**作业文档命名方式：学号\_姓名\_课堂作业7**

1. **实验目的**
2. 掌握jieba库的使用，学会中文分词、词性标注、特征提取。
3. 掌握词袋模型特征和TF-IDF模型。
4. 了解案例：垃圾邮件识别。
5. **实验环境**

操作系统：Windows

应用软件：Python, Anaconda, Jupyter Notebook

1. **实验内容与结果**（题目、源程序、运行结果）

1.文档集包含以下5条文本**：**

"我是中国人，我爱中国"

"我是上海人"

"我住在上海松江大学城"

"松江大学城有很多大学"

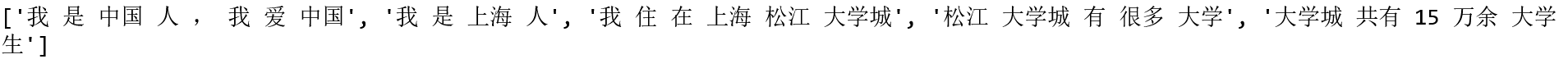
"大学城共有15万余大学生"

提取以上文档集的词袋模型特征（包含字符数为1的词作为特征词）和TF-IDF模型特征。

对于文本数据计算机无法直接处理，需要先将其数字化。特征提取的目的是将文本字符串转换为数字特征向量。其中，词袋模型（Bag of word）是常用的模型。

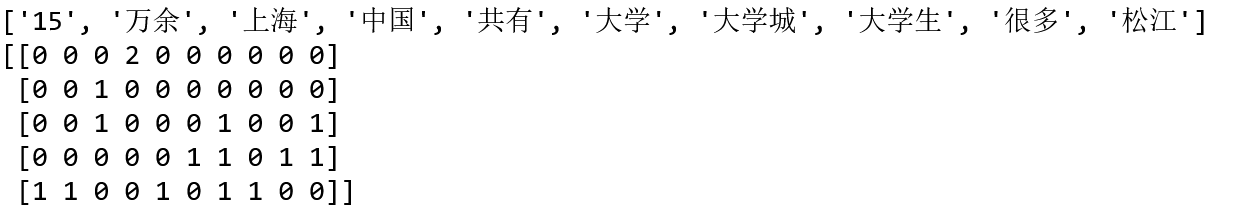
首先，提取文档集的词袋特征：



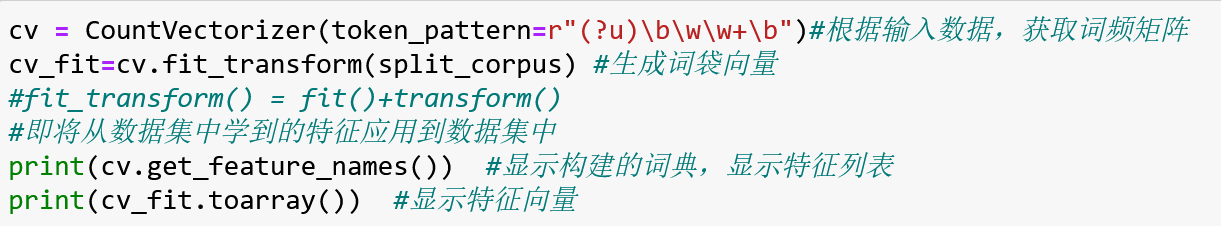


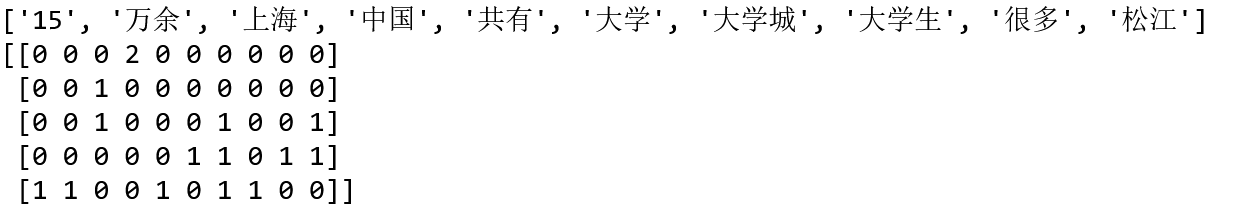
建立词袋模型：分词，构建词典，根据词典构建词袋向量：

