**实验报告**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 姓名 | 安怡歌 吴佳慧 | | 学号 | 2021301041107 | 年级 | 2021级 |
| 成绩  （考核档次或分数） |  | | 指导教师 | 吴志强 | 专业 | 图书馆学 |
| 实验类型：  □独立实验课  √含实验的理论课  □计划外自选开放实验  □学生自主式开放实验  □大学生科技竞赛 | | | 实验日期：  20 24 年 2 学期，  第 7 周 | | 实验学时数：  学时 | |
| 相关课程：《数字图书馆原理与技术》 | | | | | | |
| 相关科研项目：无 | | | | | | |
| 实验名称 | | 实验一 数字资源信息组织方法实验 | | | | |
| 1. **预习部分（实验目的、实验基本原理等）**   实验原理：  1. 数字资源信息组织  数字资源信息组织主要对资源的文本内容进行细粒度分析，选择可以表达主题概念以及具有检索意义的关键词作为特征词，用这些特征词来解释资源的内部特征，达到资源的可检索利用以及可分析性；  主要技术包括：基于文本分析的内容标引技术；基于文本分析的资源分类技术。  2. 内容标引  文本分析过程实现的就是数字资源内容标引，选取的词就是标引词。在分析信息内容的基础上，用某种语言或标识符把信息的主题概念及其具有检索意义的特征表示出来，作为信息与检索的基础。标引的结果是标引词向量。  3. 文本分类  文本分类是根据分类体系将文本进行归类的过程。需要预先确定一个分类体系，针对特定的数字资源文档集合，随机选择一部分文档作为训练集，基于分类体系对训练集文档进行类别划分，确定每一文档所属类别，这一过程通常是人工分类或人机结合分类；基于训练集文档标引词向量与类别的对应关系，训练文档分类模型；对于文档集中为分类的文档，与训练集中的所有文档进行文本相似度计算，以文本相似度为依据，利用分类算法，预测每一文档的类别。  主要仪器设备（含必要的元器件、工具）：  实验目的：   1. 掌握数字图书馆数字资源的标引原理与方法 2. 掌握数字图书馆数字资源的分类原理与方法   实验内容：  1.通过网络、文献资料调研基于文本分析（分词、特征词提取）的数字资源标引方法以及基于文本分析的数字资源分类方法，掌握大规模数字资源的标引、分类的相关原理、技术与基本流程。  2.通过网络获取用于文本分析、文本分类相关的工具、软件（或程序代码），并了解其在大规模数字资源信息组织中的功能和作用。  3.利用这些工具和程序代码以及已学过的相关专业知识，完成数字资源标引与分类的实验。  4.设计UI界面，显示数字资源标引与分类的实验过程中所涉及的流程、中间结果以及最终的实验结果。  实验要求：   1. 在调研基础上，分析描述数字资源标引、分类的基本流程 2. 分析所下载的自动分类相关工具或软件（程序代码）的基本功能，选择其中你认为适合的工具或代码作为本次实验的工具参考。 3. 利用所提供的训练集和实验集，完成大规模数字资源的标引、分类实验。 4. 总结在实验中遇到了哪些问题？如何解决的？本次实验有哪些收获？将上述问题及其心得体会总结成实验报告。实验报告以电子版方式，由学委收齐后，于本课程期末考试之前提交。   目录  [1 理论知识 4](#_Toc163916129)  [1.1 基于文本分析的数字资源标引方法 4](#_Toc163916130)  [1.1.1 标引原理 4](#_Toc163916131)  [1.1.2 标引技术 4](#_Toc163916132)  [1.1.3 标引流程 5](#_Toc163916133)  [1.2 基于文本分析的数字资源分类方法 5](#_Toc163916134)  [1.2.1 分类原理 5](#_Toc163916135)  [1.2.2 分类技术 5](#_Toc163916136)  [1.2.3 分类流程 6](#_Toc163916137)  [2 技术工具 6](#_Toc163916138)  [2.1 相关工具总结 6](#_Toc163916139)  [2.1.1 软件/平台 6](#_Toc163916140)  [2.1.2 库（Library） 6](#_Toc163916141)  [2.1.3 工具包（Toolkit） 7](#_Toc163916142)  [2.1.4 框架（Framework） 7](#_Toc163916143)  [2.2 本实验使用的工具 8](#_Toc163916144)  [2.2.1 8](#_Toc163916145)  [2.2.2 8](#_Toc163916146)  [2.2.3 8](#_Toc163916147)  [3实验操作：数字资源标引与分类 8](#_Toc163916148)  [3.1文本预处理 8](#_Toc163916149)  [3.2文本特征提取 8](#_Toc163916150)  [3.3分类模型构建 8](#_Toc163916151)  [3.4 文本相似度计算 8](#_Toc163916152)  [3.5 预测类号 8](#_Toc163916153)  [4 UI设计 8](#_Toc163916154)  [4.1 制作可视化页面展示实验流程、中间结果及最终分类结果 9](#_Toc163916155) | | | | | | |

