

自然语言处理课程大作业报告

—— 项目名称：基于sklearn实现多标签文本分类任务

姓名：李长博

班级：硕计241班

学号：220242221045

日期：2024年11月20日

1. 任务背景

在数字化时代，文本数据的爆炸性增长带来了信息过载的问题。用户和系统常常需要从海量数据中快速识别和检索信息，这就需要有效的文本分类技术来辅助。多标签文本分类是一种特殊的文本分类任务，它允许单个文档同时被归类到多个类别中，这在现实世界的应用中非常普遍。

本项目旨在开发一个基于机器学习的多标签文本分类系统，该系统能够自动、准确地为文本数据分配多个相关标签，以支持上述应用场景。通过使用Python语言和Scikit-learn库，项目将实现一个高效、可扩展的分类模型，以处理实际的多标签文本分类问题。本项目使用的数据集存储在JSON文件中，包含文本数据和对应的标签。

1. 模型原理

在本项多标签文本分类项目中，我主要采用的是基于支持向量机（SVM）的One-vs-Rest策略。以下是模型原理的详细介绍：

1、支持向量机（SVM）：

①SVM是一种监督学习算法，主要用于分类问题。它的基本原理是在特征空间中寻找一个最优的超平面，以此来区分不同类别的数据点。在二维空间中，这个超平面可以被看作是一条直线，而在更高维的空间中，它是一个平面或者更复杂的结构。

②SVM的核心思想是最大化两个类别之间的间隔，即找到使得间隔最大的分割超平面。这个间隔被称为“边缘最大化”，它提高了模型对于未知数据的泛化能力。

2、One-vs-Rest策略：

①由于SVM本身是二分类算法，为了处理多标签分类问题，我们采用了One-vs-Rest策略。这种策略将每个标签视为一个独立的二分类问题，即对于每个标签，我们训练一个SVM模型来区分该标签为正例和负例的情况。

②在预测阶段，对于每个输入样本，所有SVM模型都会给出一个预测结果，如果一个模型预测为正例，则该样本被认为属于对应的标签。

3、特征提取：

①在文本分类任务中，我们使用TfidfVectorizer来提取文本特征。TF-IDF（Term Frequency-Inverse Document Frequency）是一种统计方法，用以评估一个词语对于一个文件集或一个语料库中的其中一份文件的重要程度。

②TF-IDF向量化器将文本数据转换为TF-IDF特征向量，这些向量能够反映词语在文档中的重要性，并且减少了不同长度文档之间的影响。

4、多标签处理：

多标签分类问题中，一个样本可能同时属于多个类别。为了处理这种情况，我们使用MultiLabelBinarizer将多标签数据转换为二进制形式，每个标签对应一个二进制向量，向量中的1表示样本属于该标签，0表示不属于。

