数据挖掘课程实验最终报告

**王丹 2120151035**

**杨亮 2120151053**

**朱鹏飞 2120151075**

文本分类和主题提取

实验内容简介：

文本分为10个类别，每个类别有200篇文章，每篇文章大概3000字，属于长文本分类，对于每个类别，提取其中的主题内容。

一、预处理阶段

首先需要对文本进行预处理，

1. 去掉不相关的header，footer以及其他注释信息
2. 去除文本分行标志的“\r\n”，合并为一个段落
3. 将处理好的文件放到新的目录下，目录结构和之前的结构相同。

二分词

对上一步进行预处理的文本进行分词，分词后放到新的目录下，目录结构仍然保持和之前的目录结构一致。

三对预处理的文本进行打包

本次步骤主要是实现一个训练用语料数据结构，为做计算tf-idf向量空间模型做准备

1首先定义训练集的数据结构

定义训练集对象：data\_set

使用python的bunch类提供一种key,value的对象形式

Target\_name：所有分类集名称列表

Label:每个语篇定义分类标签列表

Filenames：分词后语篇路径

Contents:分词后语篇内容

2 从分词语料库中将所需信息读入训练集的数据结构中

3 将训练集持久化为一个数据对象文件

4 读出数据对象文件，验证持久化的正确性。

四对打包后的数据计算tf-idf权重，并持久化词包文件

1 导入训练集

2 从文件导入停用词表，并转换为list

3 创建词袋数据结构，并配置停用词表

4 统计每个词语的tf-idf权值。

使用Tdidfvectorizer计算tf-idf权值。

五对测试集进行分类

1 确定测试语料：对测试语料进行预处理

2 对测试语料进行分词

3 导入测试语料：随机选取测试语料类别并记录

4 导入训练词袋模型含vocabulary

5 计算测试语料的tf-idf权值，让两个tfidfvectorizer共享一个vocabulary

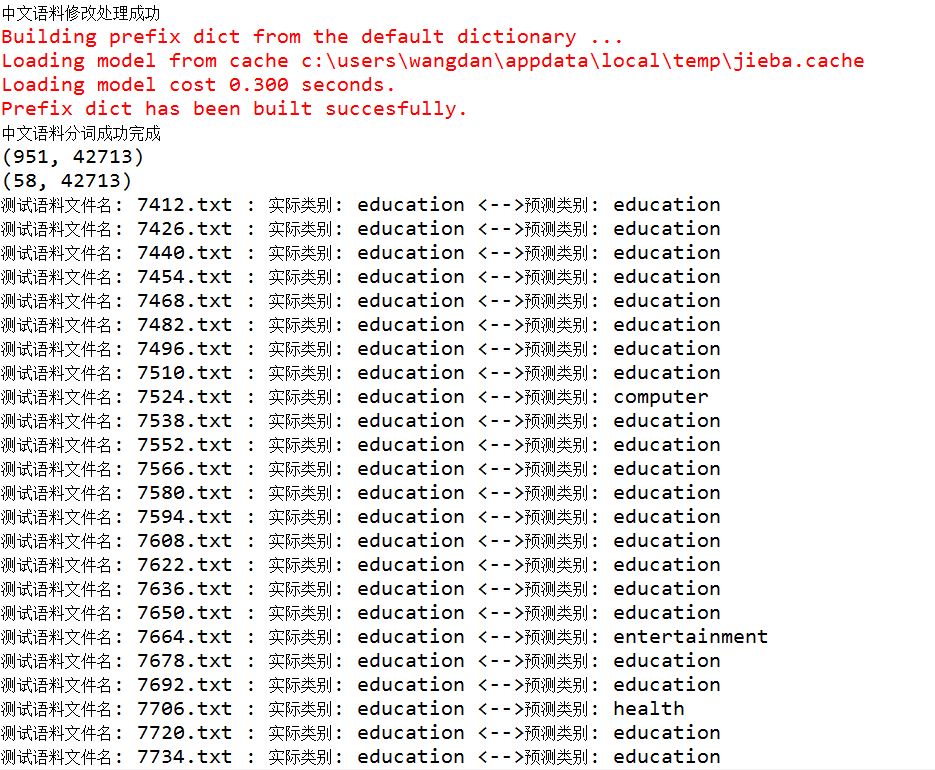
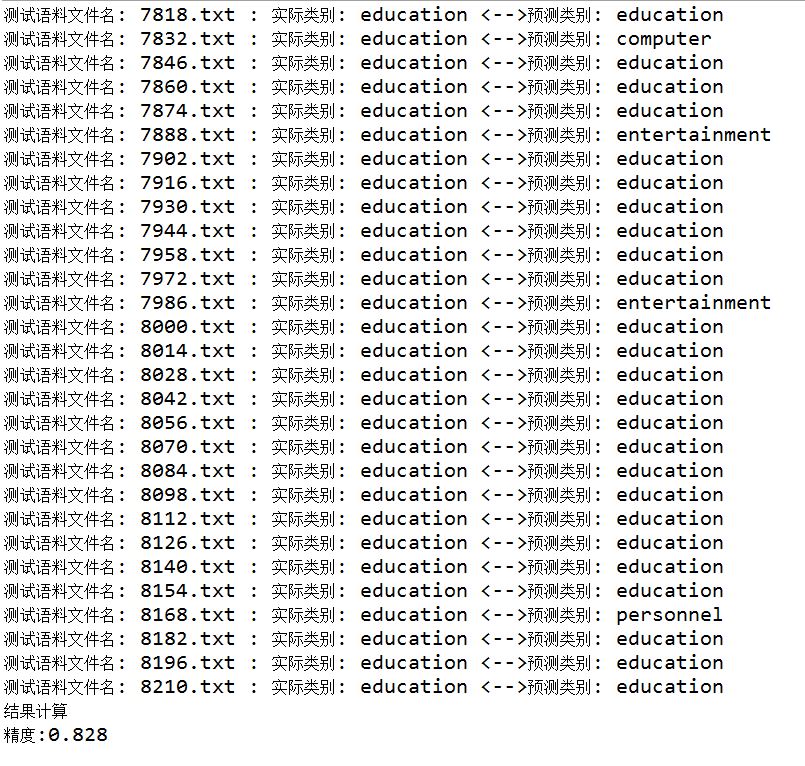
6 应用分类算法

