**题目：基于朴素贝叶斯 的文本分类实验**

**学 号： 2021211012**

**班 级： 2021211303**

**姓 名： 熊佳进**

**专 业： 计算机科学与技术**

**导 师： 王晓茹**

**学 院： 计算机学院**

**2024 年 1 月 11 日**

**一、实验目的**

1. 掌握数据预处理的方法，对训练集数据进行预处理；

2. 掌握文本建模的方法，对语料库的文档进行建模；

3. 掌握分类算法的原理，基于有监督的机器学习方法，训练文本

分类器；

4. 利用学习的文本分类器，对未知文本进行分类判别；

5. 掌握评价分类器性能的评估方法。

**二、实验类型**

数据挖掘算法的设计与编程实现。

**三、实验要求**

1. 文本类别数：>=10 类；

2. 训练集文档数：>=50000 篇；每类平均 5000 篇。

3. 测试集文档数：>=50000 篇；每类平均 5000 篇。

4. 分组完成实验，组员数量<=3，个人实现可以获得实验加分。

**四、实验内容**

利用分类算法实现对文本的数据挖掘，主要包括：

1. 语料库的构建，主要包括利用爬虫收集 Web 文档等；

2. 语料库的数据预处理，包括文档建模，如去噪，分词，建立数据字典，

使用词袋模型或主题模型表达文档等；

注：使用主题模型，如 LDA 可以获得实验加分；

3. 选择分类算法（朴素贝叶斯（必做）、SVM/其他等），训练文本分类器，

理解所选的分类算法的建模原理、实现过程和相关参数的含义；

4. 对测试集的文本进行分类

5. 对测试集的分类结果利用正确率和召回率进行分析评价：计算每类正确

率、召回率，计算总体正确率和召回率。

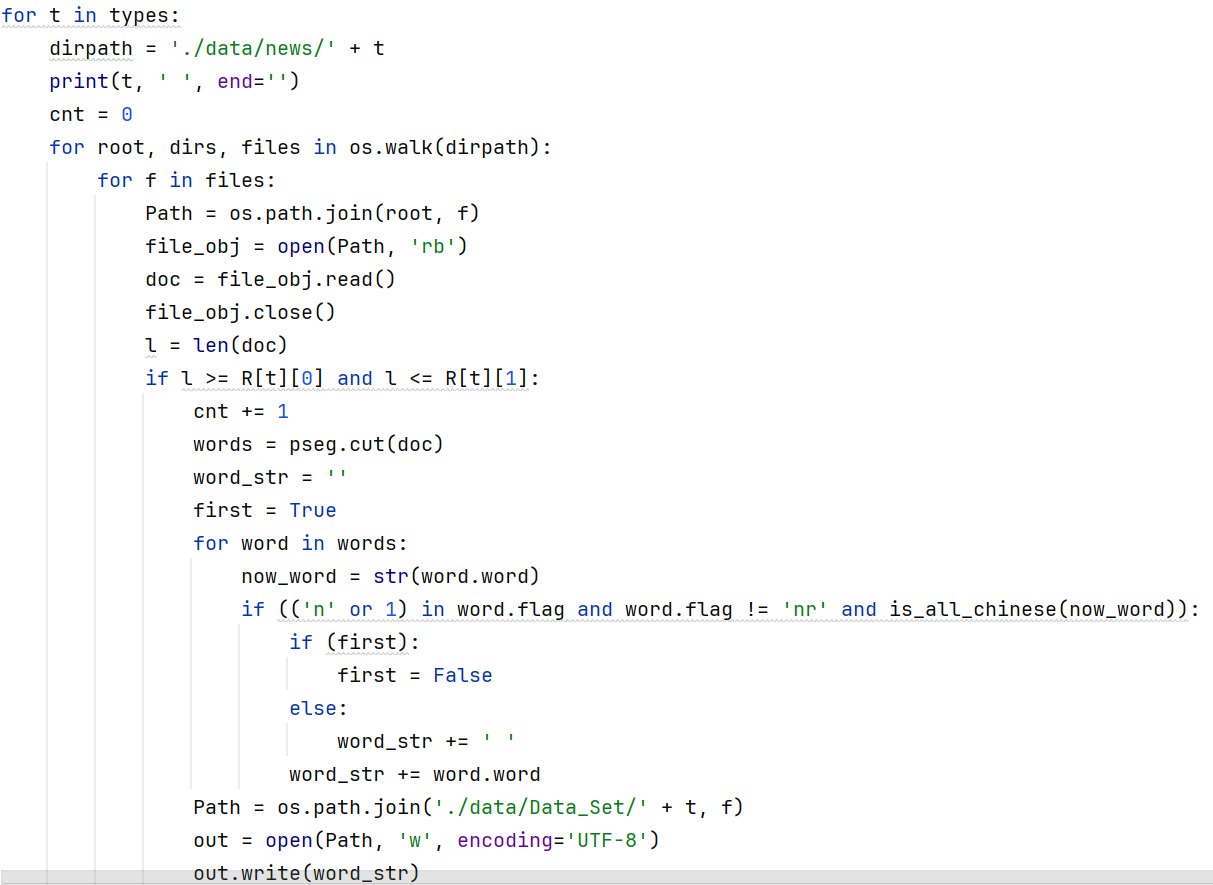
**五、实验步骤**

（一）数据收集

