程度。字词的重要性随着它在文件中出现的次数成正比增加，但同时会随着它在语料库中出现的频率成反比下降。

词频 (term frequency, TF) 指的是某一个给定的词语在该文件中出现的次数。这个数字通常会被归一化(一般是词频除以文章总词数), 以防止它偏向长的文件。计算公式为：

$$TF_w = \frac{\text{在某一类中词条}w\text{出现的次数}}{\text{该类中所有的词条数目}}$$

逆向文件频率 (inverse document frequency, IDF) IDF 的主要思想是：如果包含词条 t 的文档越少, IDF 越大，则说明词条具有很好的类别区分能力。某一特定词语的 IDF，可以由总文件数目除以包含该词语之文件的数目，再将得到的商取对数得到。计算公式为：

$$IDF = \log\left(\frac{\text{语料库的文档总数}}{\text{包含词条}w\text{的文档数} + 1}\right)$$

某一特定文件内的高词语频率，以及该词语在整个文件集合中的低文件频率，可以产生出高权重的 TF-IDF。因此，TF-IDF 值倾向于过滤掉常见的词语，保留重要的词语。TF-IDF 计算公式如下：

$$TF - IDF = TD * IDF$$

[1] 由于时间和硬件限制，第 2 和 3 部分未使用 15 万文本共 16 个话题的大数据集进行实验，采用了 16 个话题之一的“事件”类话题的文本进行训练，实际使用了 14784 条文本共 38 个子话题的小规模数据集。

2.2 核心代码描述

代码 13 - TF-IDF 计算

```
#calculate tf
tf = words.count(word) / len(words)#wordes 为该标签下所有词  word 需要统计的词
tf_list.append(tf)
#calculate idf
doc_cn = 0
for doc in corpus:#corpus 是所有文本的词，doc 为一个列表，是一个文本里的词
    if word in doc:#统计这个词出现在了多少个文本里
        doc_cn += 1
idf = math.log(len(corpus) / (doc_cn + 1))
idf_list.append(idf)
#sorted in tfidf
ti_kw_list = sorted(kw_list,key=lambda t:t.tfidf, reverse=True)[:500] #取前 500 关键词
```

2.3 结果分析

（1）结果展示

分别以话题“秋游”“骑马”“考试”为例，每个话题前 50 的关键词结果见表 2.1。

表 2.1 TF-IDF 提取关键词结果

话题	前 50 关键词
秋游	秋游,来到,公园,仙人掌,导游,同学,目的地,美丽,游戏,老师,地方,好像,食物,活动,表演,植物,学校,秋风,游乐场,刺激,开心,东西,烧烤,参观,植物园,碰碰车,沙漠,心情,出发,坐在,好玩,到达,山顶,只见,草坪,草地,动物园,动物,欣赏,香山,离开,妈妈,景色,走进,绿色,时间,枫叶,兴奋,害怕,项目
骑马	骑马,马儿,草原,缰绳,爸爸,马背上,马场,马鞍,骑着马,妈妈,上马,教练,白马,那马,下马,山海关,马屁股,感觉,高大,绳子,马倌,马官,胖弟,马背,照片,来到,温顺,哥哥,白龙马,大马,骑上去,电视,奔跑,叔叔,表姐,牧民,讲解员,大灰马,害怕,握住,强壮,驴车,内蒙古,驯马,奔驰,嘶鸣,雪水,棕色,草地,使劲
考试	考试,试卷,卷子,数学,成绩,复习,语文,老师,检查,题目,分数,考得,紧张,学生,期末考试,考好,妈妈,时间,教室,考完,答案,作文,同学,考场,考试成绩,学习,教育,监考,阅读,道题,陶行知,学校,考卷,原因,知识,心情,父母,粗心,一题,马虎,同桌,仙人掌,数学考试,单元,失败,教师,失分,学期,代表,成功

