



|  |  |
| --- | --- |
| **题目：** | 文本数据的分类与分析 |
| **学院：** | 计算机学院（国家示范性软件学院） |

小组成员：

|  |  |  |  |
| --- | --- | --- | --- |
| 姓名 | 学号 | 班级 | 分工 |
| 杨义军 | 2020140581 | 2020111302 | 爬虫，降维，朴素贝叶斯和SVM分类器 |
| 蒋玲玲 | 2020140582 | 2020111302 | 爬虫，分词去停用词 |
| 周炜 | 2020140520 | 2020111304 | 爬虫，TF-IDF |

报告日期：2020年11月29日

# 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，基于监督学习的机器学习方法，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。

# 实验要求

* 1. 文本类别>=10类。
  2. 训练集文档数>=500000篇，每类平均50000篇；
  3. 测试集文档数>=500000篇，每类平均50000篇。
  4. 分组实验，组员数量<=3。

# 实验内容

## 3.1文本数据的爬取

我们使用语言是Python，requests爬取内容，bs4解析html文档。数据由两部分组成，第一部分是网上提供的现成数据，第二部分是我们自己爬取的。

### 3.1.1现成数据集

我们使用的数据集是<http://thuctc.thunlp.org/>”网站提供的。每一类的测试集和数据集共10万，共10个类别。由于数据不足我们选取了体育、科技、股票这三类，加上娱乐9万和社会5万的数据。不足的数据我们自己爬取。

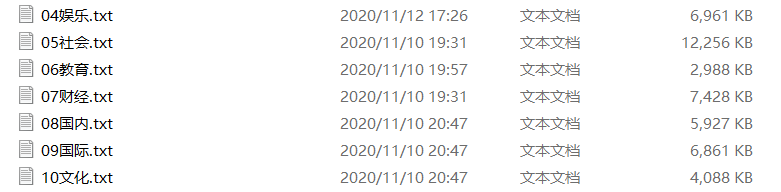
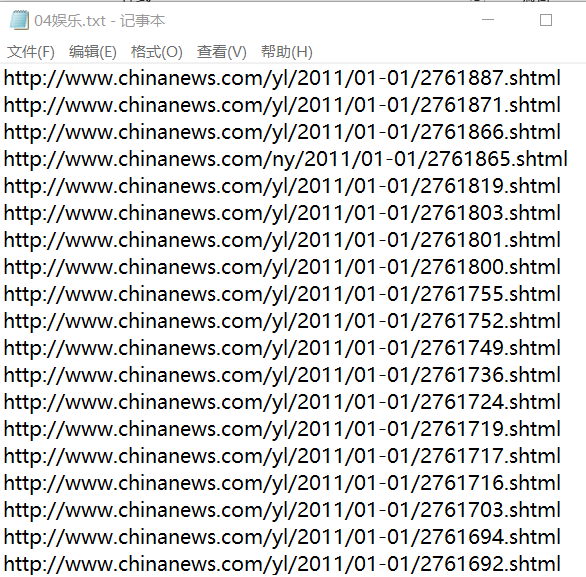
### 3.1.2爬取数据

本次实验数据来源于中国新闻网（[www.chinanews.com](http://www.chinanews.com)），从中选取七个类别的新闻，分别是娱乐1万、社会5万、教育、财经、国内、国际、文化，爬取了这七类新闻从2011年1月至2020年10月的10万文章。

**1. 爬取数据的URL**

网址的模板像这样<http://www.chinanews.com/scroll-news/年份/月份天/news.shtml>

如<http://www.chinanews.com/scroll-news/2011/0101/news.shtml>

里面的内容是当天的新闻的网址，从中选取想要的分类，保存下URL。

图表 1娱乐里面的URL

图表 2每个文件里面保存URL

2.根据URL爬取数据

从URL中爬取数据并且保存。

|  |  |
| --- | --- |
| 娱乐 | 1万 |
| 社会 | 5万 |
| 教育 | 10万 |
| 财经 | 10万 |
| 国内 | 10万 |
| 国际 | 10万 |
| 文化 | 10万 |