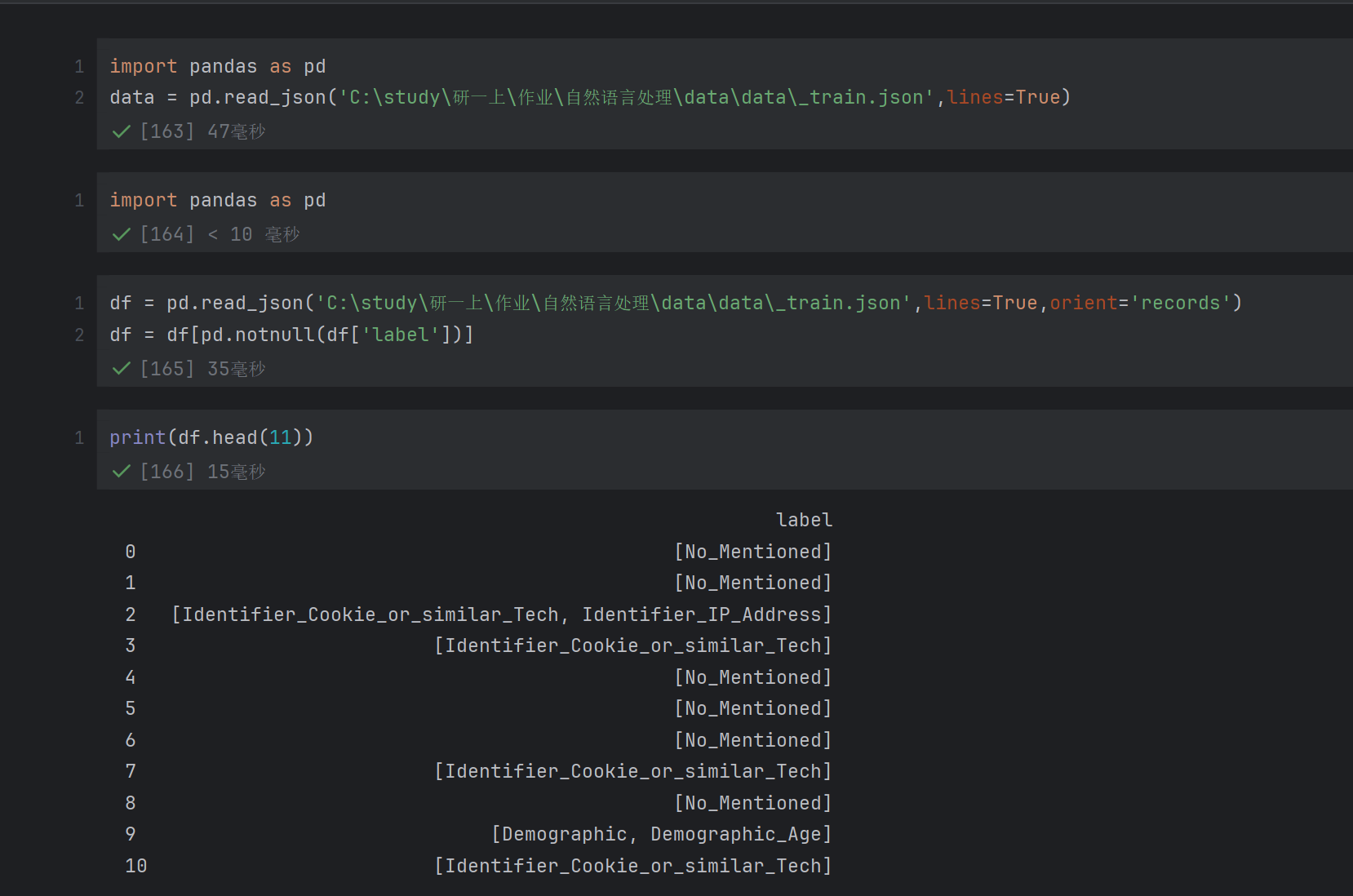
5、模型评估：

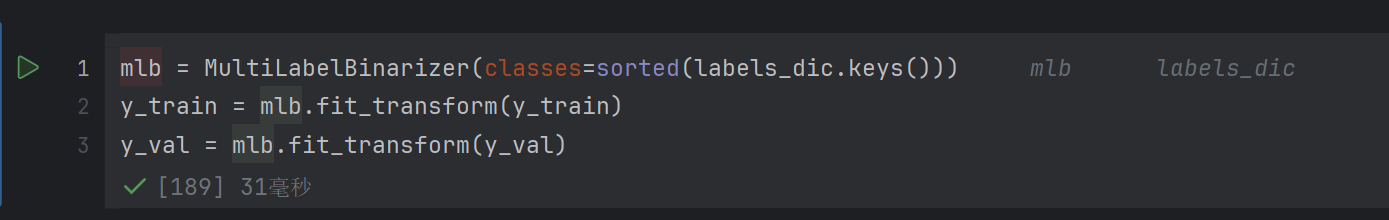
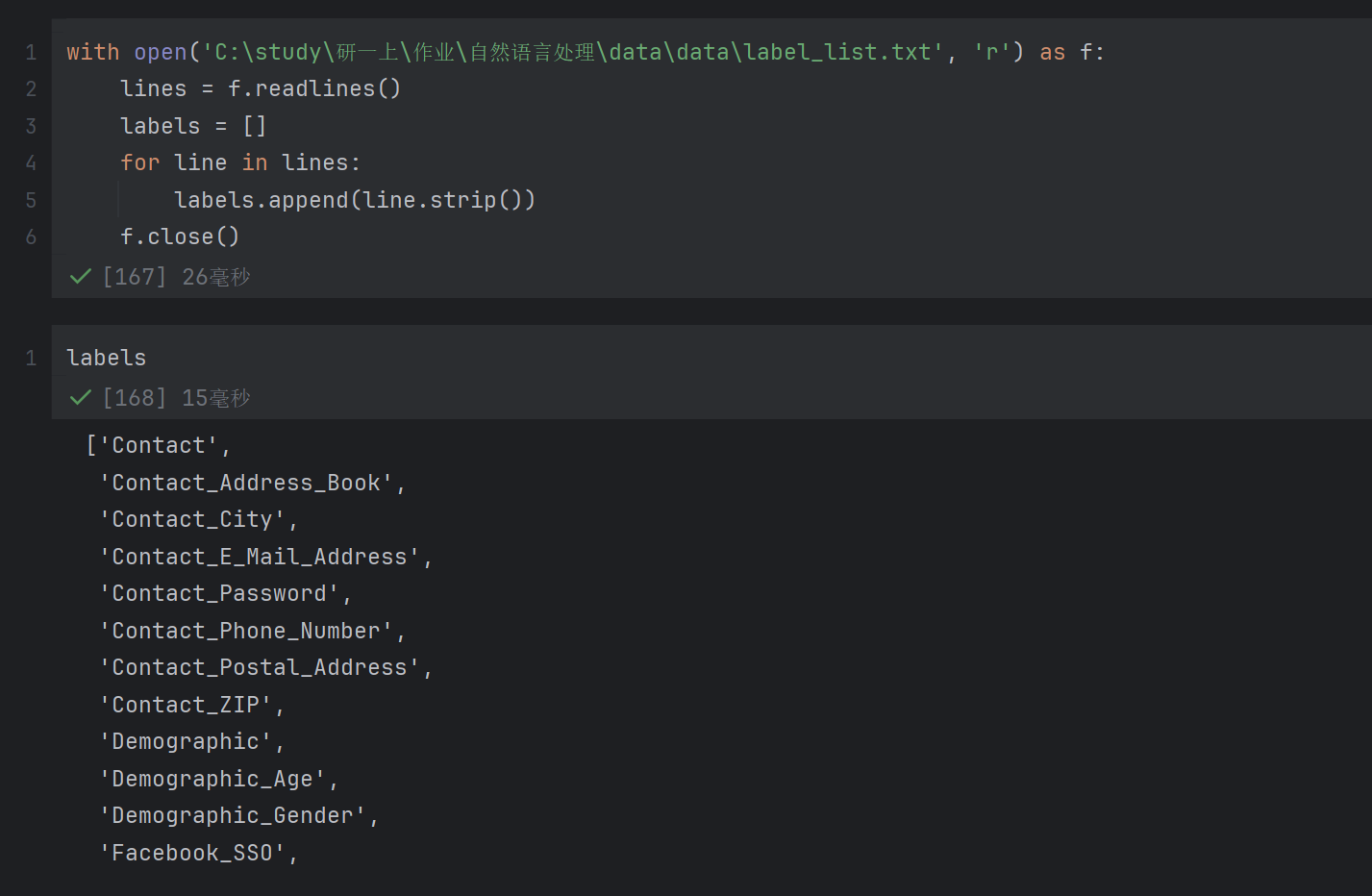
在模型评估阶段，我们使用准确率和宏平均F1分数作为评估指标。准确率衡量模型预测正确的标签比例，而宏平均F1分数衡量模型在所有类别上的整体性能，特别是对于那些不平衡的类别。