使用 WordCloud，画出“秋游”话题的关键词词云图，如图 2.1，背景选取了一张枫叶的图片。

图 2.1 “秋游”话题的关键词词云图

（2）结果分析

从表 中可以大致看出，使用 TF-IDF 提取的关键词与话题的相关度还是比较高的。但是其中也不乏有相关度低的词语，比如“考试”话题下的“仙人掌”一词。

话题关键词的提取可以用作对话题的整体描述，让用户对该类话题的写作有初步的印象，同时也能为他们写该类话题作文提供借鉴。

3 基于机器学习方法下作文库的应用

3.1 应用 1—基于 Doc2Vec 方法的话题关键词统计

Doc2Vec 作为机器学习领域中常见的文本处理算法，可以根据有标签的数据训练出文本中的词向量。进而可以根据已有的向量通过“距离”的计算，选出离“话题”最“近”的词语作为关键词。在一定程度上改进了仅依据词频的 TF-IDF 算法。本节在基本距离的基础上做了改进，通过对比五种标准下得到的话题关键词，得出关键词统计中较优的排序标准。

3.1.1 Doc2Vec 的原理

Doc2Vec 是一种非监督式算法，可以获得 sentences/paragraphs/documents 的向量表达，是 Word2Vec 的拓展。学出来的向量可以通过计算距离来找 sentences / paragraphs /

documents 之间的相似性，可以用于文本聚类，对于有标签的数据，还可以用监督学习的方法进行文本分类。其主要包含两种模型：PVDM 和 DBOW。

（1）PVDM

训练句向量的方法和词向量的方法非常类似。训练词向量的核心思想就是说可以根据每个单词的上下文预测，也就是说上下文的单词对是有影响的。那么同理，可以用同样的方法训练 Doc2Vec。例如对于一个句子 I want to drink water，如果要去预测句子中的单词 want，那么不仅可以根据其他单词生成 feature，也可以根据其他单词和句子来生成 feature 进行预测。因此 Doc2Vec 的框架如图 3.1.1 所示：

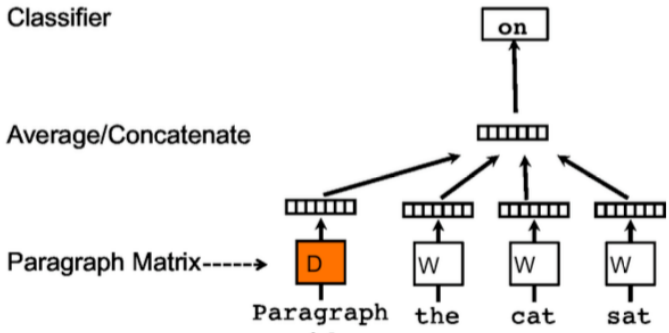

图 3.1.1　PVDM 模型

（2）DBOW

这种种训练方法是忽略输入的上下文，让模型去预测段落中的随机一个单词。就是在每次迭代的时候，从文本中采样得到一个窗口，再从这个窗口中随机采样一个单词作为预测任务，让模型去预测，输入就是段落向量。如图 3.1.2 所示。

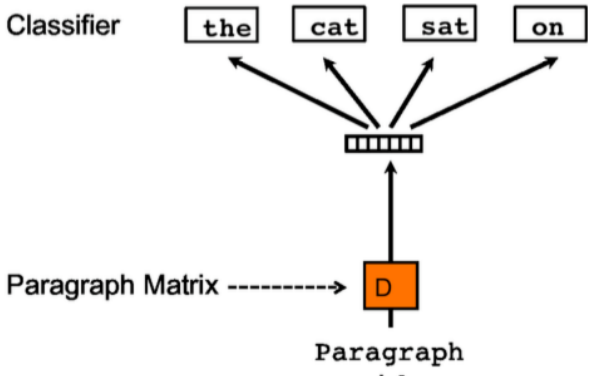

图 3.1.2　DBOW 模型

3.1.2 核心代码说明

为了提取每个话题的关键词，将把一个话题下的所有文本看成一个文本（document），话题作为训练的标签。可以使用第三方库 gensim 进行 Doc2Vec 模型的训练。由于一个话题下包含多个原始的作文，因此选择忽略输入的上下文，采用 DBOW 的训练模型。训练模型的代码见代码 14。

