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。

```
中文语料修改处理成功
Building prefix dict from the default dictionary ...
Loading model from cache c:\users\wangdan\appdata\local\temp\jieba.cache
Loading model cost 0.300 seconds.
Prefix dict has been built succesfully.
中文语料分词成功完成
(951, 42713)
(58, 42713)
测试语料文件名： 7412.txt : 实际类别： education <-->预测类别： education
测试语料文件名： 7426.txt : 实际类别： education <-->预测类别： education
测试语料文件名： 7440.txt : 实际类别： education <-->预测类别： education
测试语料文件名： 7454.txt : 实际类别： education <-->预测类别： education
测试语料文件名： 7468.txt : 实际类别： education <-->预测类别： education
测试语料文件名： 7482.txt : 实际类别： education <-->预测类别： education
测试语料文件名： 7496.txt : 实际类别： education <-->预测类别： education
测试语料文件名： 7510.txt : 实际类别： education <-->预测类别： education
测试语料文件名： 7524.txt : 实际类别： education <-->预测类别： computer
测试语料文件名： 7538.txt : 实际类别： education <-->预测类别： education
测试语料文件名： 7552.txt : 实际类别： education <-->预测类别： education
测试语料文件名： 7566.txt : 实际类别： education <-->预测类别： education
测试语料文件名： 7580.txt : 实际类别： education <-->预测类别： education
测试语料文件名： 7594.txt : 实际类别： education <-->预测类别： education
测试语料文件名： 7608.txt : 实际类别： education <-->预测类别： education
测试语料文件名： 7622.txt : 实际类别： education <-->预测类别： education
测试语料文件名： 7636.txt : 实际类别： education <-->预测类别： education
测试语料文件名： 7650.txt : 实际类别： education <-->预测类别： education
测试语料文件名： 7664.txt : 实际类别： education <-->预测类别： entertainment
测试语料文件名： 7678.txt : 实际类别： education <-->预测类别： education
测试语料文件名： 7692.txt : 实际类别： education <-->预测类别： education
测试语料文件名： 7706.txt : 实际类别： education <-->预测类别： health
测试语料文件名： 7720.txt : 实际类别： education <-->预测类别： education
测试语料文件名： 7734.txt : 实际类别： education <-->预测类别： education
测试语料文件名： 7818.txt : 实际类别： education <-->预测类别： education
测试语料文件名： 7832.txt : 实际类别： education <-->预测类别： computer
测试语料文件名： 7846.txt : 实际类别： education <-->预测类别： education
测试语料文件名： 7860.txt : 实际类别： education <-->预测类别： education
测试语料文件名： 7874.txt : 实际类别： education <-->预测类别： education
测试语料文件名： 7888.txt : 实际类别： education <-->预测类别： entertainment
测试语料文件名： 7902.txt : 实际类别： education <-->预测类别： education
测试语料文件名： 7916.txt : 实际类别： education <-->预测类别： education
测试语料文件名： 7930.txt : 实际类别： education <-->预测类别： education
测试语料文件名： 7944.txt : 实际类别： education <-->预测类别： education
测试语料文件名： 7958.txt : 实际类别： education <-->预测类别： education
测试语料文件名： 7972.txt : 实际类别： education <-->预测类别： education
测试语料文件名： 7986.txt : 实际类别： education <-->预测类别： entertainment
测试语料文件名： 8000.txt : 实际类别： education <-->预测类别： education
测试语料文件名： 8014.txt : 实际类别： education <-->预测类别： education
测试语料文件名： 8028.txt : 实际类别： education <-->预测类别： education
测试语料文件名： 8042.txt : 实际类别： education <-->预测类别： education
测试语料文件名： 8056.txt : 实际类别： education <-->预测类别： education
测试语料文件名： 8070.txt : 实际类别： education <-->预测类别： education
测试语料文件名： 8084.txt : 实际类别： education <-->预测类别： education
测试语料文件名： 8098.txt : 实际类别： education <-->预测类别： education
测试语料文件名： 8112.txt : 实际类别： education <-->预测类别： education
测试语料文件名： 8126.txt : 实际类别： education <-->预测类别： education
测试语料文件名： 8140.txt : 实际类别： education <-->预测类别： education
测试语料文件名： 8154.txt : 实际类别： education <-->预测类别： education
测试语料文件名： 8168.txt : 实际类别： education <-->预测类别： personnel
测试语料文件名： 8182.txt : 实际类别： education <-->预测类别： education
测试语料文件名： 8196.txt : 实际类别： education <-->预测类别： education
测试语料文件名： 8210.txt : 实际类别： education <-->预测类别： education
结果计算
精度：0.828
```