## 3.2数据预处理

数据预处理主要包括分词，去停用词，向量化

本项目首先借助python第三方库jieba完成文本的分词处理，然后通过停用词表去除中文停用词，通过正则表达式去除数字，再借助scikit-learn提供的TfidfVectorizer类完成向量化。得到的特征维度太大，为17416维度。使用TruncatedSVD进行降维，降维到2500。

## 3.3朴素贝叶斯算法

朴素贝叶斯原理：

1.条件概率：P( A | B) 表示事件B已经发生的前提下，事件A发生的概率，叫做事件B发生下事件A的条件概率:

2.贝叶斯公式：

3.在本次文本分类实验中，总共有10类文本，我们假设文本类别集合C={C1 , C2 , C3 , C4 , C5 , C6 , C7 , C8 , C9 , C10}，那么文档D属于类别Ci  的概率就可以使用贝叶斯公式计算：

对每一个分类来说，P(Ci) 恒等于1/10 ，P(D)都相等，所以若要比较文本属于C1、C2…C10的大小，只需要计算P(D|Ci) 的大小即可。

4.假设文档D的特征集合X有n个特征：X = {x1 x2 x3 … xn} ，那么P(D|Ci) 的计算公式是：

令，那么，我们就判断文档D属于类别Ck。

首先，我们从TFIDF计算得到的17416维词作为词袋，再从训练集的每个类别中计算每个词的先验概率，这就是训练过程。

当要预测一个文本时，选取词袋中的词相乘，选取10类中结果最大的就为预测的类别。

最后，设计函数格式化输出结果和混淆矩阵。

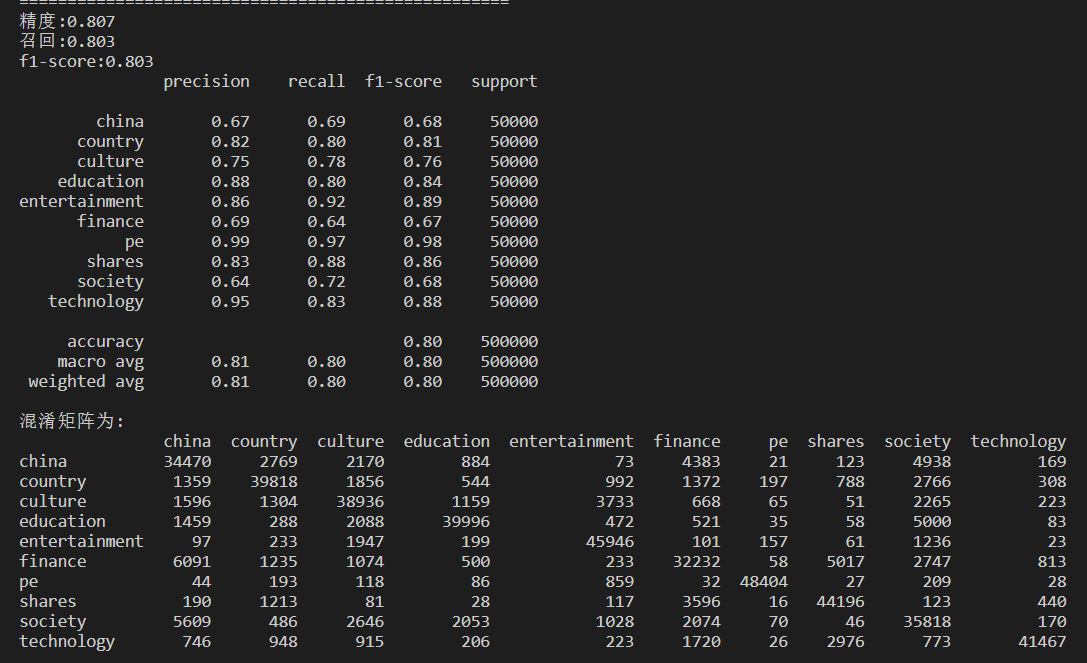
## 3.4 SVM算法

使用LinearSVC作为模型，使用降维得到的2500维特征作为输入进行训练。将训练好的模型保存。

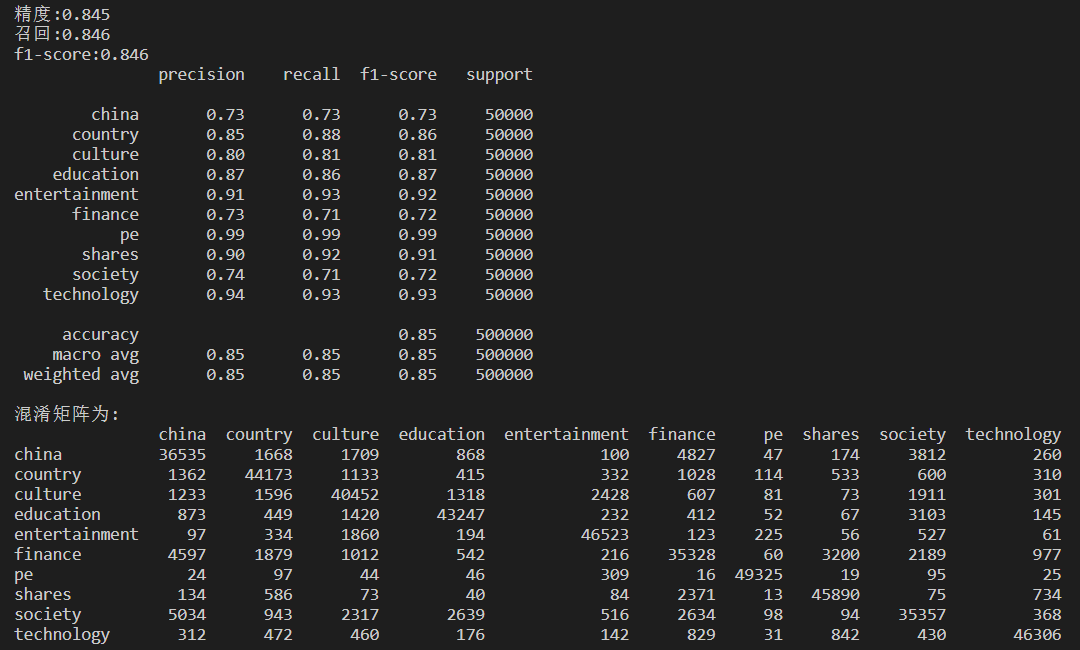
加在模型预测测试集，并且格式化输出结果和混淆矩阵。

# 实验结果

朴素贝叶斯：

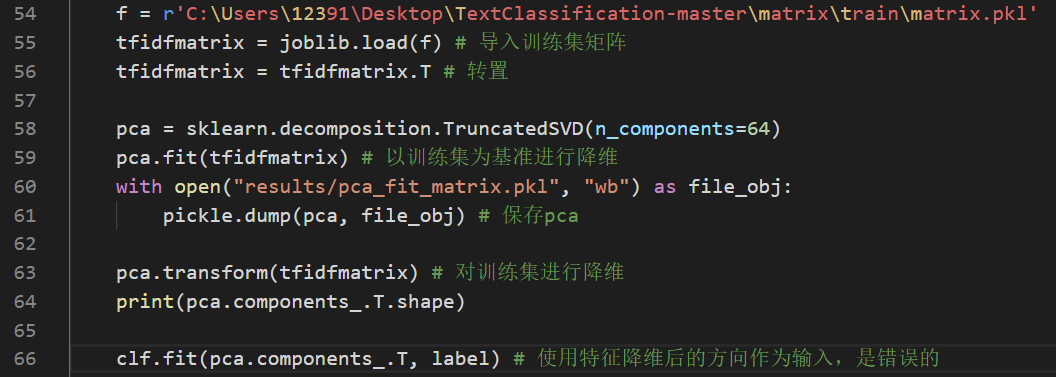


支持向量机：

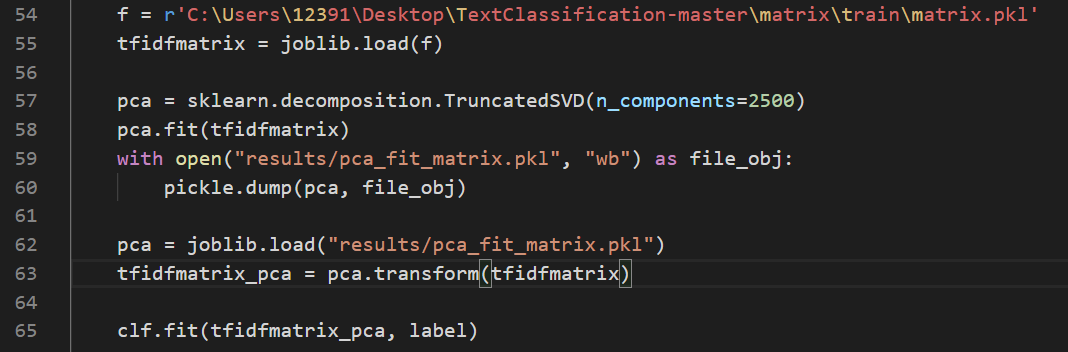


# 实验过程遇到的问题及解决办法

1. 一开始爬取URL时会报错，后来发现是原网站在2012年的时候改变了网站的布局，修改过后正常。
2. 爬取的文本中会存在一些空文本，运行程序处理到这些空文本时程序会出错，这时需要编写一段程序遍历所有文本，遇到空文本就将空文本删除。
3. 在实验验收时老师说我们的降维有问题，我的分析如下：

