网络数据挖掘课程报告

姓名：何长鸿

学号：2016141482154

时间：2018年6月21日

小组题目：scuinfo帖子情感分析

网络与数据挖掘课程报告

何长鸿 2016141482154

1. 课程项目

Scuinfo文本分类分析

1. 项目类容

爬取川大匿名社区SCUinfo在一段时间内的帖子，对其进行情感分类分析，包括情绪分类（积极，消极），帖子内容关联分析等。

1. 个人工作完成报告
   1. **工作概述**

负责数据收集、预处理以及简单的情感分析

* 1. **爬虫方案：**

scuinfo为动态加载网页，并且有移动端验证，尝试使用scrapy爬取数据失败，需要对爬虫进行大量定制优化，为了提高开发效率，最终使用了之前爬取QQ空间采用的方式：使用selenium库调用firefox浏览器驱动，该再通过代码模拟人为操作浏览器。获取到对应页面数据后，使用etree和xpath选取相应的目标节点数据。

**优点：**轻松解决网页动态加载，登录验证，移动端验证等问题。

**缺点：**该方式需要一直保持浏览器前台运行，并且只能为单线程模式，爬取效率相对较低。

**关键代码：**

|  |
| --- |
| print("准备打开浏览器")  driver = webdriver.Firefox()  print("打开浏览器")  driver.maximize\_window()  driver.get("https://scuinfo.com/")  time.sleep(1)  driver.find\_element\_by\_css\_selector('body').click()  #浏览器翻页控制  for i in range(1, 20000):  driver.find\_element\_by\_xpath("//body").send\_keys(Keys.DOWN)  time.sleep(0.001)  selector = etree.HTML(driver.page\_source)  divs = selector.xpath('//div[@id="homeList"]/div[@style="padding: 12px 25px 4px; border-bottom: 12px solid rgb(242, 242, 242);"]')  #这里使用 a 表示内容可以连续不清空写入  with open('scuinfo.txt','a',encoding='utf-8') as f:  for div in divs:  content = div.xpath('./div[2]/div[1]/span/text()')  comment\_num = div.xpath('./div[2]/div[2]/div[2]/div[1]/div[2]/text()')  name = div.xpath('./div[1]/div[1]/div[2]/text()')  good\_num = div.xpath('./div[2]/div[2]/div[2]/div[2]/div[2]/text()') |

* 1. **数据存储**

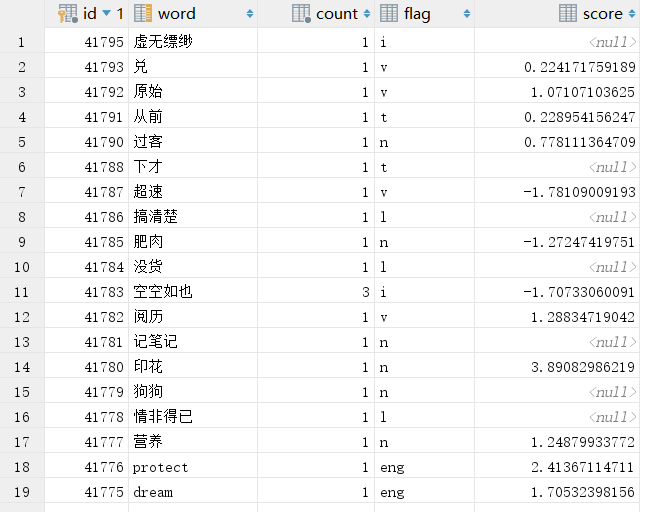
为了方便初期直观的处理数据，使用Mysql数据库保存爬虫数据集。最终数据5000余条。存储方式使用自己编写的Mysql操作类实现。每次爬虫进程创建一个数据库连接对象。

* 1. **文本分词**

使用第三方python库jieba分词，以及词汇情感标注开源语料库，将分词结果保存到数据库，每个数据项包括词汇内容，频数，词性，以及情感评分，其中，积极情感分数为正，消极情感分数为负数。分数绝对值代表情感激烈程度。

* 1. **数据展示**

分词结果合计41795条，帖子总数合计5429条。



* 1. **初步情感分析**

使用Corpus语料库，该语料库包含一个词汇词汇情感评分的标签库，该标签库数据为微博帖子训练而来，由于微博与scuinfo平台相似度较高，因此理论上数据用在本场景中会有比较不错的效果。