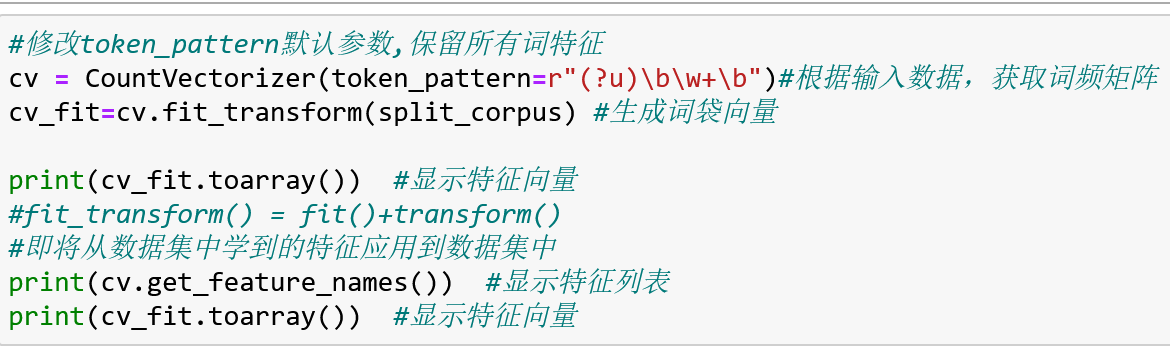


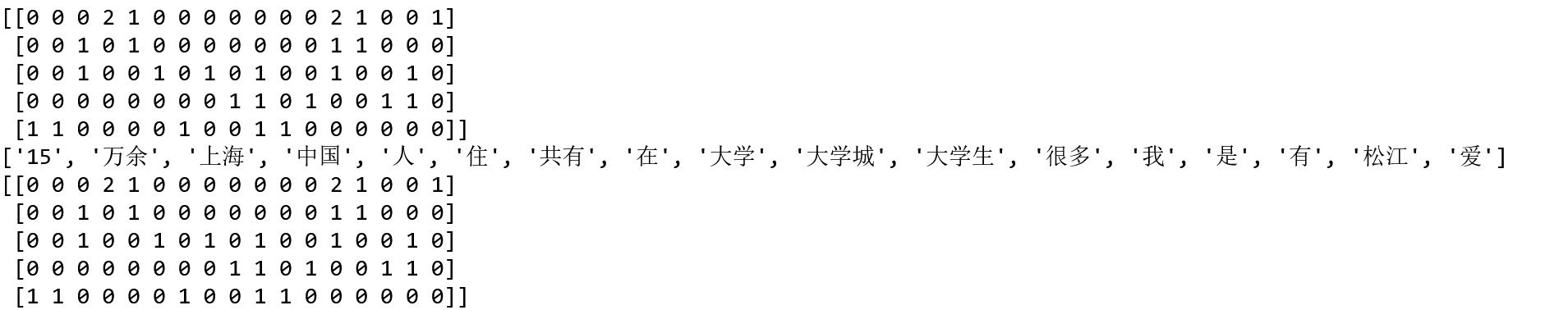


提取以上文档集的词袋模型特征

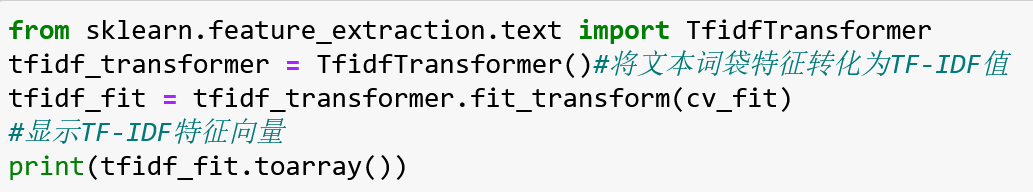


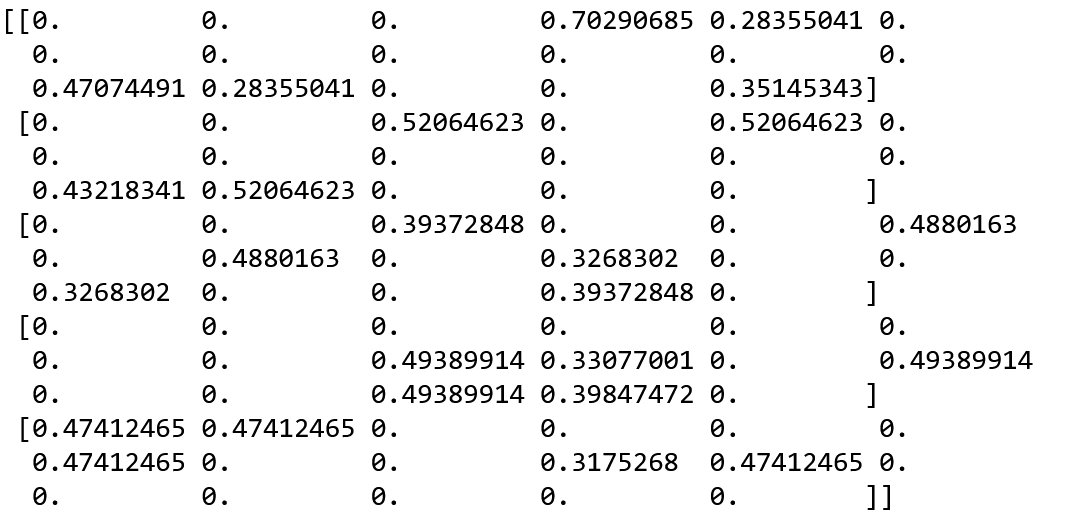






提取以上文档集的TF-IDF模型特征：





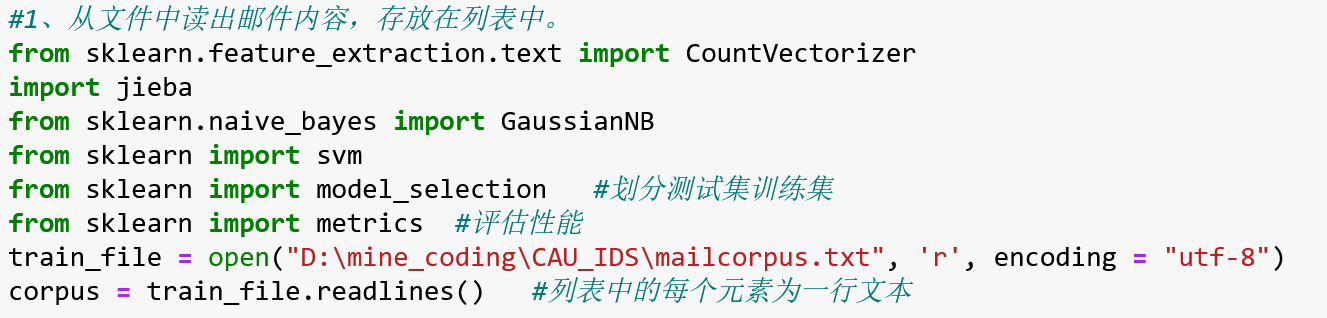
2.用TF-IDF模型提取邮件特征，朴素贝叶斯模型训练分类模型实现垃圾邮件识别并分析模型性能。

