**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： 沈晨玙 学号：2019092121 班级：19计科04**

**实验时间：2022年5月13日（周五）-2022年5月25日（周三）**

**实验报告提交时间： 2022年5月22日星期日**

**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教材《Introduction to Information Retrieval》第13章中介绍的两种特征选择方法：13.5.1节中描述的基于互信息（Mutual Information）的特征选择方法和13.5.2节中描述的基于X^2的特征选择方法。请自行从学校公文通获取2021年的新闻文档（爬取或手动下载），要求包括以下150篇新闻文档：  “党政办公室”发布的最新的30篇新闻文档，  “教务部”发布的最新的30篇新闻文档，  “招生办公室”发布的最新的30篇新闻文档，  “研究生院”发布的最新的30篇新闻文档，  “科学技术部”发布的最新的30篇新闻文档。  将“党政办公室”、“教务部”、“招生办公室”、“研究生院”和“科学技术部”作为5个class，并通过互信息和X^2为每个class选出最相关的15个特征（包含特征名称和相应的值，小数点后保留2位），并对结果作简要分析。  此外，关于以上两种特征选择方法，谈谈你的理解，并通过类比、关联或演绎的方式，举一个在日常的学习生活中可以应用的例子（要求积极向上且能自圆其说）。  请在报告中附上代码截图（不要复制源代码，请用截图的方式）、运行结果截图和详细的文字说明。程序要有详细的注释。对使用的中文分词工具作简要介绍。（60分）  (2) 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“党政办公室”、“教务部”、“招生办公室”、“研究生院”和“科学技术部”方面的信息，即从5个类别中选择最相关的一个）。  要对使用特征选择和不使用特征选择的分类效果进行对比和分析。用题(1)中的文档进行训练和测试，每个类别中的20篇用作训练，10篇用作测试。  此外，关于朴素贝叶斯分类算法，谈谈你的理解，并通过类比、关联或演绎的方式，举一个在日常的学习生活中可以应用的例子（要求积极向上且能自圆其说）  请在报告中附上系统整体设计、代码截图（不要复制源代码，请用截图的方式）、运行结果截图和详细的文字说明。程序要有详细注释。对使用的中文分词工具作简要介绍。（20分）  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2022年5月25日（周三）22:00。  （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（源程序单独打包在Blackboard中上传，不要包含外部导入的包）。  （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成绩记零分。**  （7）延迟提交，不得分；如有特殊情况，请于截至日期之后的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，以及特殊情况的说明，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教材《Introduction to Information Retrieval》第13章中介绍的两种特征选择方法：13.5.1节中描述的基于互信息（Mutual Information）的特征选择方法和13.5.2节中描述的基于X^2的特征选择方法。请自行从学校公文通获取2021年的新闻文档（爬取或手动下载），要求包括以下150篇新闻文档：  “党政办公室”发布的最新的30篇新闻文档，  “教务部”发布的最新的30篇新闻文档，  “招生办公室”发布的最新的30篇新闻文档，  “研究生院”发布的最新的30篇新闻文档，  “科学技术部”发布的最新的30篇新闻文档。  将“党政办公室”、“教务部”、“招生办公室”、“研究生院”和“科学技术部”作为5个class，并通过互信息和X^2为每个class选出最相关的15个特征（包含特征名称和相应的值，小数点后保留2位），并对结果作简要分析。  请在报告中附上代码截图（不要复制源代码，请用截图的方式）、运行结果截图和详细的文字说明。程序要有详细的注释。对使用的中文分词工具作简要介绍。（60分）  +++++++++++++++++++++++++++++++++++++++++++++++++++++++  **代码截图、运行结果截图和详细的文字说明：**  第一步：爬虫  使用python selenium自动化工具从公文通上进行文本的爬取，并且提取正文整理成文件存储。具体实现流程大致为使用chrome自动化工具自动切换至2021年各个部门的界面，然后爬虫获取前30篇新闻的链接并使用定位到相应div进行文本读取。具体实现如下：  自动化页面切换：  获取文本信息：  第二步就是进行文本信息处理   1. 读取文档，根据文件名进行分类   其中file[0-5]依次为五个类别的文章   1. 使用jieba进行分词处理，并生成词袋。   