代码 14 - 训练模型(DBOW)

```
import gensim
from gensim.models import Doc2Vec
# 加载数据
documents = [ ]#可训练 sentence,pragraph,document
for tag, content in tc_map.items():#tag 是标签，content 是标签下的所有文本
    # 切词，返回的结果是列表类型
    text = ''.join(tc_map[tag])#将一个标签下所有文本连成一个字符串
    words = list(jieba.cut(text))#切词
    # 这里 documents 里的每个元素是二元组（一个标签下所有词的列表，一个标签的列表）
    documents.append(gensim.models.doc2vec.TaggedDocument(words, [tag]))
    logging.info('%s has loaded...'%tag)#引入日志配置
# 模型训练，doc2vec 有两种模型, dm=0 dbow, dm=1 dmpv
#初始化 Doc2Vec 对象，设置神经网络的隐藏层的单元数为 200，生成的词向量的维度也与神经网络的隐藏层的单元数相同。设置处理的窗口大小为 8 个单词，出现少于 2 次数的单词会被丢弃掉，同时并发线程数 4：
model = Doc2Vec(documents, dm=0, size=200, window=8, workers=4, min_count=2)
# corpus_count 是文件个数  epochs 训练次数
model.train(documents, total_examples=model.corpus_count, epochs=50)
# 保存模型
model.save('doc2vec_dbow.model')
```

3.1.3 统计标准说明

模型训练的结果中有标签向量和每个词的向量。可以通过计算每个标签话题和该话题下每个词的“距离”，选取“距离”最近的词语作为关键词。

标准距离的计算包含“欧几里德距离”和“cosine 相似度”。还可以采用“欧几里德距离与 TF-IDF 值混合”和“cosine 相似度与 TF-IDF 值混合”的方式进行关键词的提取。

