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：2024.4.24

目录

一、实验目的

探究HPR任务的难点

二、实验环境

在python3.11下，其中安装了genism，re，numpy，sklearn等包

三、实验想法与步骤

1、Question1：HPR任务的难点

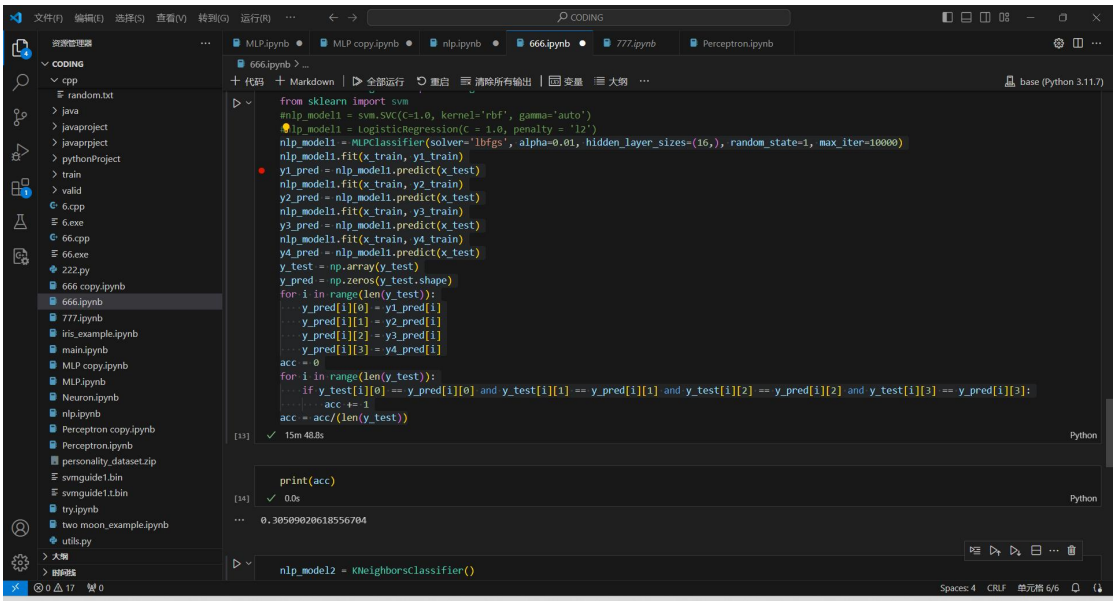
（1）找到一种有效表示文本特征的方法是困难的，简单的词袋模型中的向量过于稀疏，且对关键的词语基本不能有效筛选，就算选用了较为精确的模型如bert之类的进行预训练，也无法保证训练出来的向量与MBTI有很强大的相关性；

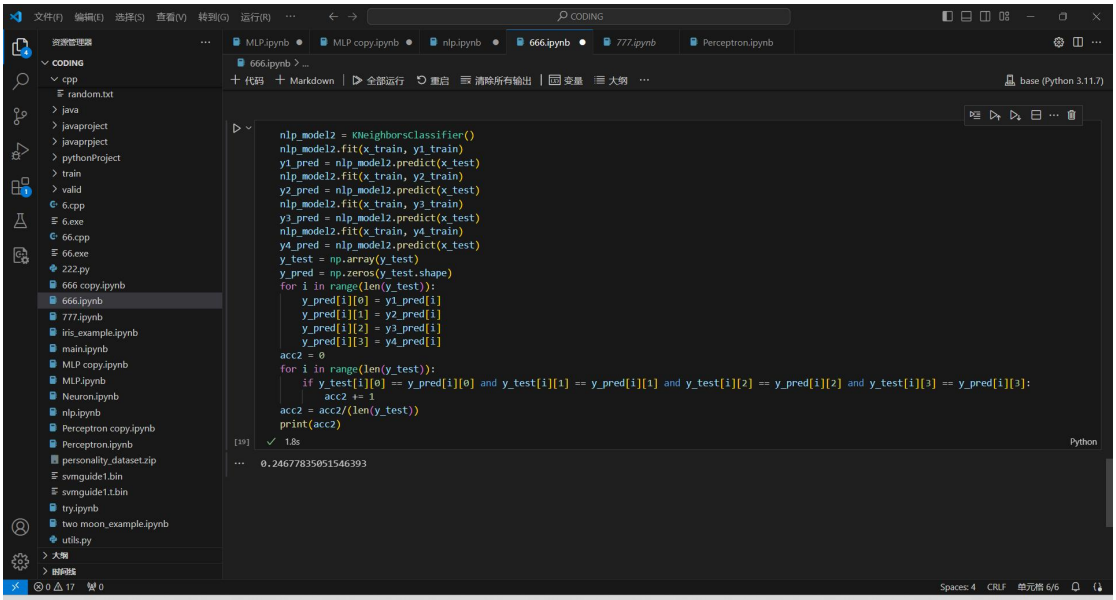
（2）mbti中的16型人格也不一定全面，存在更细划分导致同一性格内的句向量差异较大；

2、Question2：简单的实验

（1）验证想法一很简单，qq群中助教使用bert进行预训练的准确率远高于我实验代码中的0.3，更是远高于不进行预训练的；

（2）验证实验二也很粗暴，使用KNN来预测模型的准确率仅有0.24几，同样的训练集与测试集差距如此之大与样本本身的关系不可谓不大。




nlp_model2 = KNeighborsClassifier()
nlp_model2.fit(x_train, y1_train)
y1_pred = nlp_model2.predict(x_test)
nlp_model2.fit(x_train, y2_train)
y2_pred = nlp_model2.predict(x_test)
nlp_model2.fit(x_train, y3_train)
y3_pred = nlp_model2.predict(x_test)
nlp_model2.fit(x_train, y4_train)
y4_pred = nlp_model2.predict(x_test)
y_test = np.array(y_test)
y_pred = np.zeros(y_test.shape)
for i in range(len(y_test)):
y_pred[i][0] = y1_pred[i]
y_pred[i][1] = y2_pred[i]
y_pred[i][2] = y3_pred[i]
y_pred[i][3] = y4_pred[i]
acc2 = 0
for i in range(len(y_test)):
if y_test[i][0] == y_pred[i][0] and y_test[i][1] == y_pred[i][1] and y_test[i][2] == y_pred[i][2] and y_test[i][3] == y_pred[i][3]:
acc2 += 1
acc2 = acc2/(len(y_test))
print(acc2)
0.24677835051546393