分词结果：  词袋：  注：由于爬虫下来的文章会有\u3000等多余字符，所以需要进行额外处理  第三步是进行特征选择：  MI计算公式如下所示：  所以为了减少重复计算工作，我首先对于每个词项在不同类别中的出现情况进行了统计。对于每个单词，我首先统计成了以下表格形式：   |  |  |  | | --- | --- | --- | | Word / Category | 出现次数 | 未出现次数 | | Category 1 | 1 | 29 | | Category 2 | 3 | 27 | | Category 3 | 5 | 25 | | Category 4 | 7 | 23 | | Category 5 | 9 | 21 |   有了以上表格，就可以快速的计算出每个类别下N11,N10,N01,N00的四个值了，并且根据MI，X^2的计算公式得到相应结果  最终计算结果如下所示：  第四步就是进行排序了。有序题目只需要获得前15大的特征值，因此此处选择了基于小顶堆的TopK算法：  首先编写一个小顶堆重建算法：  然后是TopK排序算法：首先用前K个元素建立小顶堆，如果后续元素大于堆顶元素，则替换，并重建小顶堆。最后使用堆排序算法，对K个元素进行排序。  最终得到如下结果：  +++++++++++++++++++++++++++++++++++++++++++++++++++++++  **对使用的中文分词工具作简要介绍：**  调用jieba.cut函数进行分词  jieba分词0.4版本以上支持四种分词模式：   1. 精确模式：试图将句子最精确地切开，适合文本分析； 2. 全模式：把句子中所有的可以成词的词语都扫描出来, 速度非常快，但是不能解决歧义 3. 搜索引擎模式：在精确模式的基础上，对长词再次切分，提高召回率，适合用于搜索引擎分词 4. paddle模式：利用PaddlePaddle深度学习框架，训练序列标注（双向GRU）网络模型实现分词。同时支持词性标注。   精确模式：  全模式：  搜索引擎模式：  paddle模式：（m1 macbook暂不支持paddlepaddle安装，故没有测试）  可以看出，全模式的切词是最粗糙的，将所有的词汇都进行了返回。主要有以下几个问题：   1. 未结合语境，容易有歧义：协同过滤----> 协同+同过+过滤 2. 不了解词汇：鲁棒----> 鲁+棒   精确模式与搜索引擎模式可以结合具体需求进行选择。  +++++++++++++++++++++++++++++++++++++++++++++++++++++++  **通过互信息为每个class选出的最相关的15个特征：**   |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | |  | “党政办公室” | “教务部” | “招生办公室” | “研究生院” | “科学技术部” | | 特征1和值 | 党政0.62 | 教务部0.53 | 中学0.48 | 研究生院0.65 | 科学技术部0.67 | | 特征2和值 | 办公室0.24 | 教学0.17 | 招生0.44 | 研究生0.55 | 项目0.32 | | 特征3和值 | 定于0.16 | 课程0.12 | 走进0.38 | 硕士0.21 | 科研0.21 | | 特征4和值 | 全校0.13 | 办公室0.10 | 名师0.32 | 导师0.19 | 国家自然科学基金0.20 | | 特征5和值 | 党委0.12 | 党政0.08 | 进0.31 | 学位0.15 | 领域0.19 | | 特征6和值 | 培养0.10 | 也0.08 | 介绍0.29 | 培养0.12 | 李洁0.18 | | 特征7和值 | 总支0.09 | 门0.08 | 随后0.27 | 博士0.12 | 研究0.17 | | 特征8和值 | 基层0.09 | 招生0.08 | 主任0.25 | 邹娟0.09 | 科技部0.17 | | 特征9和值 | 筹0.08 | 研究生院0.07 | 圆满结束0.25 | 专业学位0.09 | 学生0.15 | | 特征10和值 | 南山区0.08 | 广东省教育厅0.07 | 通知0.25 | 项目0.08 | 国际0.13 | | 特征11和值 | 保密工作0.08 | 科学技术部0.07 | 生源0.24 | 所有0.08 | 科技0.13 | | 特征12和值 | 博物馆0.08 | 四六级0.06 | 伊始0.24 | 硕士学位0.08 | 自然科学0.13 | | 特征13和值 | 一名0.08 | 思政0.06 | 办公室0.24 | 岗位职责0.08 | 日0.13 | | 特征14和值 | 申请0.07 | 教高0.06 | 高考0.23 | 科学技术部0.07 | 办公室0.12 | | 特征15和值 | 研究0.07 | 上网0.06 | 们0.22 | 创新0.07 | 经费0.12 |   +++++++++++++++++++++++++++++++++++++++++++++++++++++++  **对通过互信息为每个class选出的最相关的15个特征作简要分析：**   1. 党政办公室：筛选出的信息比较符合，党政、党委、基层等词项很符合。 