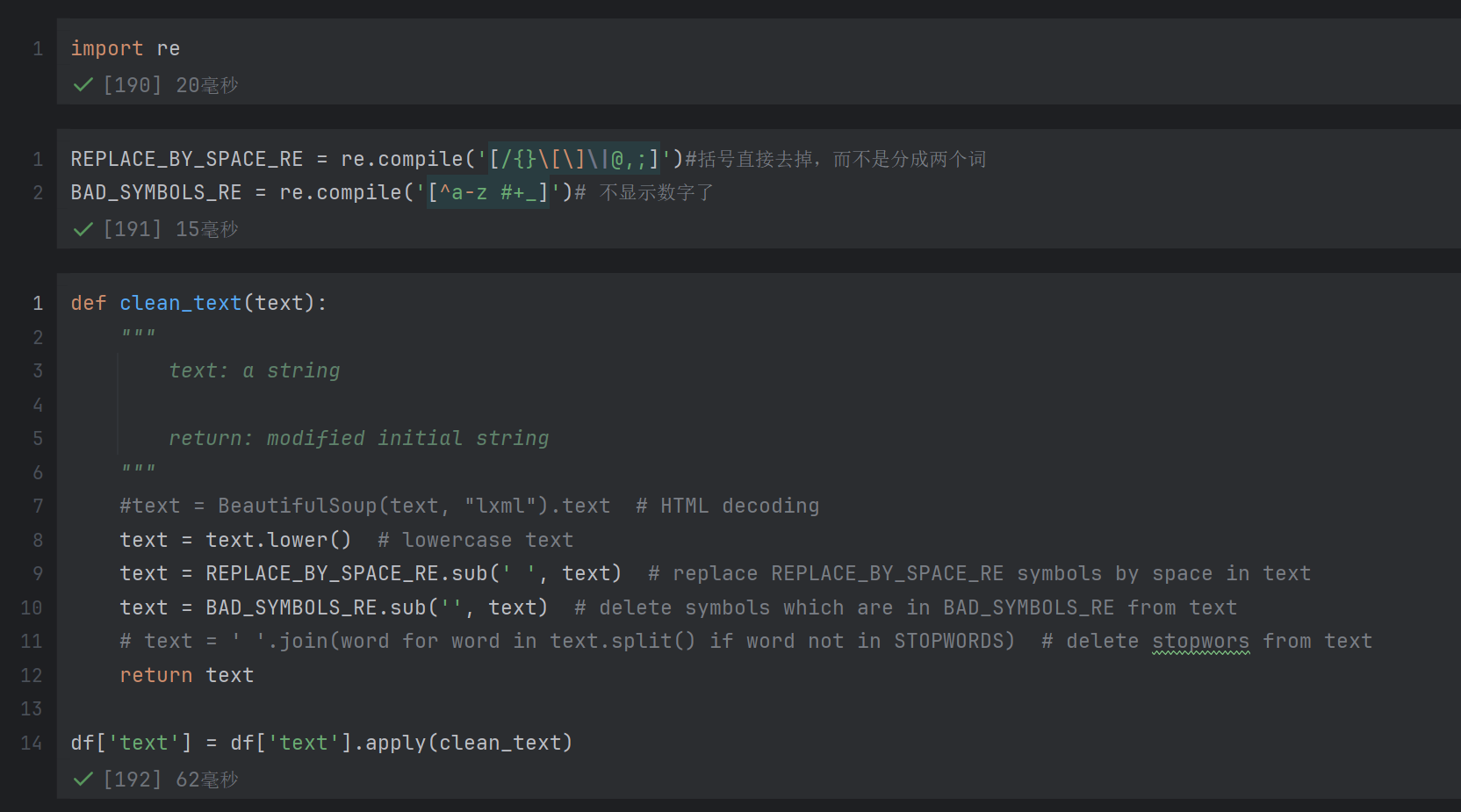
1. 实验设置

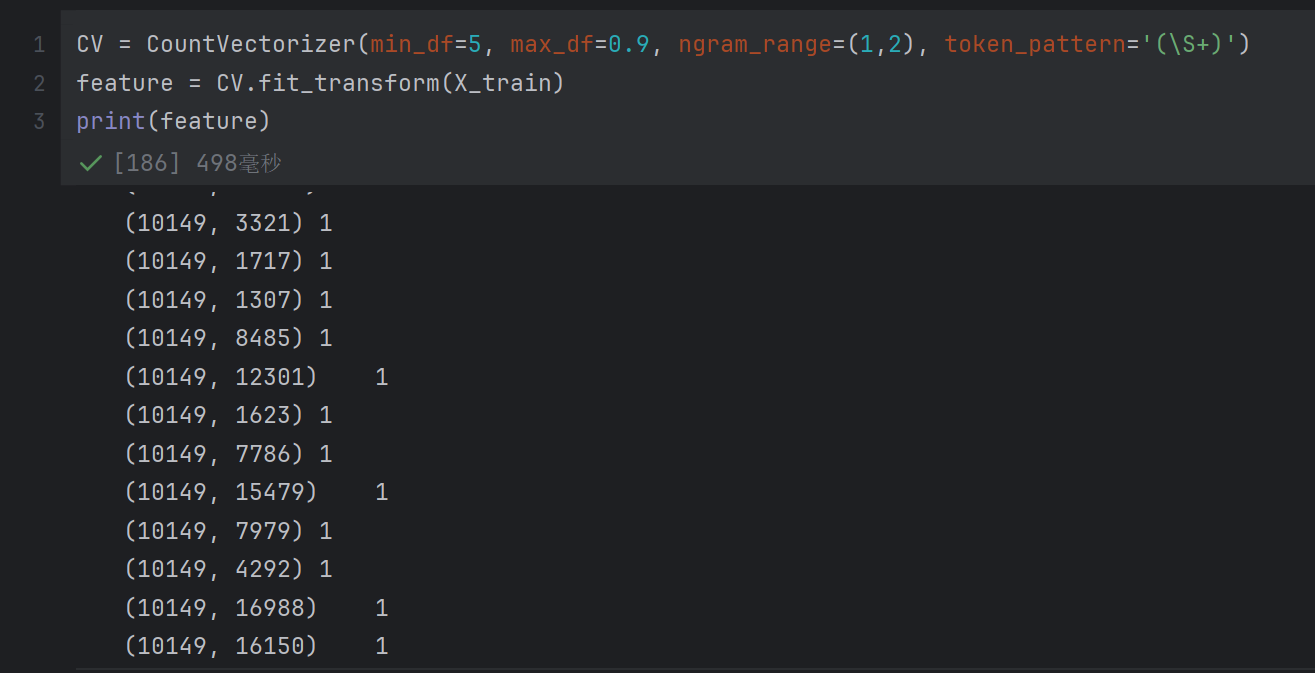
实验设置分为以下几个详细步骤：

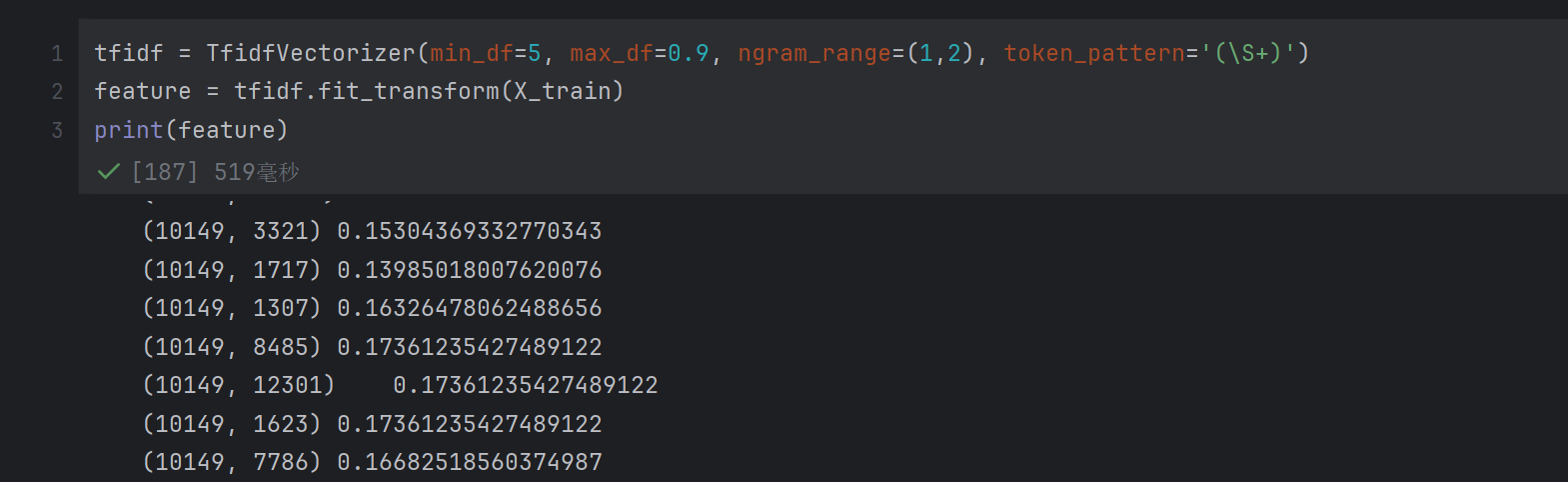