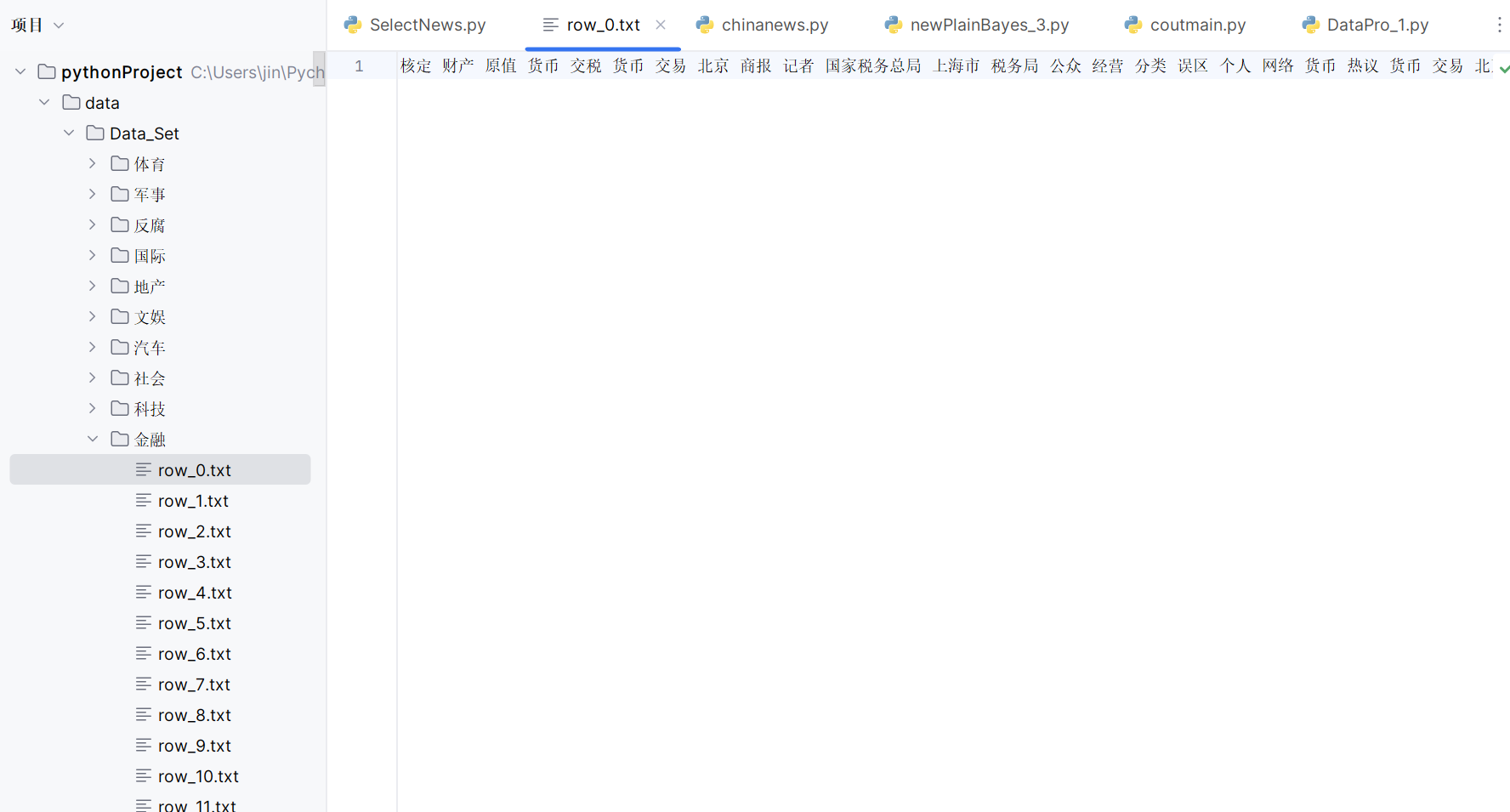
在本次实验中，我通过自编爬虫，获取中国新闻网的新闻文本。



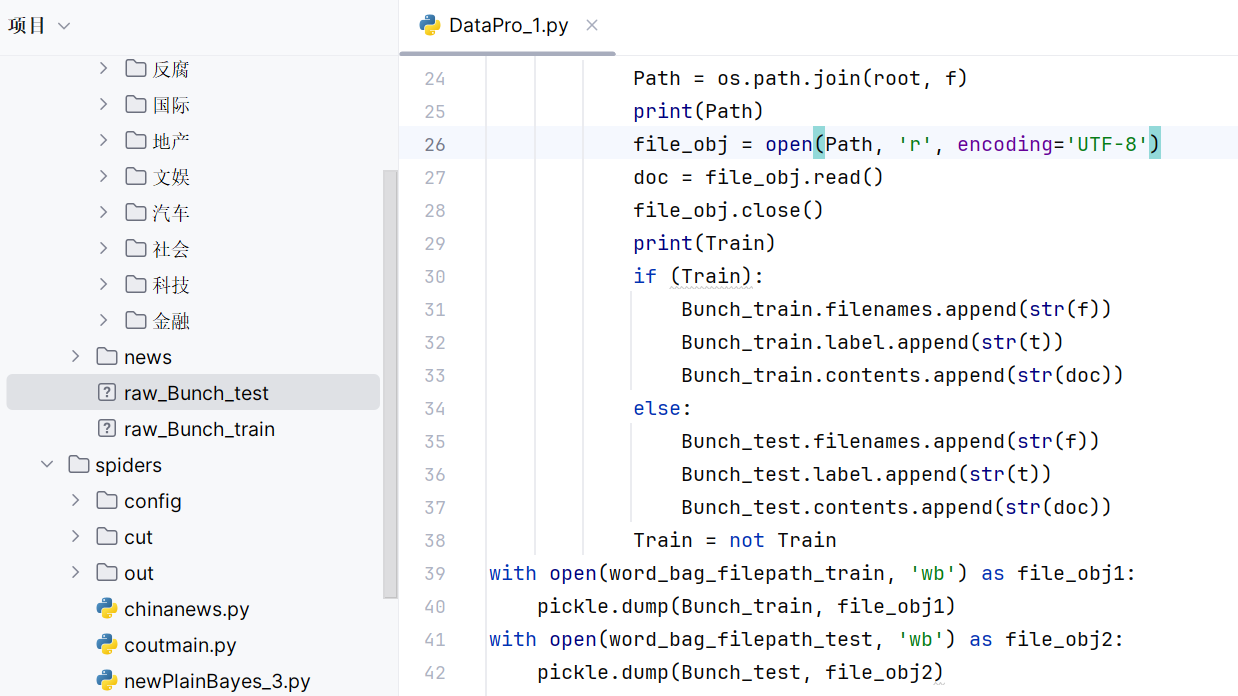
（二）数据预处理



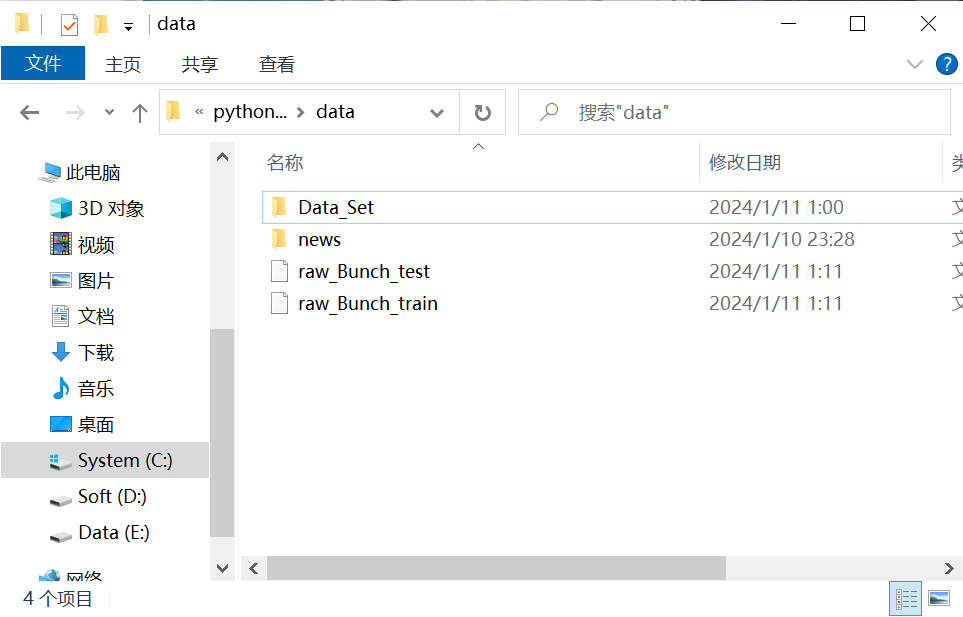
得到完成分词的10类文本，每类一万篇。



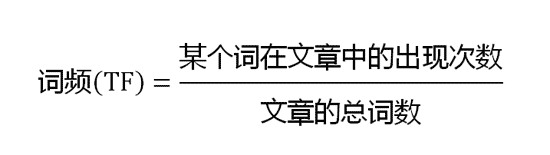
将每类文档交替存入两个词袋模型，用于划分相等数量的训练集与测试集。

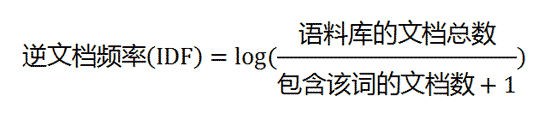


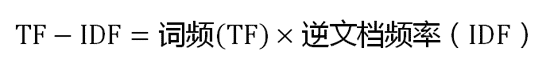
结果得到raw\_Bunch\_test与raw\_Bunch\_train两个原始词袋文件