本次情感评分方式为直接将每个帖子分词后，将所有分词的情感评分相加。此处分别展示积极和消极情绪的前3条。

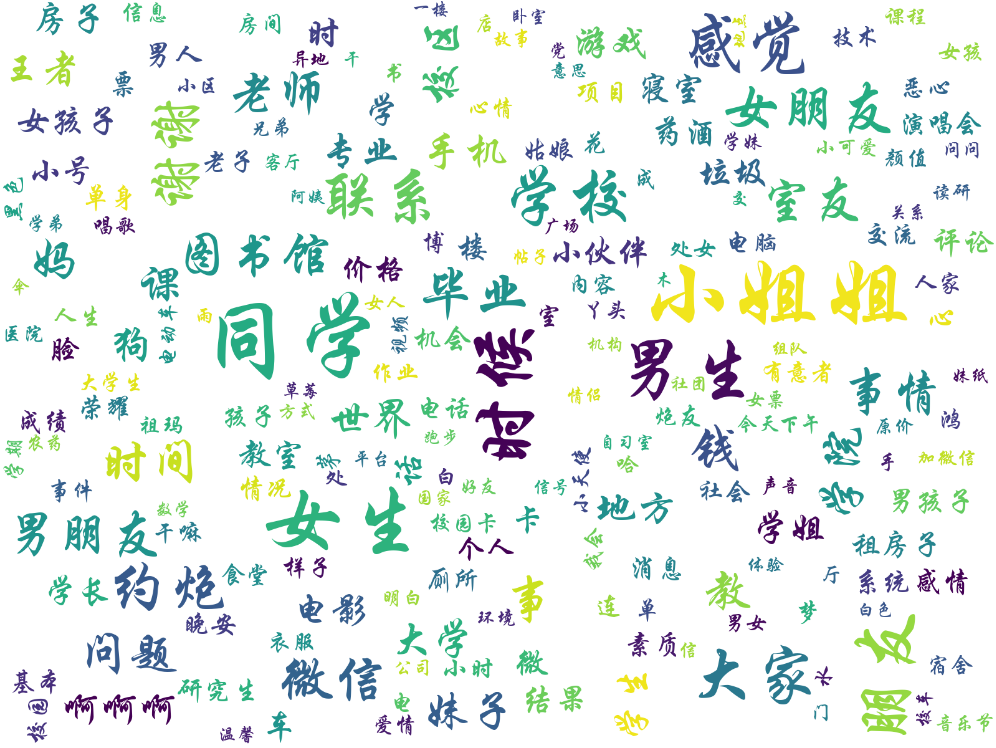
1. 真心讨厌双标\* 有室友晚上写论文第二天一个劲的说别人吵到自己，别人睡觉又一个劲的说话 有病了解一下？真的今天说话一直被打断只想说敲里嘛敲里马，真的很没素质ok，老是说打断你说话的人有问题，你自己有点 B，真的火大了 -33.0377546758296
2. 十二舍有个女人，吹头发三角插头，于是我就走过去插了那个两脚插头，于是她吹完后狠狠一拔她的插头，把我的吹风机一下子拔出来然后没有一句话拔腿就跑，我他妈转过头去想问她为啥这样结果人已经跑远，请问你为什么要这么恶意？我妨碍到你了？两个人可以同时用并且不妨碍你你为什么非要我一个人霸占？果然绝大多数长得好看...... -30.85746272228401
3. t在校大学生的苦逼，四川大学校车站的庸车，尾号为3512，3102，1339的车子，害得我出了丢钱包事故，把我搞的半死不活。 在打电话之前虽然很难过但是知道明明钱包就在三辆车的其中之一上。而现在是打电话后司机找都不找就说钱包肯定不在车上，这样的庸车留着有何用？ -29.93353905734

1. 华为财务精英挑战赛，粤港澳赛区区域赛招募新队员1至3人，截止日期为华为那边案例下发日期之前。我们是Nikolas队伍，原团队成员为三个中山，两个川大，一个暨南，由于一些老队员由于期末考、实习等原因无法参赛，先重新招募新队员和我们一起参加粤港澳赛区的比赛。我们的队伍通过了暨南赛区线上筛选14进4，之...... 54.50014566772352
2. 2018年腾讯游戏安全技术竞赛，由腾讯游戏安全主办的2018年游戏安全技术竞赛，作为国内首个专注于“游戏安全”领域的技术赛事，旨在提高高校学生以及游戏从业者对游戏安全领域的认知，并通过竞赛的形式，挖掘技术人才，共同完善与捍卫游戏安全的防线。本次竞赛不仅提供了丰厚的现金奖励，参赛者还有机会与专...... 54.50014566772352
3. 同道共识，方谓同仁；志同道合，共赴理想。地产top10的绿城中国暑期实习生校园招聘“同道人2.0计划”开启啦！现面向川大，招募2名校园大使，我们希望你：·执行力超强，能按照招聘节点计划，完成相应任务；·亲和力、沟通力与团队协作能力nice，能协同其他大使小伙伴，愉快高效地完成任务；·19届...... 54.1780398103219

由以上结果可知，前三条为消极情绪的帖子，识别结果是比较准确的，而对于积极情绪的帖子，前面多数为招聘广告，从内容来看，广告内容的确多为积极词汇，这也是影响分析结果的主要原因。

* 1. **词云生成**

**为了直观感受川大学生在匿名平台讨论的主题，使用python库按词频生成词云，结果如下图。**



1. 心得体会

本学期网络数据挖掘课程上了解了很多机器学习算法，虽然没有很透彻的体会到它们的精妙，但激发了我学习机器学习的兴趣。期末项目选题也不是很理想，但在一定程度上提高了我的代码能力。希望以后能够在数据挖掘和分析的道路上走得更远。