1、数据准备：导入本地文件并读取train.json文件来进行数据收集。

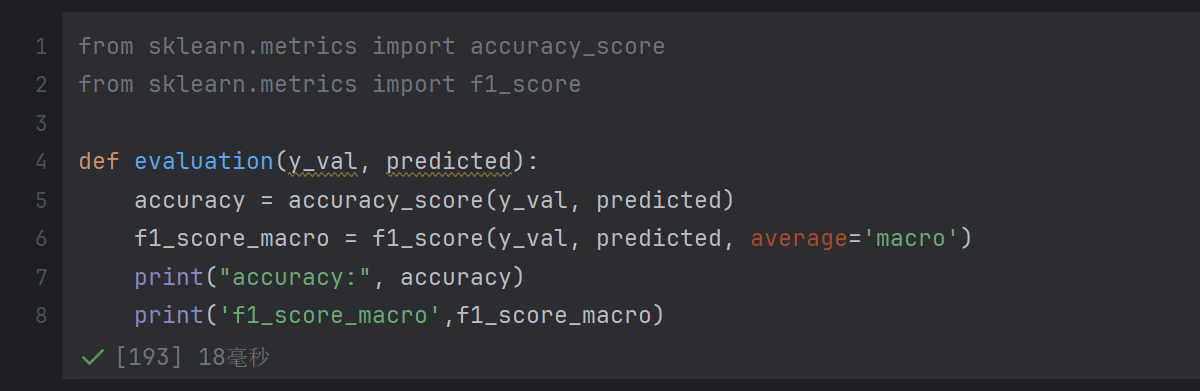
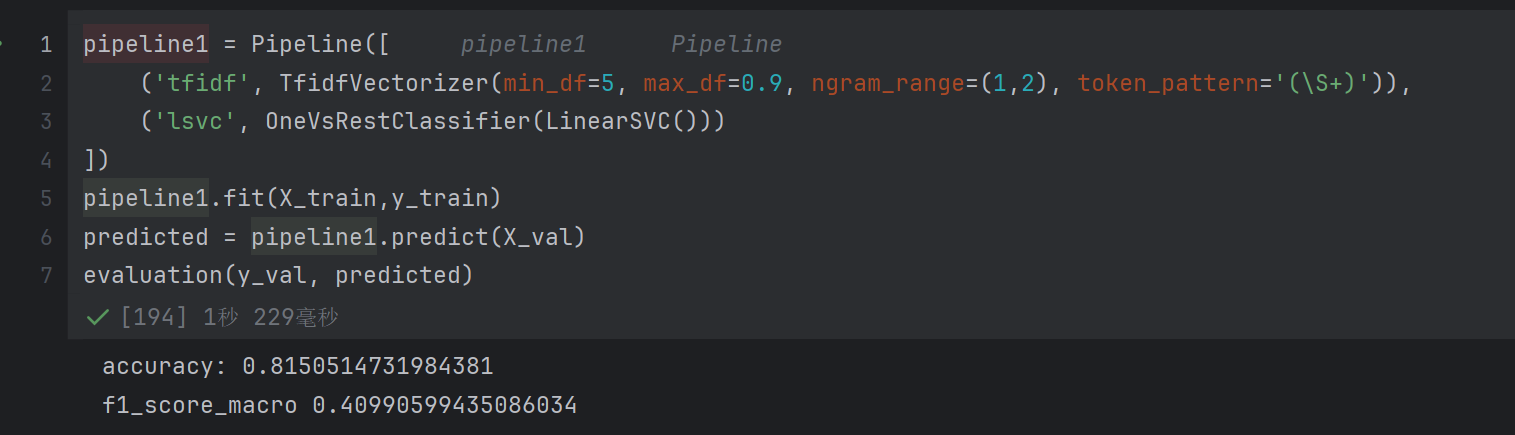