2. 教务部：筛选的文章大多集中于12月份，当时大部分文章都在于总结年度工作。经过对比，可以大致概括出当月的主要工作内容。   3、招生办公室：比较符合。可以很明显看出当时发布了许多走进高中宣传深大的咨询。经验证，发现确实发布了大量主题为《名师进中学》的系列文章。  4、研究生院：筛选出的信息比较符合，硕士、导师、博士等词项很符合。  5、科学技术部：筛选出的信息比较符合，国家自然科学基金、自然科学、经费等词项很符合。  +++++++++++++++++++++++++++++++++++++++++++++++++++++++  **通过X^2为每个class选出的最相关的15个特征：**   |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | |  | “党政办公室” | “教务部” | “招生办公室” | “研究生院” | “科学技术部” | | 特征1和值 | 党政122.79 | 教务部102.44 | 中学104.46 | 研究生院127.61 | 科学技术部132.72 | | 特征2和值 | 定于37.05 | 教学38.28 | 走进83.72 | 研究生112.52 | 项目69.80 | | 特征3和值 | 办公室35.99 | 课程27.28 | 招生81.37 | 硕士50.03 | 科研50.72 | | 特征4和值 | 全校29.14 | 办公室18.17 | 名师73.58 | 导师45.55 | 领域46.27 | | 特征5和值 | 党委28.76 | 广东省教育厅18.01 | 进68.996 | 学位37.65 | 国家自然科学基金45.88 | | 特征6和值 | 基层20.42 | 门16.433 | 介绍66.13 | 博士29.44 | 李洁41.47 | | 特征7和值 | 总支20.42 | 网上14.54 | 随后64.64 | 培养27.11 | 科技部41.34 | | 特征8和值 | 本页16.52 | 大厅14.54 | 主任60.73 | 专业学位20.42 | 研究40.53 | | 特征9和值 | 入口16.52 | 学期13.10 | 圆满结束59.22 | 邹娟20.42 | 国际33.49 | | 特征10和值 | 筹16.43 | 课12.80 | 生源54.92 | 所有19.05 | 科技31.67 | | 特征11和值 | 一名16.43 | 教学改革12.80 | 伊始54.92 | 岗位职责16.43 | 经费29.44 | | 特征12和值 | 博物馆16.43 | 教材12.8 | 高考54.54 | 硕士学位16.43 | 资助28.76 | | 特征13和值 | 南山区16.43 | 任课老师12.80 | 们52.30 | 是否16.24 | 指南28.76 | | 特征14和值 | 保密工作16.43 | 上网12.53 | 接着50.19 | 名单15.35 | 自然科学28.63 | | 特征15和值 | 请假14.50 | 四六级12.53 | 大学49.37 | 现将13.50 | 依托28.42 |   +++++++++++++++++++++++++++++++++++++++++++++++++++++++  **对通过X^2为每个class选出的最相关的15个特征作简要分析：**   1. 党政办公室：筛选出的信息比较符合，党政、党委、基层等词项很符合。 2. 教务部：筛选的文章大多集中于12月份，当时大部分文章都在于总结年度工作。经过对比，可以大致概括出当月的主要工作内容。   3、招生办公室：比较符合。可以很明显看出当时发布了许多走进高中宣传深大的咨询。经验证，发现确实发布了大量主题为《名师进中学》的系列文章。  4、研究生院：筛选出的信息比较符合，硕士、导师、博士等词项很符合。  5、科学技术部：筛选出的信息比较符合，国家自然科学基金、自然科学、经费等词项很符合。  +++++++++++++++++++++++++++++++++++++++++++++++++++++++  **对通过互信息和X^2为每个class选出的最相关的15个特征作简要的对比分析：**  因为爬虫的关系，在所有文章中都会出现类似（本文最近更新于2021/12/29 19:05:00　累计点击数:877）的语句。但是两种算法都可以很好的过滤掉这种在所有类别中都重复出现的信息，原因是这种语句中term的N11和N10都很高，可以较好的进行过滤。  另外两种计算方法，前几个特征的选择及排序都相对一致。后面几个特征会有不同的侧重，这是因为X^2基于显著统计性进行选择，因此他会比MI选择出更多的罕见项，而这些词项对于分类是不太可靠的。当然，MI也不一定就能选出是的分类精度最大化的词项。因此我认为加大样本量才是更好的方法。  +++++++++++++++++++++++++++++++++++++++++++++++++++++++  **此外，关于以上两种特征选择方法，谈谈你的理解，并通过类比、关联或演绎的方式，举一个在日常的学习生活中可以应用的例子（要求积极向上且能自圆其说）。**  这两种特征选择方法可以很好地总结出不同类别的关键词，并且随着文章数目的增多，关键词的准确性会更加的准确。