关键词统计标准及计算公式如图 3.1.3 所示。

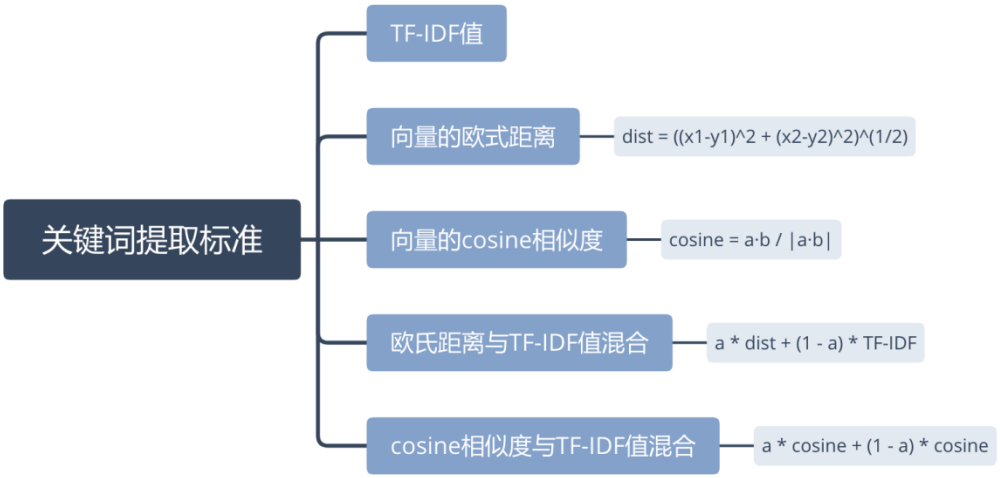

图 3.1.3 几种关键词统计标准及计算公式

3.1.4 结果分析

在参数 a=0.6 的情况下，五种统计方式得到话题“秋游”前十关键词的结果见表 3.1。

表 3.1 五种方法下前十个关键词

TF-IDF	DIST	COSINE	DIST-TFIDF	COSINE-TFIDF
秋游	询问	询问	秋游	询问
来到	烧烤	烧烤	花馍	烧烤
公园	仪器	仪器	吸管	仪器
仙人掌	跳来跳去	跳来跳去	镶着	跳来跳去
导游	入水	入水	颤抖	入水
同学	喝彩	灵动	新生	灵动
目的地	躲过	躲过	鸟语花香	躲过
美丽	灵动	激情	更美	激情
游戏	骑车	骑车	枫	骑车
老师	登山	拍手	棕红色	拍手

不同的方法会产生不同的关键词统计结果。在实际应用中，网站管理者可以结合用户需求和现实作文教育中的应用场景，通过调节参数和训练更大规模的数据，选取最佳的关键词统计方法，为用户提供更有效的话题关键词。

3.2 应用 2—推荐高级表达词

本节通过 Doc2Vec 训练出来的向量，为用户输入的词语推荐作文中常用的“高级表达”，帮助用户拓展写作思路、提升作文的表达技巧。

3.2.1 算法思想

用户输入一个词语（在作文词语库中存在），首先判断该词所属的话题，接着在该话题下的所有词语中寻找距离用户输入词最近的“高级”词语。这里为简单起见，定义“高级表达”为字数大于 3 的词语。算法思想见图 3.2.1.

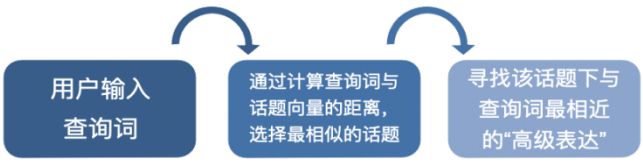

图 3.2.1 推荐高级表达词算法思路

3.2.2 结果展示

推荐词结果如表 3.2 所示。可以看到，在推荐的词语中，多数词语是与输入词语义相关的成语或短语，但也包含不少的“噪音”。分析原因，一方面由于训练数据集不足，仅有一万多条文本；另一方面由于算法具有局限性，采用距离计算将用户输入词归类不能完全体现词语和话题之间的相关度，算法本身需要改进。

表 3.2 推荐词结果

用户输入词	相关话题	推荐词列表
蝴蝶	春游	'新鲜空气', '饱餐一顿', '恭喜发财', '宰相肚里', '情不自禁', '欺人太甚', '天安门广场', '激动不已', '依依不舍', '百花争艳', '军事训练', '迅雷不及掩耳', '辉煌成就', '东躲西藏', '所向无敌', '隐隐约约', '惹人注目', '琳琅满目', '七十多年', '翻来覆去', '春光明媚', '气喘嘘嘘', '井然有序', '感到高兴', '上气不接下气', '欢迎仪式', '捧腹大笑', '一筹莫展', '近在咫尺', '置身其中', '红杏出墙', '耐人寻味', '美味可口', '少数民族', '高耸入云', '大好河山', '有用之才', '日本鬼子', '热带雨林', '不相上下', '娇生惯养', '望梅止渴', '五颜六色', '发号施令', '誓不罢休', '千辛万苦', '百花盛开', '楚楚动人', '古香古色', '添砖加瓦'