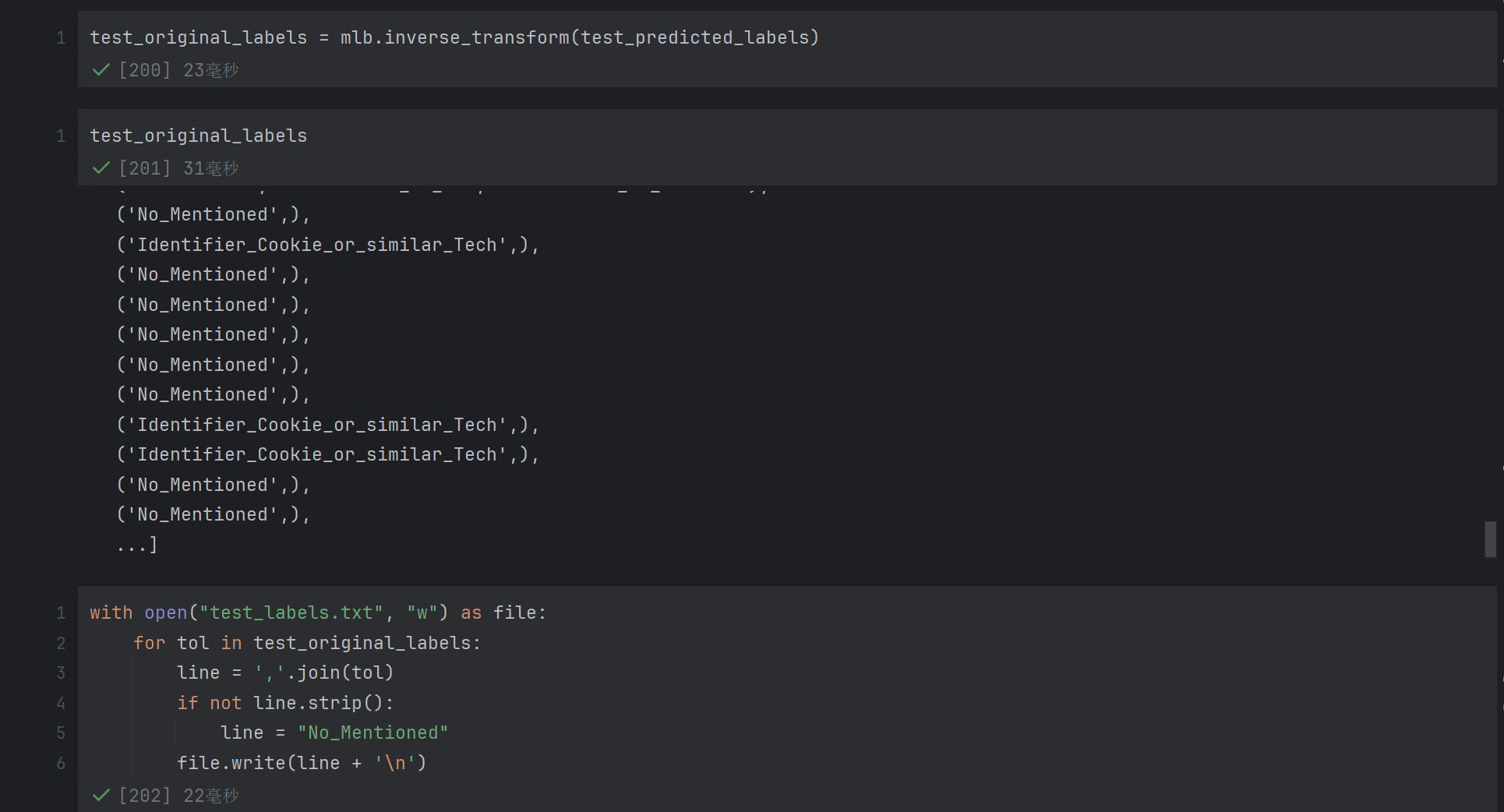
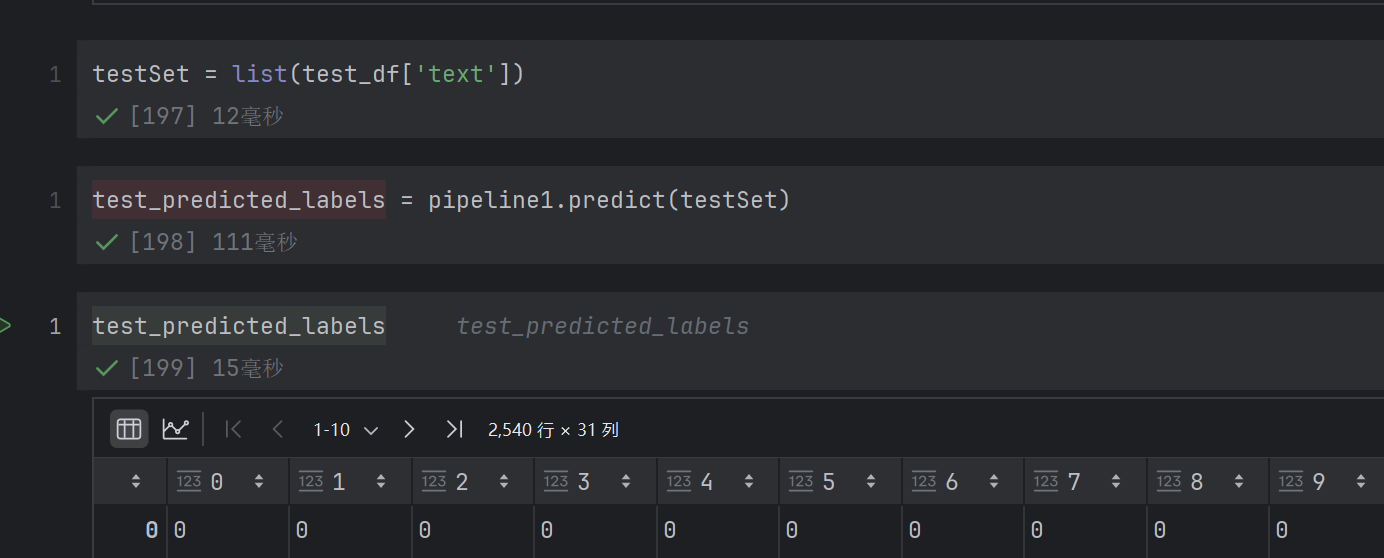