7 预测和输出分类结果

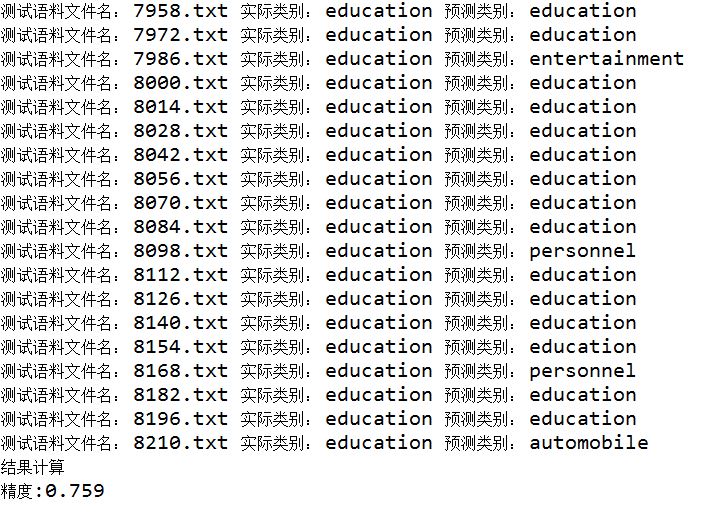
8 计算分类精度

KNN算法分类

KNN算法原理：通过训练好模型，当有新的文章来时，统计它周围k个类别的文章的类型，距离采用的是计算tf-idf矩阵间的距离，由于每篇文章是平等的，由于每类文章的数量基本是一致的，所以不存在有权重大小问题。



朴素贝叶斯分类：



六 主题提取

1 导入语料集

2 从文件导入停用词表

3 从文件导入数据包

4 统计每个类别中tf-idf排名靠前的几个词语，可以大概了解该类文本的主题。



七、实验结论

通过对文本进行分类，学习了NLP和机器学习的有关知识，本次实验分类采用了knn和朴素贝叶斯两种方法，其中knn方法的效果好一些，当然也可以采用包括Kmeans,svm等方法，以后会尝试一下，看看效果怎么样。文本主题的提取方法比较简单，只是提取出了tf-idf靠前的几个单词，结果中发现会有一些没有实际意义的词，并不能直观的从这些词中判断该类别的主题是什么，但是还是有一些具有代表性的词提取出来了，比如sports类别，里面有姚明、nba、曼联，这些词语还是可以很直观的表达出sports这个主题。

对于数据挖掘，在接下来的日子，还是会继续的学习下去，将来也想从事有关数据挖掘的工作，感谢老师的悉心教诲，谢谢！