使用TF-IDF模型提取训练集中的特征词。TFIDF的主要思想是：如果某个词或短语在一篇文章中出现的频率TF高，并且在其他文章中很少出现，则认为此词或者短语具有很好的类别区分能力，适合用来分类。计算数据集中词频（TF）和逆文档频率（IDF），两者相乘得到TF-IDF。



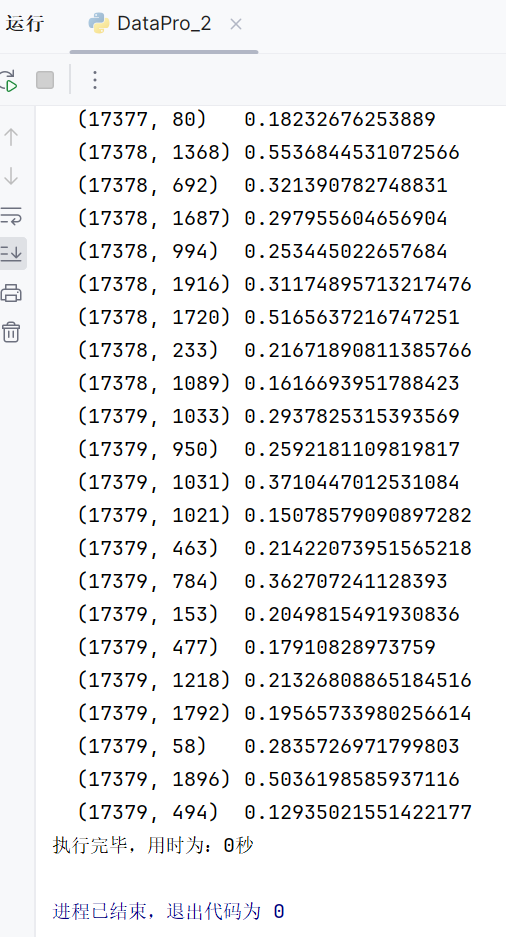




处理为TF-IDF模型后，利用TfidfVectorizer方法得到指定维度的字典为2000，得到数据字典，并生成训练集、测试集词袋模型。其中在生成TF-IDF的TfidfVectorizer指定去除在少于0.1%的文档中出现的词，在高于15%的文档中出现的词。本实验共选择10类文本，每类占比10%，如果一个词在15%以上的文档中都出现，说明它在不止一个类别中多次，那么该词语不具有好的代表性，应该去掉。对于少于0.1%的词，同样不具有代表性。



