武汉大学国家网络安全学院

实验报告

课程名称 社会计算的基本方法与应用

专业年级 2021 级网安

姓名(学号)

实验学期 2022 学年 1 学期

课堂时数 33 课外时数 0

填写时间 2022 年 11 月 9 日

实验概述
【实验项目名称】： 词语级情感倾向分析

【实验目的】：**1.** 计算数据集中最有倾向的前五十个词

2.学习使用 **python**

3.学习社会计算的基本方法

【实验环境】（使用的软件）：**visual studio code 2022**

(1) 硬件环境：**CPU AMD RYZEN 5900HX GPU RTX 3070**

(2) 操作系统环境：**Windows 11**

(3) 测试脚本编程语言：**Python**

(4) 被测系统编程语言：

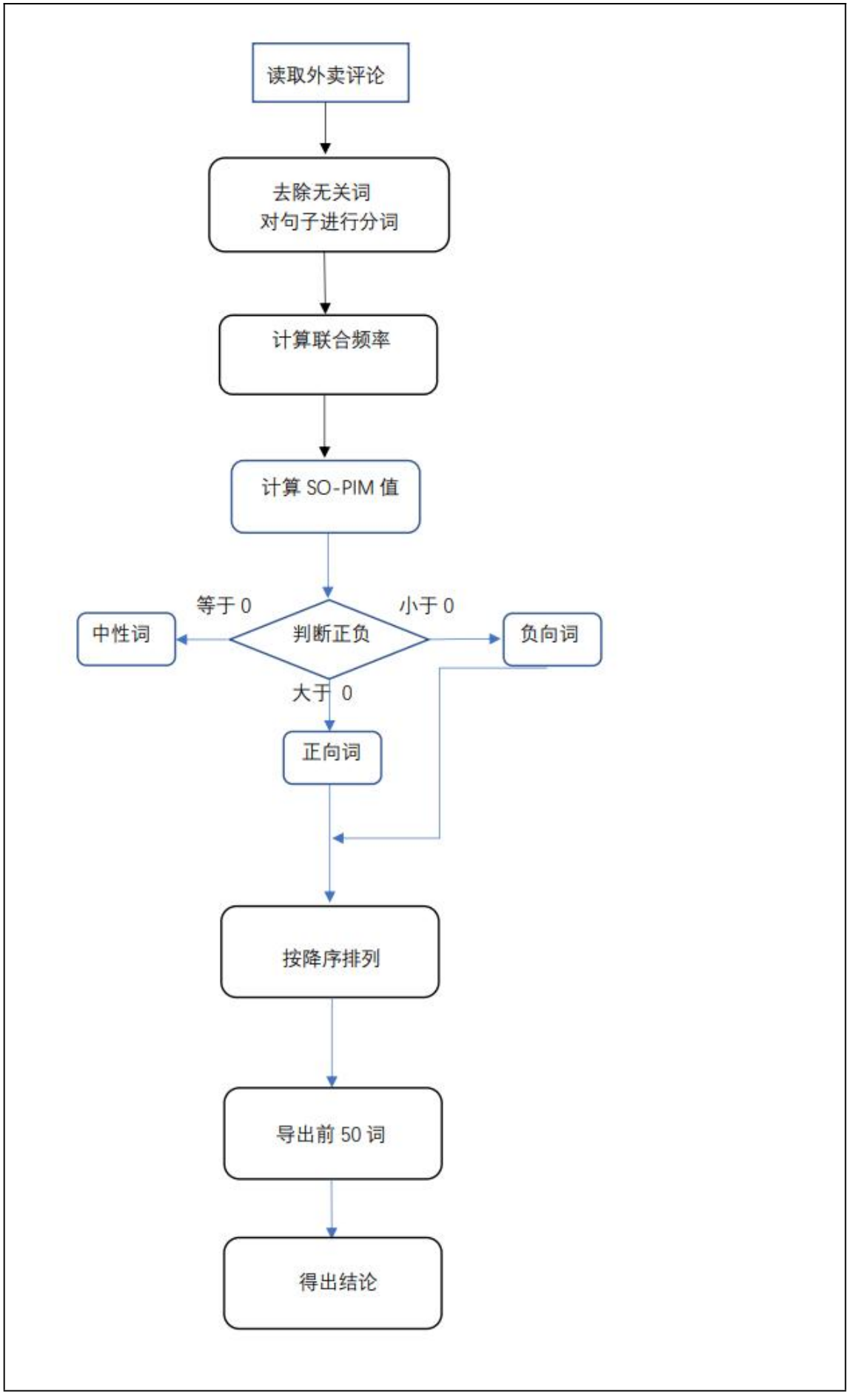
(5) 网络环境：无

(6) 其他环境：无

【参考资料】：**github nlp_so-pmi,Balding-Lee**

实验内容

【实验方案设计】：


读取外卖评论
去除无关词
对句子进行分词
计算联合频率
计算 SO-PIM 值
等于 0
小于 0
中性词
判断正负
负向词
大于 0
正向词
按降序排列
导出前 50 词
得出结论