我认为这种算法可以应用于快速总结某一不熟悉领域，例如我搜集了大量关于不同领域的深度学习的论文，我可以通过特征选择方法快速地得到不同方向文章的共性特征或者近期研究热点，而不需要一篇篇地人工看摘要，从而更快更省时地指导方向。  (2) 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“党政办公室”、“教务部”、“招生办公室”、“研究生院”和“科学技术部”方面的信息，即从5个类别中选择最相关的一个）。  要对使用特征选择和不使用特征选择的分类效果进行对比和分析。用题(1)中的文档进行训练和测试，每个类别中的20篇用作训练，10篇用作测试。  请在报告中附上系统整体设计、代码截图（不要复制源代码，请用截图的方式）、运行结果截图和详细的文字说明。程序要有详细注释。对使用的中文分词工具作简要介绍。（20分）  +++++++++++++++++++++++++++++++++++++++++++++++++++++++  **系统整体设计：**  **整体设计：**  +++++++++++++++++++++++++++++++++++++++++++++++++++++++  **代码截图、运行结果截图和详细的文字说明：**  **第一步：读取文章数据集**  读取过程中，需要读文本进行处理。其中包括由于爬虫下来的文章会有\u3000等多余字符，所以需要进行额外处理。然后利用jieba分词生成文章列表。postingList和classVec一一对应，为文本与标记的正确分类。  **根据文章列表生成词袋**  **接下来就是训练NB分类器的过程了**  **朴素贝叶斯计算公式如下：**  **具体为如下训练伪代码：**  **具体实现如下：**  **最终可以得到每个词项的条件概率**  **condprob[term][c]代表term在类别c中的条件概率**  **朴素贝叶斯算法应用：**  **计算公式如下所示：可以加上log函数解决小数丢失问题**  **具体为如下训练伪代码：**  **applyMultinomialNB会返回概率最大的文档类别。**  +++++++++++++++++++++++++++++++++++++++++++++++++++++++  **对使用的中文分词工具作简要介绍：**  调用jieba.cut函数进行分词  jieba分词0.4版本以上支持四种分词模式：   1. 精确模式：试图将句子最精确地切开，适合文本分析； 2. 全模式：把句子中所有的可以成词的词语都扫描出来, 速度非常快，但是不能解决歧义 3. 搜索引擎模式：在精确模式的基础上，对长词再次切分，提高召回率，适合用于搜索引擎分词 4. paddle模式：利用PaddlePaddle深度学习框架，训练序列标注（双向GRU）网络模型实现分词。同时支持词性标注。   +++++++++++++++++++++++++++++++++++++++++++++++++++++++  **使用特征选择时的分类效果：**  整体正确率94%  +++++++++++++++++++++++++++++++++++++++++++++++++++++++  **不使用特征选择时的分类效果：**  整体正确率86%  +++++++++++++++++++++++++++++++++++++++++++++++++++++++  **使用特征选择和不使用特征选择的分类效果进行对比和分析：**  可以看到，在使用了特征选择后的分类效果更佳。这是因为使用了特征选择后，可以更准确的对类别进行关键词区分，而未使用的过程中，则会有更多冗余词项进行干扰。  另外两种方法下，对于教务部的分类准确率都不太理想。结合具体文章来看，我觉得可能原因是整理的文章类型太多，数据量太少没有很好的符合规律。因此会导致分类准确度的下降。我猜测的一个可行的方法是增加样本容量，丰富相应的词项。  +++++++++++++++++++++++++++++++++++++++++++++++++++++++  **此外，关于朴素贝叶斯分类算法，谈谈你的理解，并通过类比、关联或演绎的方式，举一个在日常的学习生活中可以应用的例子（要求积极向上且能自圆其说）**  这个算法有一些类似于机器学习中的文本分类任务，只不过是将机器学习方法转化为更基础的统计学方法。可以看到，从一个简单的数学公式出发就可以完成一项看似复杂的文本分类任务，概率论在计算机中的应用真的好神奇。NB算法可以用于小样本下的文本分类任务，这也是一个不太消耗资源的计算方法。  ++++++++++++++++++++++++++++++++++++++++++++++++++++++  **其他（例如感想、建议等等）。**  本次实验完成并掌握基本的分类和聚类算法。说实话，一开始我对于数学公式不是特别理解，但是实现后才发现它真的管用。也是这次实验，让我进一步发现了数学和算法之间的紧密联系，这真的很神奇。有些公式只有自己亲手去实现了，才能更好的去理解，仅从数学角度去理解，真的太抽象了。 |

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：  2022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。