朴素贝叶斯分类：

测试语料文件名： 7958.txt 实际类别： education 预测类别： education
测试语料文件名： 7972.txt 实际类别： education 预测类别： education
测试语料文件名： 7986.txt 实际类别： education 预测类别： entertainment
测试语料文件名： 8000.txt 实际类别： education 预测类别： education
测试语料文件名： 8014.txt 实际类别： education 预测类别： education
测试语料文件名： 8028.txt 实际类别： education 预测类别： education
测试语料文件名： 8042.txt 实际类别： education 预测类别： education
测试语料文件名： 8056.txt 实际类别： education 预测类别： education
测试语料文件名： 8070.txt 实际类别： education 预测类别： education
测试语料文件名： 8084.txt 实际类别： education 预测类别： education
测试语料文件名： 8098.txt 实际类别： education 预测类别： personnel
测试语料文件名： 8112.txt 实际类别： education 预测类别： education
测试语料文件名： 8126.txt 实际类别： education 预测类别： education
测试语料文件名： 8140.txt 实际类别： education 预测类别： education
测试语料文件名： 8154.txt 实际类别： education 预测类别： education
测试语料文件名： 8168.txt 实际类别： education 预测类别： personnel
测试语料文件名： 8182.txt 实际类别： education 预测类别： education
测试语料文件名： 8196.txt 实际类别： education 预测类别： education
测试语料文件名： 8210.txt 实际类别： education 预测类别： automobile
结果计算
精度:0.759

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。

```
共10 种类别['automobile', 'computer', 'education', 'entertainment', 'estate', 'finance', 'health', 'personnel', 'sports', 'technolog
topic automobile
汽车 投标 拥车证 新车 车子 价格 人数 新加坡 购买 成价 如果 万元 认为 本地 可以 驾驶 公司 跑车 下跌 人们
topic computer
病毒 手机 公司 微软 电脑 用户 软件 黑客 中国 游戏 市场 网络 服务 可以 3g 技术 推出 视窗 使用 系统
topic education
考试 学生 考生 记者 专业 毕业生 高考 招生 人才 一个 企业 北京 公司 工作 学校 录取 今年 一些 增加 自己
topic entertainment
电影 一个 观众 我们 音乐 演出 他们 导演 自己 专辑 没有 这个 中国 就是 这部 记者 觉得 电视剧 美国 因为
topic estate
单位 项目 公寓 平方英尺 组屋 推出 售价 地段 价格 市场 万元 房地产 发展商 地契 今年 平均 位于 私人 这个 洋房
topic finance
汇价 公司 银行 市场 增长 经济 股市 央行 利率 记者 产品 中国 日本 支撑 显示 贷款 震荡 基金 问题 美元
topic health
病人 研究 药物 医生 运动 健康 细胞 治疗 糖尿病 手术 问题 可以 肥胖 发现 男性 一种 癌症 显示 关节 可能
topic personnel
员工 企业 工作 公司 他们 简历 管理 一个 人才 自己 本网 部门 人力资源部 我们 培训 问题 管理者 hr 需要 招聘
topic sports
比赛 姚明 球员 nba 球队 我们 分钟 曼联 他们 取得 切尔西 没有 表现 中国 奥尼尔 最后 联赛 球迷 利物浦 湖人队
```

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！