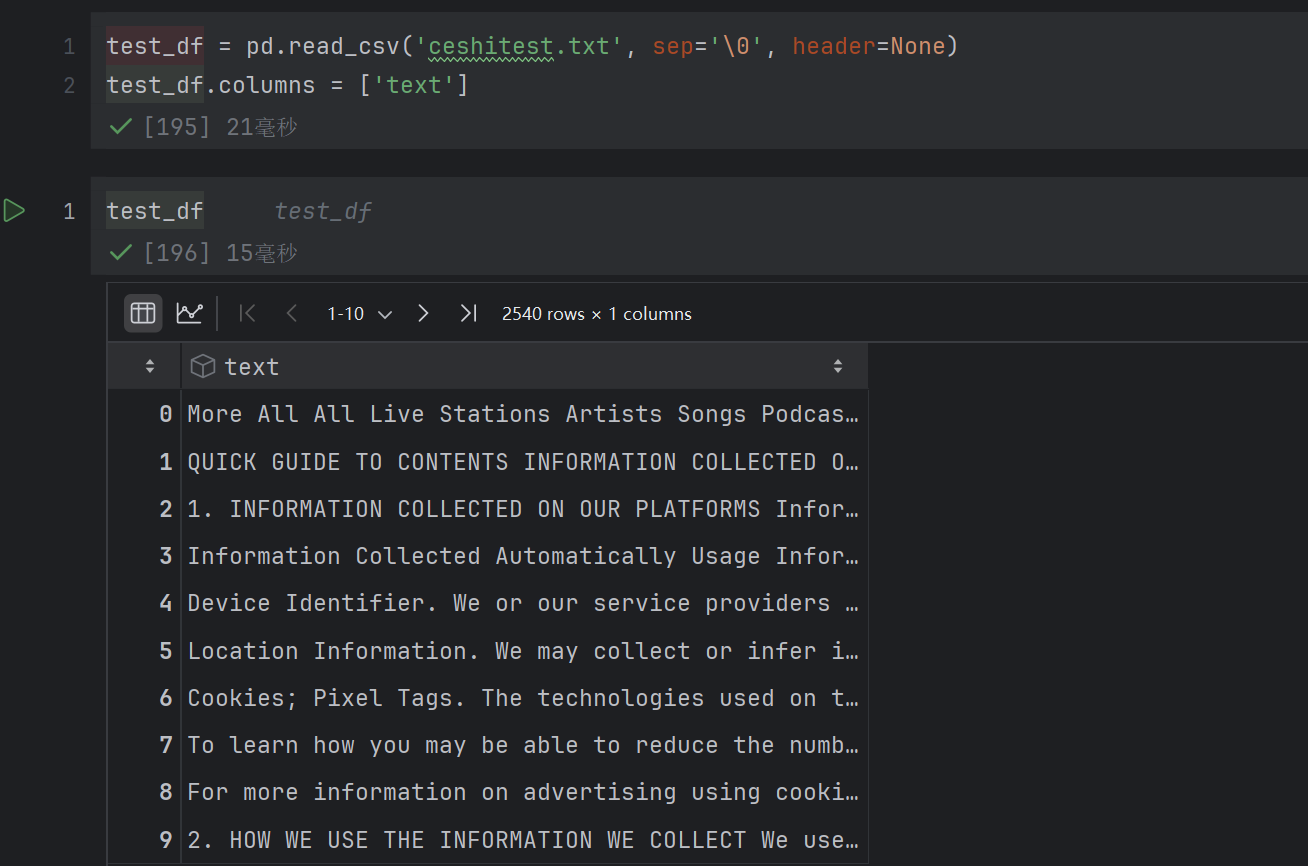
2、标签处理：包括构建标签字典labels\_dic以及对标签二值化

