实验四：基于词典与基于深度学习的文本情感分类

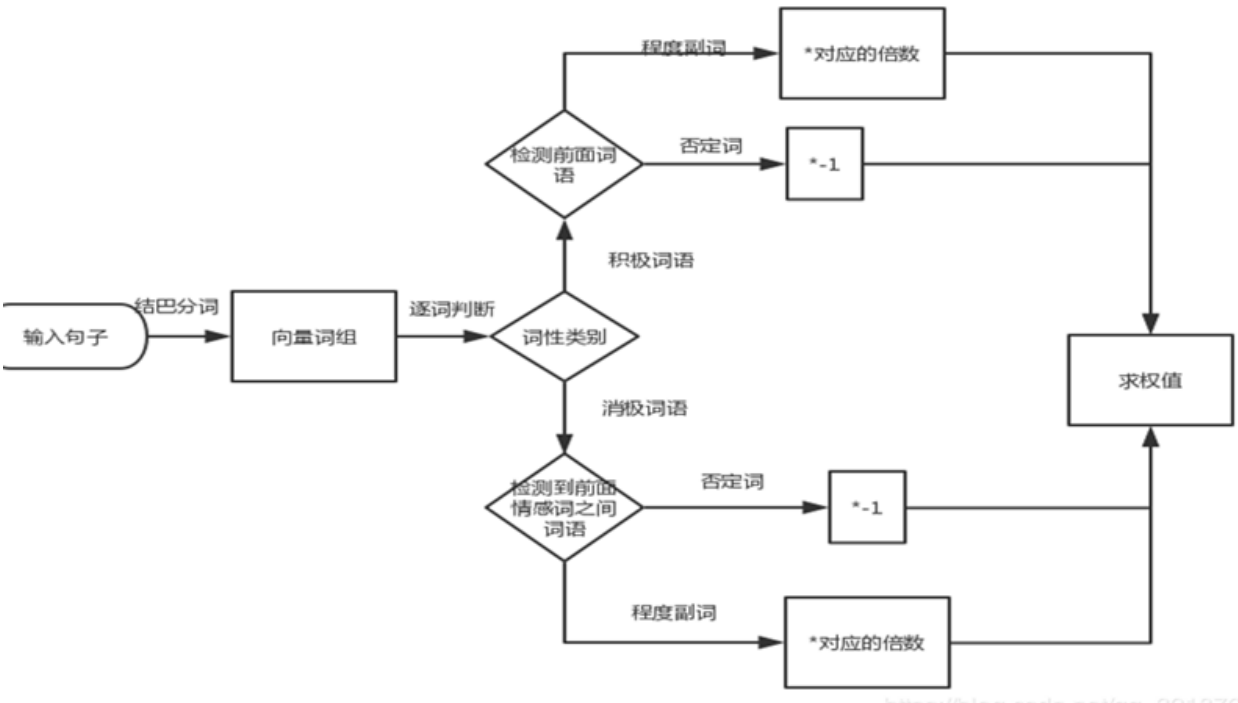
姓名：李鹏
学号：10175501102

一、基于词典的情感分类

1. 基于的词典

本次实验基于台湾大学提供的NTUSD中文简体情感词与知网提供的情感词典，对他们进行整合处理，提出情感词与程度副词，用于基于词典的情感分类。

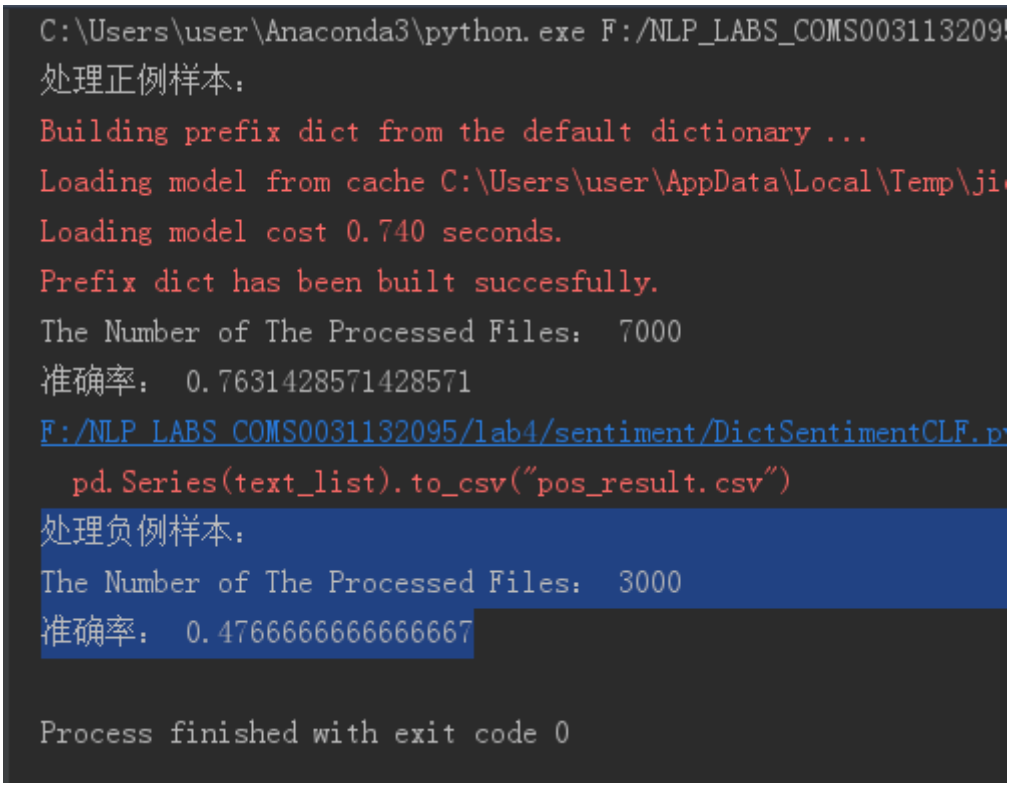
2. 实验流程



本次实验首先对文本进行了清洗，然后基本参照上述实验流程完成实验，只是在查找程度副词的时候全采用的是图的下半部分的流程。

3. 实验结果与结果分析

1) 实验结果



准确率：

- 正例样本：76.6%
- 负例样本：47.6%

2) 结果分析

虽然实验结果超出了随机判断的水平（33.3%，含中性），但是实验的准确率并不理想，实验过程下来，分析的主要原因有：

- **语料问题**：由于是爬取的语料，语料中的文本并不标准，有不少错别字，有的情感也比较复杂，很难判断总体是正面情感还是负面情感；
- **标注问题**：如在数据集中“ *距离川沙公路较近,但是公交指示不对,如果是”蔡陆线”的话,会非常麻烦.建议用别的路线.房间较为简单.* ”一句被标注为“正面”，个人觉得不算正面情感；
- **分词问题**：由于判断情感需要问题，但是现有的分词工具如“结巴分词”很有可能把表示情感的词给错误地划分，很难与情感词典很好地对应，情感词典由于没有其他标注，也很难作为自定义词典加入到分词工具之中去；
- **词典问题**：虽然已经对词典进行了整合，但是词典中的词并不是非常适合这个任务，有一些非常少用的词很少在酒店评论中用到，而评论中的一些表示情感的词在词典中也没有，这样就会影响实验结果；

二、基于机器学习的情感分类

基于机器学习的情感分类主要有特征提取与模型训练两个部分，在特征提取阶段，本次实验尝试提取了1-gram特征，1-gram与2-gram特征，1-gram、2-gram与3-gram特征三种TF-IDF特征，基于

大型语料预训练的词向量的，基于本身的语料训练的词向量几种特征，在模型训练方面，实验简单采用逻辑回归模型，一方面因为可以对比不同特征的表示效果，另一方面也是因为逻辑回归是深度学习的基础，鲁棒性也比较好。由于训练时间问题，本次实验没有采用Bi-LSTM、BERT等现在流行的深度学习的模型。

1. 数据集分析

在对正面情感的数据集进行分析的时候，一个主要的问题是数据集评论长度差异较大，最长的能够有954个词，最短的在数据清洗后只有1个词，平均有67个左右的词，这很大程度上影响了问题的建模，必须很难使用LSTM模型进行训练，这是一种短文本与长文本相间的问题。

```