TF-IDF权重矩阵

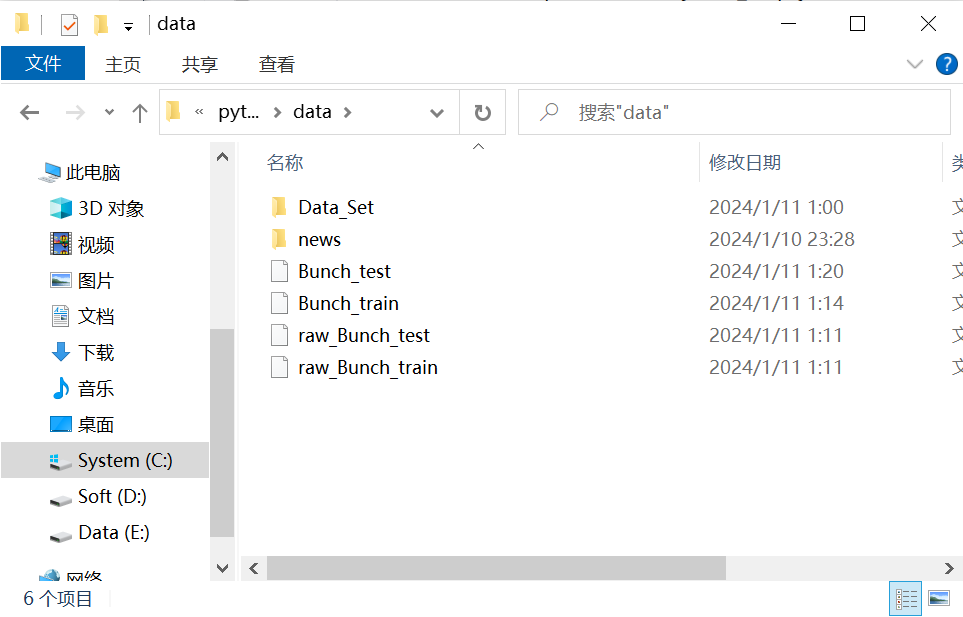


（用numpy稀疏矩阵表示，(i,j)代表第i篇测试文档在第j维向量上的特征值）

用训练集得到的字典生成测试集的特征向量。

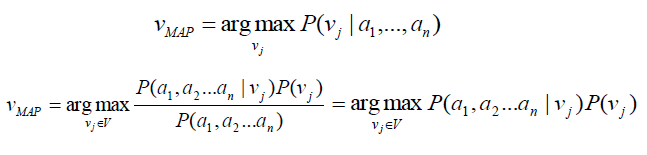


得到训练集与测试集的词袋文件



（三）朴素贝叶斯分类器

利用贝叶斯公式，在给定描述实例的属性值，得到最可能的目标值。



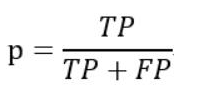
在此基础上，利用TF-IDF进行改进，使用TF-IDF将测试集中单词出现的次数转为一个0~1的TF-IDF权值，它在原贝叶斯方法的基础上额外考虑的单词的出现频率（重要程度）。与此同时取对数减少乘法运算次数，并在对数内加1以处理零概率。因此，计算后验概率式子变为



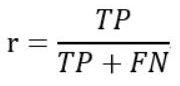
（为当前测试集第i个词的TF-IDF值）。

对于正确率、召回率、F1-score的计算。

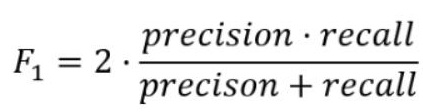
正确率（precision）指的是预测值为1且真实值也为1的样本在预测值为1的所有样本中所占的比例。