考试	考试	'期末考试', '何时何地', '期中考试', '失魂落魄', '大声喊叫', '铁石心肠', '七上八下', '扑通一声', '必不可少', '睁大眼睛', '小菜一碟', '犹豫不决', '纸上谈兵', '朗读课文', '百科全书', '伟大工程', '闷闷不乐', '大闹天宫', '迥然不同', '天壤之别', '咬牙切齿', '职业道德', '临渴掘井', '皱着眉头', '漫不经心', '有史以来', '远远不够', '耳濡目染', '数学公式', '公共场合', '不能平静', '老大徒伤', '天下大同', '圆满结束', '面对现实', '内心深处', '独自一人', '想想也是', '氢氧化钠', '百战不殆', '解决问题', '凉了半截', '计算错误', 'nxxx', '如梦初醒', '比比皆是', '来之不易', '不敢相信', '如此说来', '一鸣惊人'

4 建议

根据对乐乐课堂作文库的数据分析，对在线教育的发展提出以下几条建议：

（1）经济发展影响用户在线学习

经济因素对用户学习产生了巨大的影响，经济发达的地区不仅对学校教育重视，对在线教育的关注度也是比较高的，而在西南、西北等一些经济发展较慢的地方，人们的在线学习意识较为薄弱，相关教育部门及在线教育平台应该加大在西北、西南地区的推广、宣传，使更多的人通过在线教育平台进行学习，获得更多的学习机会和更丰富的学习资源。

（2）社会需求影响用户学习

作文库的建设应该和线下教学的方向结合起来，尤其和实时的考试方向结合起来，为用户提供多元的资源。还可以将作文库统计的一些结果展现在用户面前，譬如最近一个月提交最多的话题类型、不同难度等级的写作话题，让用户能够及时了解最新的资源和方向，合理安排自己的学习计划，提高自身能力。

（3）在线教育平台应该有针对性地开展课程

不同年龄段、学历层次的用户对学习的需求是不一样的，对付费学习的意愿、金额也是不同的。对于中小学生而言，应充分考虑其经济实力，降低付费金额，多增加免费的优质课程。对于年龄较小、学历层次较低的用户而言，他们的学习自主性较弱，易受各种因素的影响，在线教育平台应通过不同的方式，督促、提醒其进行学习，增长其在线学习时间。再者，根据一段时间的数据可以分析出如今用户整体的薄弱项，可以制定相应的训练计划，帮助用户在薄弱的写作环节有所提升。

5 总结

在大数据的驱动下，在线教育呈现出蓬勃发展的趋势，同时由数据带来的一系列问题也需要通过不断地探究进行解决。新技术，新挑战，大数据带给我们的除了机遇，还有挑战，要善于运用技术，解决学习中产生的问题。利用大数据对教育数据进行全方位与全程性采集，对多维教育大数据进行深度分析；运用可视化的分析工具，将在线教育过程中产生的问题用可视化数据呈现出来，使在线教育平台优化自己的课程；运用一些前沿的方式，譬如机器学习、深度学习的模型与理论，对数据进行更加深入的探索，研发一些新的功能，针对性地推广、开展课程，让学习者更好地参与学习，从知学转变为好学，由好学进一步变为优学。

数据时代，要学会正确地获取数据、分析数据、应用数据，利用数据服务学习生活，不断学习新技术，解决数据产生的问题，培养数据素养，树立数据意识。

6 参考文献

[1] 杨明刚,孙启超,朱韦茜.可视化大数据在在线教育教学中的应用研究[J].设计,2015(04):157-158.

[2]冯珊珊.乐乐课堂:回归教育本质[J].首席财务官,2016(Z1):106-108.

[3]魏梦楠,马燕.基于大数据分析的在线教育用户基本属性的研究——以“腾讯课堂”为例[J].中国信息技术教育,2018(19):103-105.

[4]胡振兴.基于“互联网+”的 K12 教育发展趋势探究[J].现代经济信息,2018(01):406-407.

[5] 蔡樱.大数据技术下个性化在线教育互动式教学探索[J].高等建筑教育,2018,27(04):131-134.

[6]颜秉刚.论大数据时代背景下的在线教育[J].西部素质教育,2018,4(13):124.

[7]何璐.面向在线教育领域的大数据研究及应用[J].信息与电脑(理论版),2017(22):139-140.