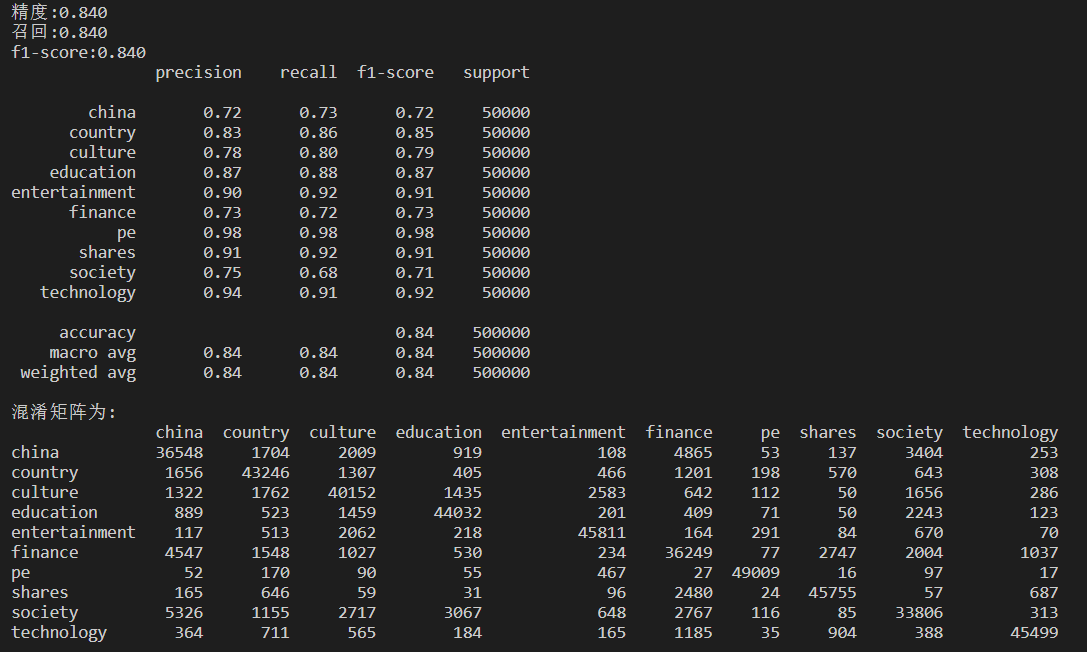
**错误代码如下**：

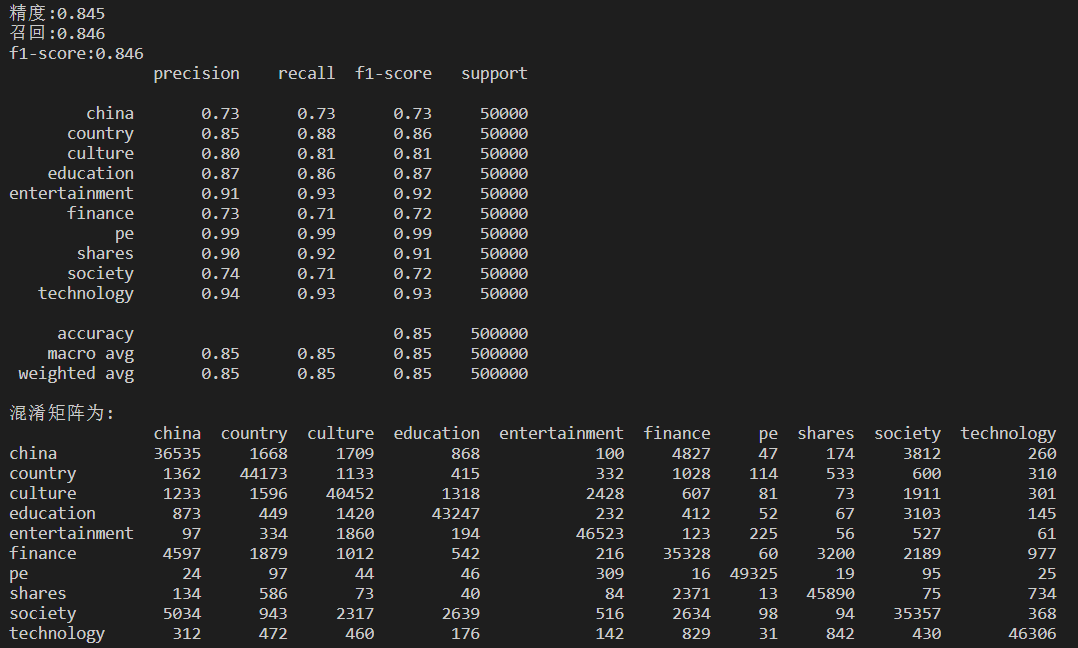
错误原因就是在第63行中，我理解错了transform函数，因为是类所以我先入为主的认为结果保存到了对象中的属性中，后来我看了官方文档，发现结果是transform函数的返回。而第66行分类器的输入则是降维后的每一个成分在原来特征空间中的方向。如第56行转置后的shape为（17416,500000），分类器的输入pca.components\_.T的shape为（500000，64）。第56行转置就是把500000个数据当作了特征，并且降维到64，所以当时我训练时间非常长，因为输入就是错误的。

**修改代码如下**：

删除了56行的转置，57行维度改为2500维，分类器的输入为transform的返回值。训练时长为6937秒（1.9小时）。

**结果对比**：

错误代码

正确代码

正确率并没有提升多少

**分析**：

首先为什么错误代码会有结果，因为分类器的训练用的是pca.components\_.T预测也是这个，所以才会有结果，而transform函数并没有任何作用。

第二，为什么正确代码正确率不是很高。从混淆矩阵中看到chian, finance, society这三类容易分错类别，而entertainment, pe, shares, technology这四类正确率就很高。从现成数据集中得到的的数据分类很好，而自己爬取的数据分类效果并不是很好。

# 实验总结

通过本次实验，我们初步了解了通过机器学习来训练一个分类模型的方法，即构造数据集、通过特征提取得到特征向量、通过机器学习算法训练模型、测试模型的泛化能力。了解了文本特征提取的方法有TF-IDF值方法。了解了朴素贝叶斯和SVM的基本原理。了解了特征向量降维和模型调参的基本方法。

程序地址：<https://github.com/jetyang98/TextClassification>