1.读取文件：读取文件时按行来读取文件中的内容，并通过切片的方法来去除每一行前面的数字，并单独去除第一行的内容来完成第一步的读取。

2. 分词：分词前先进行去除无关词的操作，比如数字、字符等，事先准备一个文件来保存要去除的内容，其中包括了一些无关的高频词汇。其后再使用 **jieba** 分词的精确模式来分割为一个个词汇并保存在列表中。

3. 计算 **so_pmi**：在基础公式的基础上，因为分子分母可能为 **0**，所以通过**+1** 平滑的方式来避免 **0** 的错误；并且采用了改进后的公式，能够以连加的方式来计算，简化了代码量。

4. 分类：依据计算出的词汇的 **pmi** 值的正负来分出正向词汇和负向词汇，然后按照绝对值降序的方式来排出前五十名。

5. 写入文件

```
#去除无关词
def removeword(words):
    removelist = []
    newlist = []
    with open('./removewords.txt',mode='r',encoding='utf-8') as f:
        removewords = f.read()
        removelist = removewords.split('\n')
    for word in words:
        if not(word in removelist):
            newlist.append(word)
    return newlist

#jieba分词
def cutsentence(sentences):
    words = []
    for sentence in sentences:
        cutword = list(jieba.cut(sentence,cut_all=False))
        words.append(cutword)
    return words
```

```python
#计算联合频率
def count_intersection(word1,word2,sentences):
    count = 0
    for sentence in sentences:
        if (word1 in sentence) and (word2 in sentence):
            count += 1
    return count
#计算word的SO-PMI值
def SO_PMI(word,P_seed,N_seed,sentences):
    PMI = 0.
    num_seed = len(P_seed)
    for i in range(num_seed):
        P_COUNT = count_intersection(word,P_seed[i],sentences)
        N_COUNT = count_intersection(word,N_seed[i],sentences)
        PMI += math.log((P_COUNT+1)/(N_COUNT+1))#防止为0都+1
    return PMI


P_seed = ['很好','不错','很赞','辛苦','很快','喜欢','好评','谢谢','满意','好喝']
N_seed = ['垃圾','太慢','凉','不行','很差','太少','差评','太晚','不好','服了']
```

【小结】：

以下分别是基准词从 10 个到 6 个递减的结果

No_1,赞
No_2,好
No_3,感谢
No_4,准时
No_5,棒棒
No_6,足
No_7,超级
No_8,也好
No_9,好吃
No_10,及时
No_11,实惠
No_12,快
No_13,热情
No_14,超快
No_15,他家
No_16,值得
No_17,给力
No_18,特别
No_19,有点
No_20,非常感谢
No_21,大
No_22,更好
No_23,美味
No_24,最
No_25,很棒
No_26,五星
No_27,饮料
No_28,不过
No_29,餐员
No_30,分量
No_31,但
No_32,支持
No_33,一如既往

No_1,难吃
No_2,慢
No_3,饿
No_4,差劲
No_5,不要
No_6,再也不会
No_7,差
No_8,尼玛
No_9,连
No_10,是不是
No_11,下午
No_12,单
No_13,送错
No_14,公里
No_15,无
No_16,影响
No_17,马上
No_18,歉意
No_19,任何
No_20,退款
No_21,超慢
No_22,不仅
No_23,他们
No_24,钱
No_25,不说
No_26,路程
No_27,不到
No_28,接到
No_29,两
No_30,我家
No_31,没到
No_32,足足
No_33,饺子

No_1,难吃	No_1,赞
No_2,慢	No_2,好
No_3,饿	No_3,感谢
No_4,差劲	No_4,足
No_5,不要	No_5,好喝
No_6,再也不会	No_6,准时
No_7,差	No_7,也好
No_8,尼玛	No_8,实惠
No_9,连	No_9,棒棒
No_10,是不是	No_10,及时
No_11,下午	No_11,超快
No_12,单	No_12,热情
No_13,送错	No_13,他家
No_14,公里	No_14,美味
No_15,无	No_15,值得
No_16,影响	No_16,好吃
No_17,马上	No_17,五星
No_18,歉意	No_18,超级
No_19,任何	No_19,给力
No_20,退款	No_20,支持
No_21,超慢	No_21,更好
No_22,不仅	No_22,非常感谢
No_23,他们	No_23,有点
No_24,钱	No_24,大
No_25,不说	No_25,方便
No_26,路程	No_26,很棒
No_27,不到	No_27,挺好吃
No_28,接到	No_28,真不错
No_29,两	No_29,挺快
No_30,我家	No_30,服务
No_31,没到	No_31,快
No_32,足足	No_32,不过
No_33,饺子	No_33,最

