nlp实验一

2024春季学期

院系：人工智能学院

姓名：刘梦杰

学号：211300022

邮箱：2757400745@qq.com

实验时间：2024.4.20

目录

一、实验目的

实现对文本的人格分类测试

二、实验环境

在python3.11下，其中安装了nltk，genism，re，numpy，sklearn等包，其中nlp.py为未经预训练的产物，666.py为经过预训练的产物。

三、实验想法与步骤

1、Question1：用传统机器学习方法训练

由于将每个出现的词的词频统计作为维度过于浪费计算资源，且耗费时间较长。基于不同性格分类的人有不同的说话习惯，譬如xxxx人格可能更倾向于多用形容词，xxxx人格可能更喜欢简单的陈述，假设不同人格的人所代表的文本含有各种词性的单词的占比是不同的。采取统计每个文本中英文单词的词性的方法，将不同的词性作为训练集的不同维度，导入nltk包作为鉴别词性的工具，与此同时，利用MLP训练所得训练集，将四种分类分别用于二分类，最后合并在一起，所得准确率在0.18左右，低于baseline，不过缩短了训练时间。

2、Question2：预训练

本来希望使用bert来表示出向量，但是不知道是否是下载的数据集问题，在简单尝试使用bert时发生了如下情况：

```
from bert_serving.client import BertClient
print("666")
bc = BertClient()
print(bc.encode(['First do it', 'then do it right', 'then do it better']))
```

[1] 244m 40.2s

```
666
```

+ 代码 + Markdown

于是转而使用genism中的doc2vec对文本进行向量化，具体为：

```
model = Doc2Vec(documents, vector_size=100, window=5, min_count=3, workers=4)
model.train(documents, total_examples=model.corpus_count, epochs=20)
```

代码中向量化的数据集其实就是训练集，再使用函数model.infer_vector()对测试集中文本进行向量化（如果采用更大的向量化的数据集或许结果会更好），最后使用逻辑斯蒂回归的准确率为0.23左右。

3、question3：

预训练会比传统的机器学习方法好。一方面，直接采用词袋模型，在存储和运算维度上的数据量要远远高于预训练以后的数据，

另一方面准确率也有所上升，说明预训练以后对原文本的表达更准确。

五、一些思考

在处理文本的过程中，我们都是将标点去掉，但是实际上标点符号是否也在判别过程中起到一些作用呢，尤其是在情感这方面，譬如反问句的语气更加强烈，是否应该将之作为思考的一部分？