**深 圳 大 学 实 验 报 告**

**课程名称： 信息检索**

**实验项目名称：分类和聚类的实验**

**学院： 计算机与软件学院**

**专业： 计算机科学与技术**

**指导教师：潘微科**

**报告人： 伍耀梁 学号： 2015150166 班级： 5**

**实验时间：2019年5月10日（周五）-2019年5月18日（周六）**

**实验报告提交时间： 2019年5月17日**

**教务部制**

|  |
| --- |
| **实验目的与要求：**  **实验目的：**掌握基本的分类（classification）和聚类（clustering）算法。  **实验要求：**  (1) 用Java语言（或其他常用语言）实现一个基于朴素贝叶斯分类算法（Naive Bayes algorithm）的文档分类简易系统（判断某个公文通的通知是不是关于本科生学科竞赛、创新创业项目等方面的信息，是一个二分类的任务）。  实现朴素贝叶斯分类算法（Naive Bayes algorithm）40分；  支持13.5.1节中的基于互信息（Mutual Information）的特征选择15分；  支持13.5.2节中的基于X^2的特征选择15分；  支持中英文混合的文档10分（一般公文通的通知中都有中文和英文）。  请在报告中附上系统整体设计、代码截图、运行结果截图、简要文字说明和对结果的详细分析。程序要有详细注释。使用特征选择和不使用特征选择的分类效果要进行对比和分析。  注：请自行从公文通中获取一些文档，训练集至少有200篇文档（.txt格式，相关的文档和不相关的文档比例大约为1：9），测试集至少有100篇文档（.txt格式，相关的文档和不相关的文档比例大约为1：9）。  报告写作。要求：主要思路有明确的说明，重点代码有详细的注释，行文逻辑清晰可读性强，报告整体写作较为专业。（20分）  **说明：**  （1）本次实验课作业满分为100分。  （2）本次实验课作业截至时间2019年5月18日（周六）23:59。  （3）报告正文：请在指定位置填写，本次实验需要单独提交源程序文件和数据集（打包在Blackboard中上传）。  （4）个人信息：WORD文件名中的“姓名”、“学号”，请改为你的姓名和学号；实验报告的首页，请准确填写“学院”、“专业”、“报告人”、“学号”、“班级”、“实验报告提交时间”等信息。  （5）提交方式：截至时间前，请在Blackboard平台中提交。  （6）发现抄袭（包括复制&粘贴整句话、整张图），**抄袭者和被抄袭者的所有平时作业成绩记零分（含抄袭往届同学的作业），即这门课不及格。**  （7）延迟提交，不得分；如有特殊情况，请于截至日期之后48小时内发邮件到panweike@szu.edu.cn，并在邮件中注明课程名称、作业名称、姓名、学号等信息，以及特殊情况说明，我收到后会及时回复。  （8）期末考试阶段补交无效。 |

|  |
| --- |
| **(1) 用Java语言（或其他常用语言）实现一个基于朴素贝叶斯分类算法（Naive Bayes algorithm）的文档分类简易系统（判断某个公文通的通知是不是关于****本科生学科竞赛、创新创业项目等方面的信息，是一个二分类的任务）。**  **实现朴素贝叶斯分类算法（Naive Bayes algorithm）40分；**  **支持13.5.1节中的基于互信息（Mutual Information）的特征选择15分；**  **支持13.5.2节中的基于X^2的特征选择15分；**  **支持中英文混合的文档10分（一般公文通的通知中都有中文和英文）。**  **请在报告中附上系统整体设计、代码截图、运行结果截图、简要文字说明和对结果的详细分析。程序要有详细注释。使用特征选择和不使用特征选择的分类效果要进行对比和分析。**  **注：请自行从公文通中获取一些文档，训练集至少有200篇文档（.txt格式，相关的文档和不相关的文档比例大约为1：9），测试集至少有100篇文档（.txt格式，相关的文档和不相关的文档比例大约为1：9）。**  1.首先是数据集的获取，数据集的获取是通过爬虫获得的。爬虫的相关代码如下：    python的爬虫，基于requests库来发送http请求，然后定义好请求头部的cookie（用于校内公文通的学生身份验证）后，发送一个近30天公文通的POST请求，获得近30天内公文通的列表（30天的公文通个数为1724个结果），取得每一个公文通的链接。  2.朴素贝叶斯分类算法  取得公文通链接的列表之后，对其中的内容进行过滤，把标题和正文部分写入到文件中，下图为其中一篇公文通的内容。  然后对数据集合进行预处理，也就是给符合条件的文章打上label。最后分类成训练集和测试集。    训练集中一共有19篇公文通是有关本科生学科竞赛、创新创业项目等方面的信息，有180篇是无关的，也就是一共有200篇公文通，其中有关和无关的比例大约是1：9，符合报告中对于训练集的要求。    然后用训练集来进行训练，并把训练出来的结果（单词表V，先验概率prior，词汇的条件概率condprob）写入到model中，作为一个预测文章类别的模型。  训练完毕后用测试集对数据进行测试，测试集的大小为100,其中10篇有关，90篇无关。测试部分先对model进行读取，然后计算分数，由于对这个分数取了对数，所以当分数大于1的时候，文章是有关的，反之，小于1的时候是无关的，所以当测试集中，有关的文章输出结果小于1则不正确，反之，无关文章输出结果大于1则正确，由于训练集合中，有关文章的集合大小有限，所以输出结果中，有关文章的准确率判断比较差，可以在输出结果中看到，误判的文档都是相关文章，而无关文章都是符合训练的模型的，最后得到整个模型的准确率为97.45%。    测试部分的相关代码如下：    3.基于互信息的特征选择  计算出对应token的N00，N01，N11，N10  ++++++++++++++++++++++++++++++++++++++++++++++++++++++  **其他（例如感想、建议等等）。** |

深圳大学学生实验报告用纸

|  |
| --- |
| 指导教师批阅意见：  成绩评定：  指导教师签字：  2019年 月 日 |
| 备注： |

注：1、报告内的项目或内容设置，可根据实际情况加以调整和补充。

2、教师批改学生实验报告时间应在学生提交实验报告时间后10日内。