**提示：** from sklearn.naive\_bayes import GaussianNB

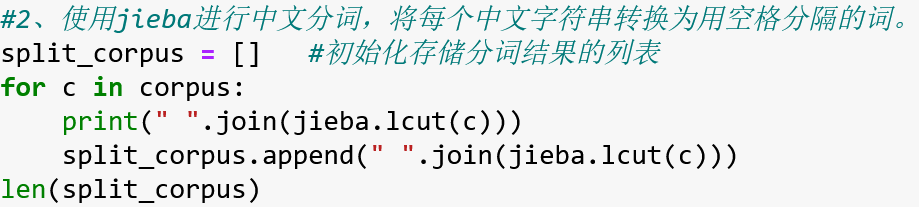
clf = GaussianNB() #模型初始化

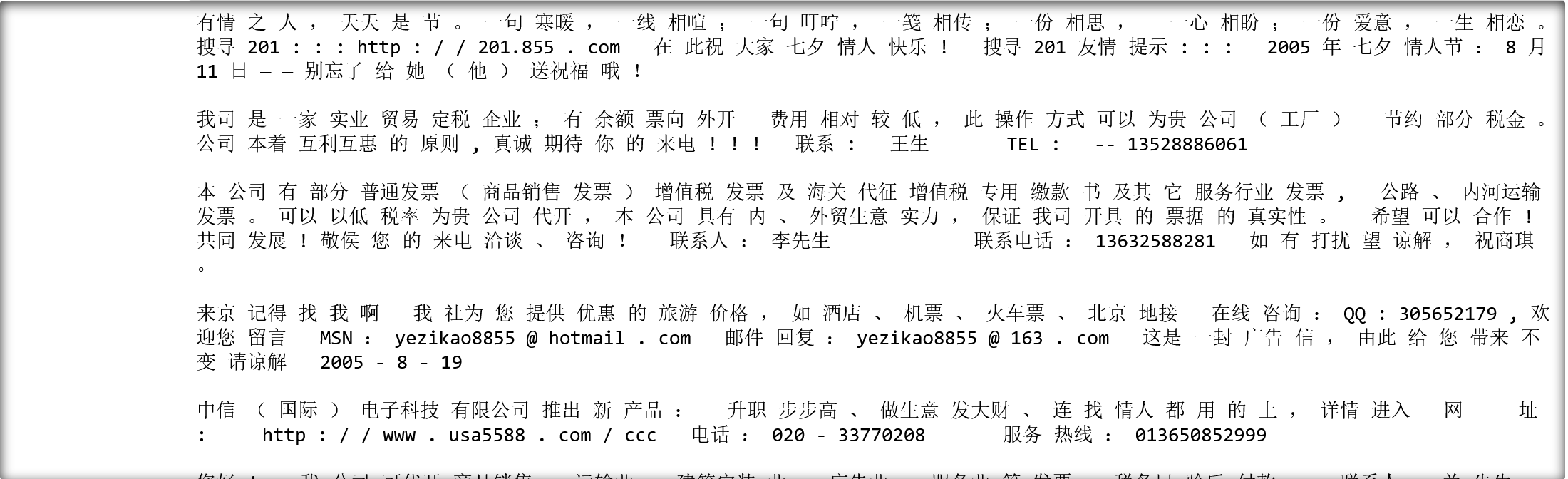
clf.fit(X, y) #模型训练

1、从文件中读出邮件内容，存放在列表中。

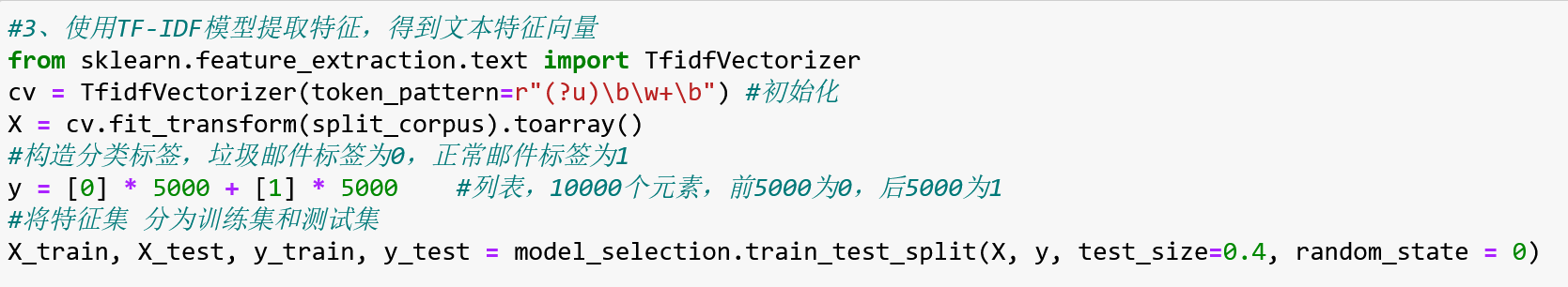


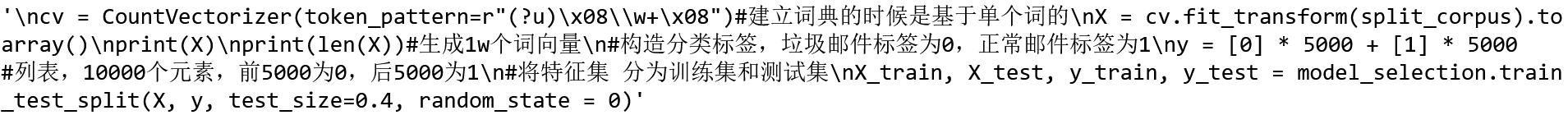
2、使用jieba进行中文分词，将每个中文字符串转换为用空格分隔的词。



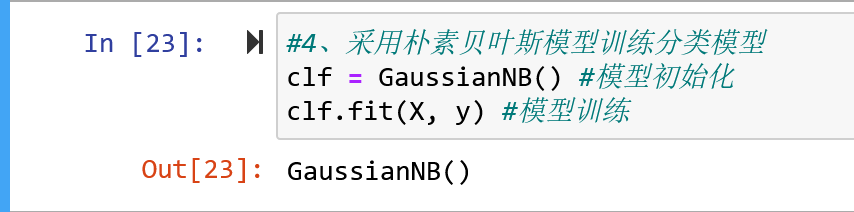


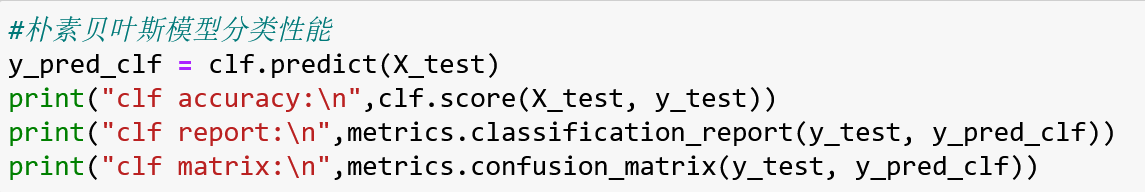
3、使用TF-IDF模型提取特征，得到文本特征向量

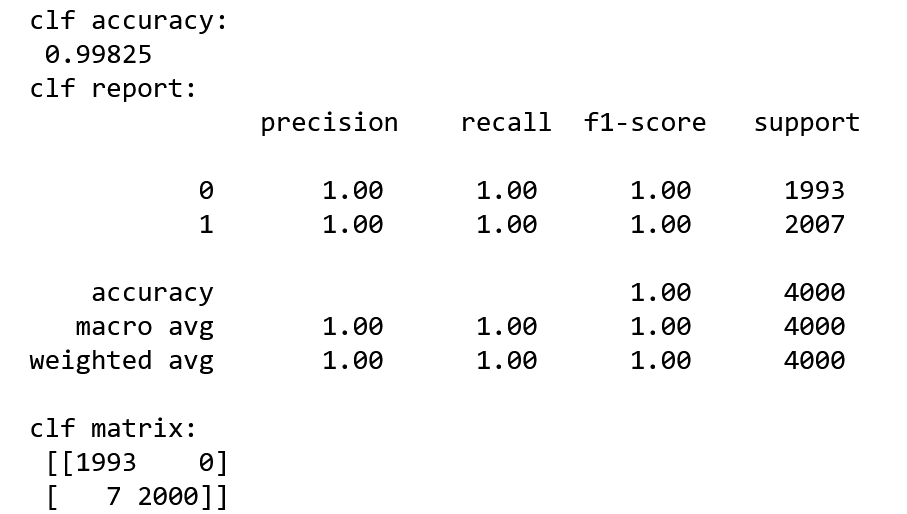




4、采用朴素贝叶斯模型训练分类模型







1. **实验总结**

对于文本数据计算机无法直接处理，需要先将其数字化。特征提取的目的是将文本字符串转换为数字特征向量。其中，词袋模型（Bag of word）和TF-IDF（term frequency–inverse document frequency）是常用的模型。

词袋模型（英语：Bag-of-words model）是一个在自然语言处理和信息检索（IR）下被简化的表达模型。此模型下，一段文本（比如一个句子或是一个文档）可以用一个装着这些词的袋子来表示，这种表示方式不考虑文法以及词的顺序。最近词袋模型也被应用在电脑视觉领域。词袋模型被广泛应用在文件分类，词出现的频率可以用来当作训练分类器的特征。

TF-IDF（term frequency–inverse document frequency，词频-逆向文件频率）是一种用于信息检索（information retrieval）与文本挖掘（text mining）的常用加权技术。TF-IDF是一种统计方法，用以评估一字/词对于一个文件集或一个语料库中的某份文件的重要程度。字词的重要性与它在文件中出现的次数成正比升高，但同时会随着它在语料库中出现的频率成反比下降。TF-IDF的主要思想是：如果某个单词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。