3、特征提取：包括文本预处理以及特征向量化

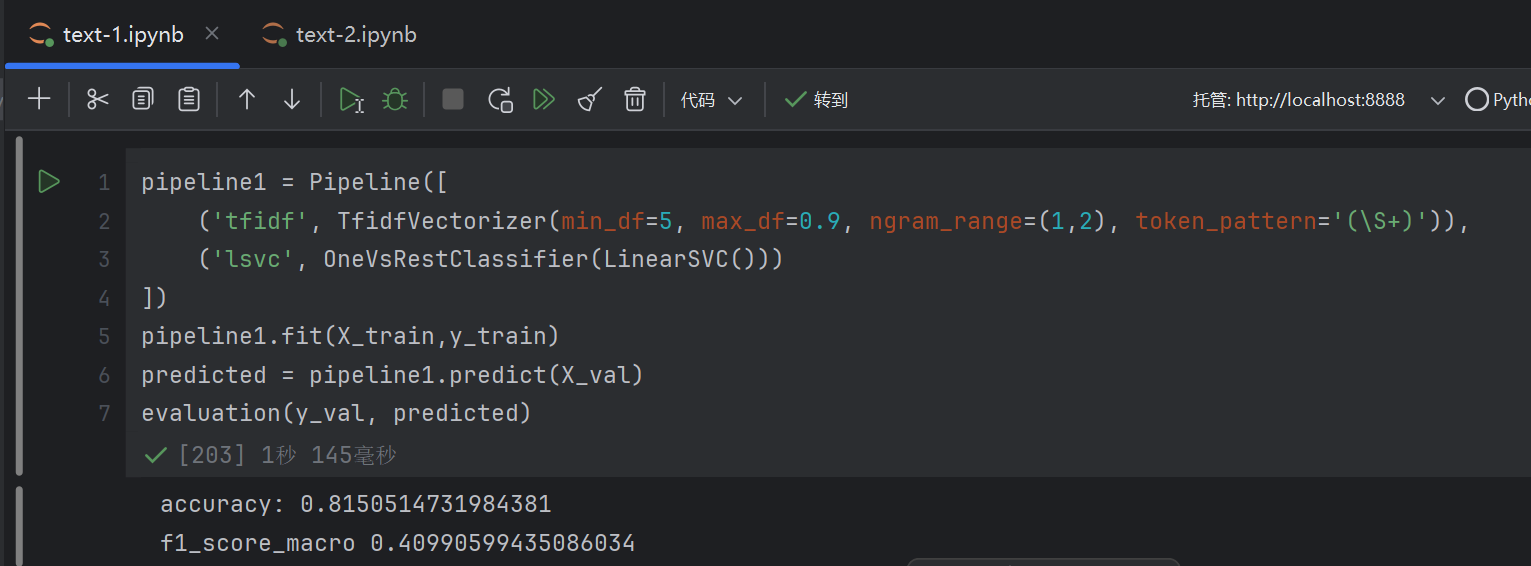
计算标签及对应个数，将train的文本和标签都转化为list，之后将文本中的词语转换为词频矩阵，计算各个词语出现的次数。

