浙江水利水电学院

实验指导书

（ 2018-2019 学年 2 学期）

|  |
| --- |
| 实验名称： 使用Python分析文本情感 |
| 教学单位： 信息工程与艺术设计学院 |
| 专 业： 软件工程 |
| 班 级： S18-1,2 |
| 指导教师： 张鹏亮 |

2019年 3 月 15 日

# 目的和意义

1. 掌握Python的函数和组合数据类型的使用。

2. 训练基本程序逻辑能力。

3. 了解Python应用场景之一，拓宽视野。

4. 简单了解文本情感分析。

# 基本要求

## 技术要求

操作系统：Linux/mac OS/Windows

Python版本：Python3.6级以上，建议使用Anaconda虚拟环境

其他工具：中文褒贬义词典（清华李军）

## 成果要求

（1）Python代码

（2）算法分析设计（算法描述）

（3）实验报告

# 实验内容

维基百科的情感分析的定义是：文本情感分析（也称为意见挖掘）是指用自然语言处理、文本挖掘以及计算机语言学等方法来识别和提取原素材中的主观信息。

现在情感分析被广泛的应用于挖掘互联网上的主观信息，文本数据来源包括微博、博客、微信、社交平台、文章、评述评论等等。分析技术也比较多样化包括自然语言处理（NLP）、数理统计、机器学习等等。企业可以使用挖掘出来的情感信息寻找新的营销机会，改善营销方式，针对人群采取精准营销等等。

本次实验只做简单的统计，暂不牵涉机器学习等方面的内容。

**1. 文本情感分析简介[1]**

现有的文本情感分析的途径大致可以集合成四类：关键词识别、词汇关联、统计方法和概念级技术。关键词识别是利用文本中出现的清楚定义的影响词（affect words），例如“开心”、“难过”、“伤心”、“害怕”、“无聊”等等，来影响分类。词汇关联除了侦查影响词以外，还附于词汇一个和某项情绪的“关联”值。统计方法通过调控机器学习中的元素，比如潜在语意分析（latent semantic analysis），SVM（support vector machines），词袋（bag of words），等等。一些更智能的方法意在探测出情感持有者（保持情绪状态的那个人）和情感目标（让情感持有者产生情绪的实体）。要想挖掘在某语境下的意见，或是获取被给予意见的某项功能，需要使用到语法之间的关系。语法之间互相的关联性经常需要通过深度解析文本来获取。与单纯的语义技术不同的是，概念级的算法思路权衡了知识表达（knowledge representation）的元素，比如知识本体 （ontologies）、语意网络（semantic networks），因此这种算法也可以探查到文字间比较微妙的情绪表达。例如，分析一些没有明确表达相关信息的概念，但是通过他们对于明确概念的不明显联系来获取所求信息。

**2. 实验任务**

本次实验做简单的中文文本情感分析，即分析一段文本表达的正面（褒义）还是负面（贬义）情绪。

本次任务分为两阶段。

**第一阶段**为预处理阶段，**是必做任务**。主要训练Python基本语法、函数、组合数据类型的使用。

第一阶段步骤：

1. 分句。分析每一句的情感度，最后再综合整段文本情感倾向。中文分句以中文的特定的标点符号间隔，如（。！？等），请统计好间隔符号后，编写文本分句函数，该函数返回每一句的列表。中文标点符号的处理，可以使用第三包如zhon，或自己编写相关代码。

2. 中文分词。可以使用jieba库来对每语句进行分词。可参考教材p169页，jieba的相关介绍。请编写函数，实现分词功能。注意，如有自定义词，需要添加。

**第二阶段**为词性统计阶段。**本阶段选做**。

简单步骤描述如下：

1. 统计每一句情感词。只统计褒义和贬义，可根据教师提供的中文褒贬义词典来统计。对情感度进行量化，找一个褒义词，褒义度加1，找到贬义词，贬义度加1.

附：打开文本文件使用第7章的open函数（p183），以及文件的读写方式。

2. 查找程度词。比如：很、非常、极度、无比、太等等。对每一褒贬义词前查找程度词，如果存在就增减褒贬义度。注意对程度词分等级，不同等级的程度词体现不同情绪强烈度。可以自行查找和定义程度词。

3.向后查找情绪符号。如感叹号，或者网络语言中常用的表情包。为简单期间，可以只查找感叹号。找到感叹号后，需要增加情绪度。

4.先前查找否定词。否定词可以翻转情绪极性。注意语言中的否定之否定。

5.综合各分句的情绪度。请确定整体情绪度的算法，并讨论它的合理性。

6.摘抄一些情绪倾向明显（褒义和贬义都需要），以及一些中性文本，检验的代码的表现。

请设计本阶段的算法，并以各种形式在实验报告中描述。

附加阶段：深入。有兴趣的同学可以参考自然语言分析（NLP）中的情感分析（sentiment）技术，需要牵涉到一些机器学习和数据挖掘的算法，如朴素贝叶斯，MaxEnt分类器，SVM，神经网络（深度学习）等等。

# 实验评分标准

本实验按以下要求评分：

1、完成度（40分）：第一阶段占60%，第二阶段占40%

2、实验报告（50分）：形式完整度30%，方案设计、分析和总结能力70%

3、拓展能力（10分）：按当前实验课题深入程度评分。

# 参考文献

[1] https://baike.baidu.com/item/%E6%96%87%E6%9C%AC%E6%83%85%E6%84%9F%E5%88%86%E6%9E%90/19431243?fr=aladdin