    c.lenCounter(c.neg_sentence_len)
AttributeError: 'MLSentimentCLF' object has no attr
文档长度统计：
文档数：  7000
长度均值：  67.33028571428571
长度最大值：  954
长度最小值：  1

Process finished with exit code 1
```

2. 训练过程

- 数据集预处理：数据清洗、去除停用词等
- 训练集、验证集、测试集划分
- 训练TFIDF模型、训练词向量模型
- 对样本进行过采样（Oversampling）
- 使用逻辑回归模型对样本进行分类

其中细节请参考附件代码，本次实验中需要重点考虑的问题是词向量的维度的问题以及word2vec建模的时间问题

3. 实验结果与结果分析

1) 实验结果

本次实验测试了多组实验结果：

- 总数据：10000
- 训练集：4900
- 验证集：2100
- 测试集：3000

特征提取	训练集准确率	验证集准确率
1-gram + 1000特征 + TF-IDF	88.12%	86.33%
1-gram + 10000特征 + TF-IDF	91.04%	86.14%
1-gram、2-gram + 10000特征 + TF-IDF	91.27%	86.14%
1-gram、2-gram + 20000特征 + TF-IDF	91.84%	85.81%
1-gram、2-gram、3-gram + 10000特征 + TF-IDF	91.16%	86%
1-gram + 10000特征 + TF-IDF + 随机过采样	**92.61%**	**87.38%**
1-gram + 10000特征 + TF-IDF + SMOTE过采样	90.88%	85.76%
使用数据集进行训练的词向量，然后平均	69.61%	70.19%

选取上表中加粗的方式进行测试，最终的结果：

最终的测试集准确率为85.33%

注：

- 特征长度指的是tf-idf建模提取的单词数目；

2) 结果分析

- 逻辑回归模型比较简单，相对于现在的深度学习模型拟合能力还是比较差，在训练集上训练的结果都一般，这是结果一般的底层原因；
- 对于不平衡的数据，进行采样能够显著提升模型效果，但是不同的模型适合不同的采样方式，需要进行选择；

- 影响自训练的词向量的效果的主要是，语料数目太少，这一点语料很难训练出很好的向量表示；
- 本次实验代码中虽然实现了基于大型语料库的word2vec的编码表示，但是实验过程耗时太长，最终无法提供最后的准确率，但是传统的模型更适合模型建模与迭代；
- 虽然3-gram模型看起来更能体现出 “程度副词+情感词” 这种组合，但是实验表明在特征维度确定的情况下这种方式效果一般；