No_1,赞	No_1,难吃
No_2,好	No_2,饿
No_3,感谢	No_3,连
No_4,好喝	No_4,单
No_5,及时	No_5,不说
No_6,准时	No_6,慢
No_7,实惠	No_7,差
No_8,超快	No_8,差劲
No_9,也好	No_9,了
No_10,超级	No_10,不要
No_11,非常感谢	No_11,再也不会
No_12,棒棒	No_12,叫
No_13,热情	No_13,下午
No_14,值得	No_14,影响
No_15,天气	No_15,足足
No_16,五星	No_16,无
No_17,足	No_17,尼玛
No_18,好吃	No_18,回复
No_19,方便	No_19,汤
No_20,大	No_20,不
No_21,更好	No_21,任何
No_22,很棒	No_22,改进
No_23,不过	No_23,退款
No_24,他家	No_24,没有
No_25,服务	No_25,完全
No_26,美味	No_26,超慢
No_27,给力	No_27,钱
No_28,有点	No_28,路程
No_29,特别	No_29,送错
No_30,支持	No_30,是
No_31,比较	No_31,不到
No_32,但	No_32,来
No_33,餐员	No_33,两
No_34,这种	No_34,退到

No_1,难吃	No_1,赞
No_2,差	No_2,谢谢
No_3,了	No_3,好喝
No_4,不要	No_4,感谢
No_5,叫	No_5,及时
No_6,连	No_6,好
No_7,单	No_7,值得
No_8,不说	No_8,五星
No_9,慢	No_9,实惠
No_10,差劲	No_10,也好
No_11,饿	No_11,非常感谢
No_12,不	No_12,棒棒
No_13,再也不会	No_13,热情
No_14,下午	No_14,超快
No_15,完全	No_15,足
No_16,钱	No_16,准时
No_17,足足	No_17,更好
No_18,没有	No_18,很棒
No_19,把	No_19,方便
No_20,是	No_20,美味
No_21,无	No_21,超级
No_22,尼玛	No_22,他家
No_23,回复	No_23,点赞
No_24,汤	No_24,但
No_25,任何	No_25,支持
No_26,接单	No_26,不过
No_27,我	No_27,比较
No_28,最后	No_28,量足
No_29,来	No_29,好吃
No_30,退款	No_30,给力
No_31,是不是	No_31,服务
No_32,就	No_32,礼貌
No_33,简直	No_33,重要

No_1,赞	No_1,难吃
No_2,好喝	No_2,差劲
No_3,谢谢	No_3,了
No_4,感谢	No_4,慢
No_5,好	No_5,差
No_6,好评	No_6,饿
No_7,方便	No_7,差评
No_8,值得	No_8,足足
No_9,他家	No_9,连
No_10,及时	No_10,不
No_11,觉得	No_11,不说
No_12,超级	No_12,叫
No_13,也好	No_13,地
No_14,比较	No_14,上班
No_15,更好	No_15,下楼
No_16,非常感谢	No_16,还有
No_17,棒棒	No_17,是不是
No_18,很棒	No_18,是
No_19,实惠	No_19,不要
No_20,点赞	No_20,还
No_21,热情	No_21,超慢
No_22,大	No_22,尼玛
No_23,必须	No_23,80
No_24,不是	No_24,不仅
No_25,忘	No_25,送错
No_26,支持	No_26,失望
No_27,牛肉	No_27,把
No_28,足	No_28,最后
No_29,不过	No_29,来
No_30,但	No_30,接单
No_31,礼貌	No_31,两
No_32,超快	No_32,单
No_33,尖椒	No_33,饺子
No_34,精致	No_34,我

可以看到随着基准词的减少在结果中逐渐出现了一些奇怪的无关词汇，说明此算法的精确性会随着基准词的增多而上升。

在实验过程中也大量学习了 python 相关知识，基本具备了使用 python 语言的能力，对于社会计算相关内容的理解更进一步。

遇到的问题：1.python 环境的特殊性，python 与之前所学过的语言有很大不同，它需要更复杂的环境配置来运行，最终通过查阅资料学习来解决。

2. so_pmi 算法的相关问题：算法还存在改进空间，通过查阅资料发现了改进的算法。

心得：1.通过实验基本掌握了 python 语言及其思维方式，面向对象语言编程的便利性。 2.社会计算方法简单的实现就已经需要借助大量工具，且结果不尽如人意，更加复杂的算法是如何实现的更加吸引着我们去了解社会计算相关内容。
指导教师评语及成绩
【评语】：   **成绩：　　指导教师签名：** **批阅日期：**

附件：

实验报告说明

1．实验项目名称：要用最简练的语言反映实验的内容。要求与实验指导书中相一致。

2．实验目的：目的要明确，要抓住重点，符合实验任务书中的要求。

3．实验环境：实验用的软硬件环境（配置）。

4．实验方案设计（思路、步骤和方法等）：这是实验报告极其重要的内容。包括概要设计、详细设计和核心算法说明及分析，系统开发工具等。应同时提交程序或设计电子版。

对于**设计型和综合型实验**，在上述内容基础上还应该画出流程图、设计思路和设计方法，再配以相应的文字说明。

对于**创新型实验**，还应注明其创新点、特色。

5．结论（结果）：即根据实验过程中所见到的现象和测得的数据，做出结论（可以将部分测试结果进行截屏）。

6．小结：对本次实验的心得体会，所遇到的问题及解决方法，其他思考和建议。

7．指导教师评语及成绩：指导教师依据学生的实际报告内容，用简练语言给出本次实验报告的评价和价值。