特征向量化，使用TfidfVectorizer将预处理后的文本转换为TF-IDF特征向量

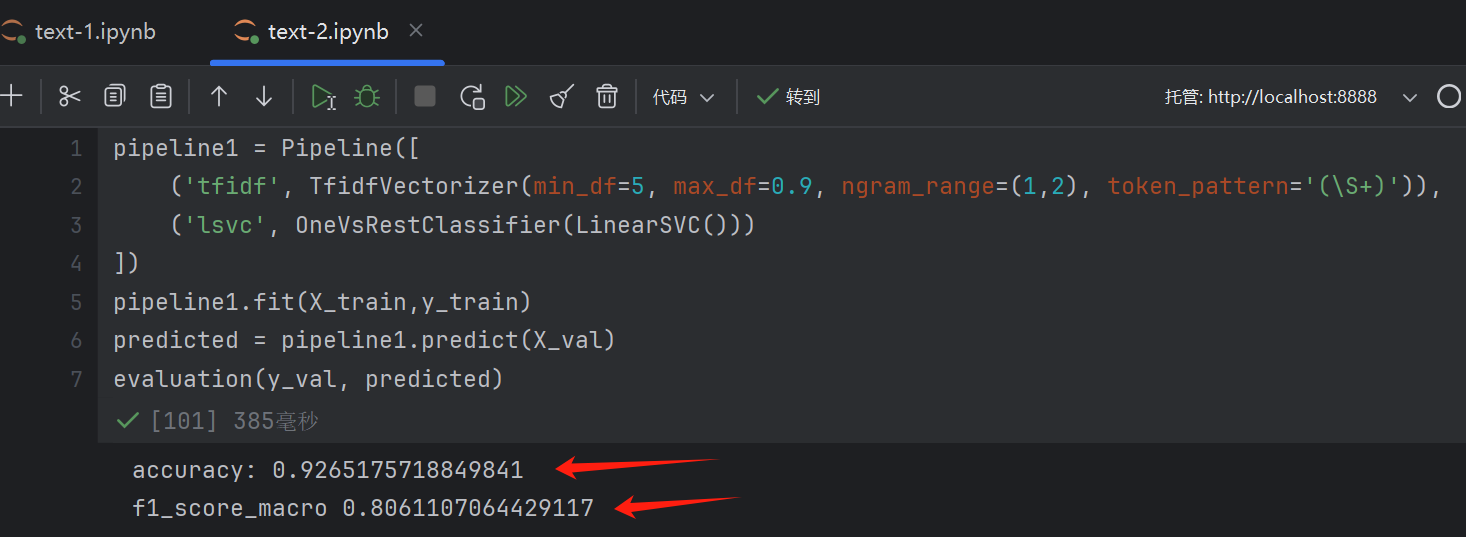
4、模型构建：选择LinearSVC作为基础分类器，使用OneVsRestClassifier将LinearSVC包装起来，构建一个Pipeline，将TfidfVectorizer和OneVsRestClassifier包装在一起，并且编写评估函数，计算accuracy

5、读取text、并且进行预测并将结果写入新的txt文件

四、结果分析

在训练集第一次运行时，模型的准确度和宏平均F1分数相对较低，如图

也尝试更改过模型中的参数，但还是ngram\_range=(1,2)，TfidfVectorizer(min\_df=5, max\_df=0.9, 准确度相对较高

在验证集上测试该模型的性能，结果其实还算可以

五、总结

本项目通过构建一个基于SVM的多标签文本分类系统，实现了对文本数据的自动化分类。通过实验设置和结果分析，我验证了模型的有效性，并尝试过多个参数组合后，找到了相对较高的参数配置。

Ps:学生的学艺不精，项目最后并没有找到最优的参数设置，这个项目作为我学习机器学习并且亲手做的第一个小项目，虽然是很简单的一个项目，但是仍然收获很多。虽然我的研究领域并不是机器学习方面的，但同样非常感兴趣，我在本门课收获了很多知识，而且在完成这项作业中也收获了很多技能，受益匪浅。继续加油！