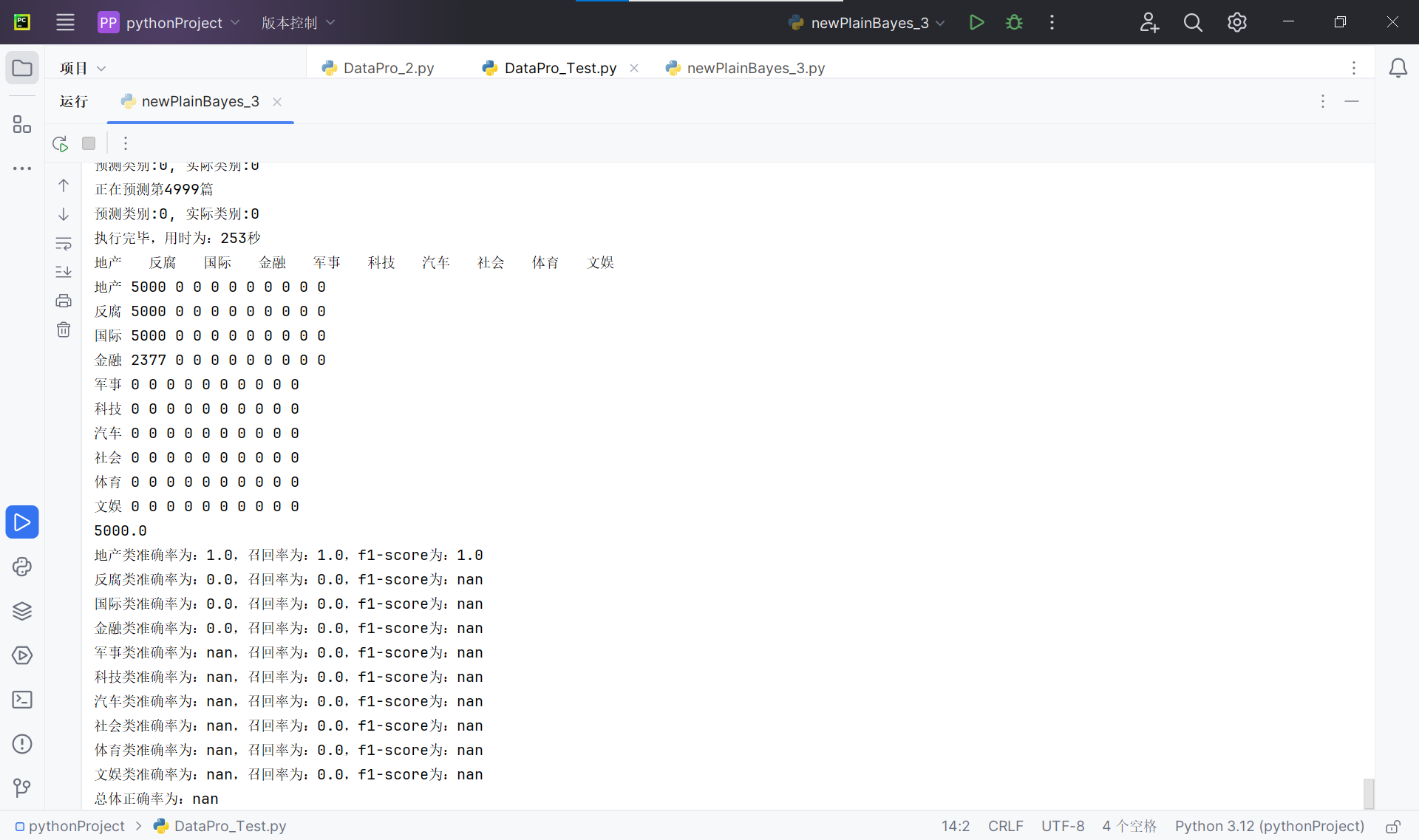


召回率（recall）指的是预测值为1且真实值也为1的样本在真实值为1的所有样本中所占的比例。



F1-score，又称为平衡F分数，是正确率和召回率的调和平均。





（四）思考与体会  
在本次文本分类实验中，我对于朴素贝叶斯的原理及实现过程有了更加深刻的理解，在编程实现朴素贝叶斯的过程中，通过对其进行相关改进的尝试，体会到了数据预处理对于实验繁杂程度及实验结果的损失程度的影响，认识到了人工智能对于数据的依赖能力；在调包实现SVM的过程中，通过对比朴素贝叶斯和libsvm，了解了不同分类器的不同思路，培养了对人工智能、数据科学的兴趣，提高了学习、汲取、进步的能力，受益匪浅。