|  |  |
| --- | --- |
| **二、实验操作部分（可续页）**  *1. 通过网络、文献资料调研基于文本分析（分词、特征词提取）的数字资源标引方法以及基于文本分析的数字资源分类方法，掌握大规模数字资源的标引、分类的相关原理、技术与基本流程。* 1 理论知识1.1 基于文本分析的数字资源标引方法1.1.1 标引原理 文本分词和特征词选取共同构成数字资源内容标引过程。将文本转化为可处理的结构化形式，并利用数据挖掘技术，通过分词算法和词频统计方法将文本转化为一系列的特征项；采用向量空间模型来表示文本向量，每个特征项则为向量中的一个维。 1.1.2 标引技术  1. 分词：分词就是将连续的字序列按照一定的规范重新组合成词序列的过程。   常用分词方法：  ① 基于词典的分词方法：又称机械分词方法，按照一定策略将待处理的中文字符串与一个“充分大”的机器词典中的词条进行匹配，若在词典中找到某个字符串，则匹配成功，词典分词方法是目前应用最广泛的分词方法。  ② 基于理解的分词方法：在分词的同时进行句法、语义分析，利用句法信息和语义信息来处理歧义。  ③ 基于统计的分词方法：利用字与字之间的相邻共现频率作为分词的统计准则，通过对语料库中相邻贡献的各个字组合的频率进行统计，计算他们的共现信息，当共现信息达到一定的阈值，便可认为他们可以构成一个词。   1. 特征词提取：原始特征词集中并非所有的特征词都可以作为标引词，特征词选取技术就是对表达文本内容特征的标引词选取技术。   特征词选取技术：  ① 频度/逆文本频度法：TF/IDF——t 在文档中综合权重  如果t 在某一特定文档内是高语词频率，但在整个文档集合中是低文档频率，他就具有高权重的TF/IDF 值。t 的重要性随着它在文档中出现的次数成正比增加，但同时会随着它在整个文档集中出现的频率成反比下降。  TF/IDF 法中的t 综合权重计算公式如下：    ② 信息增益法  用于文本分类领域的信息增益，主要通过某个特征项在一篇文档出现或不出现的次数来预测文档的类别。在文档特征的选择上，如果某一特征项的信息增益大，它对文档主题概念的贡献作用就大，以此来作为标引词选择的标准。  在特征项选取中，信息增益指将某一特征项的重要性放在分类问题中考察，考察该特征项对文档分类的重要性，进而考察该特征项对表达文档主题概念的重要性。特征项t 的信息增益为：  ③ CHI法  CHI特征选择算法首先假设特征词语类别直接是不相关的，如果利用CHI 分布计算出的检验值偏离阈值越大，那么更有信息否定原假设，从而接受特征词和类别有很高的关联度。  CHI 算法中，根据特征项t 与类c 之间的相关性来判别t 的贡献度：   1.1.3 标引流程  1.2 基于文本分析的数字资源分类方法1.2.1 分类原理  1. 预先确定一个分类体系 2. 针对特定的数字资源文档集   ① 对数字资源文档的分类要运用有关文本分析的方法和技术，并以文档的标引词向量作为分类的基础，从文档数据集中提取描述文档类的模型，并把数据集中的每个文档归入到某个已知的对象类中。其核心是比较文档的文本相似度。需要合适的分类算法。  ② 利用文本分析的方法和技术，将文档集的文档转换成标引词向量；基于分类体系对训练集文档进行类别划分，人工分类或者人机结合确定每一文档所属类别；基于训练集文档标引词向量与类别的对应关系，训练文档分类模型；对于文档集中未分类的文档，与训练集中的所有文档进行文本相似度计算，以文本相似度为依据，利用分类算法，预测每一文档的类别。  ③ 将文档集训练为一个分类模型，通过比较待分类文档与分类模型中所有文档的文本相似度，根据一定算法，将待分类文档归到合适的类中。 1.2.2 分类技术  1. KNN 分类算法   通过计算测试集和训练集所有样例的相似度（或邻近度）并根据其相似度的大小确定k个最近邻居，根据k-最近邻居的类别，采用多数表决方案，将测试样例指派到其k-最近邻居的多数类中。   1. 决策树分类算法   决策树分类算法将文档视为一个数据对象，文档的各个标引词作为数据对象的属性。首先，所有数据对象都分为一类；计算每个属性的信息增益，选择最大信息增益的属性作为分类节点，将数据进行分割；该过程递归进行，确保每个节点上的数据都属于同一个类别，知道没有属性可以进行数据分割时就停止分割。   1. 贝叶斯分类算法   就文本分类而言，贝叶斯定理认为，一个文本对象特征向量中的每个维（特征词）都是相互独立的。  X={a1,a2,....,am}为一个待分类文档，ai为x的一个特征属性。有类别集合C={y1,y2,....,yn}，计算P(y1|x),P(y2|x),...,P(yn|x)。  如果P(yk|x)=max{P(y1|x),P(y2|x),...,P(yn|x)},x∈yk，其中，因为分母对于所有类别为常数，只要将分子最大化皆可，由特征属性的条件独立性，则有：   1.2.3 分类流程   *2. 通过网络获取用于文本分析、文本分类相关的工具、软件（或程序代码），并了解其在大规模数字资源信息组织中的功能和作用。* 2 技术工具2.1 相关工具总结2.1.1 软件/平台  1. Weka   Weka是一个开源的机器学习软件，提供了各种文本分类和文本挖掘算法的实现，提供了大量的机器学习算法和数据预处理工具，涵盖了分类、回归、聚类、关联规则挖掘等多种任务。用户可以使用图形界面或命令行界面来运行各种文本分析任务。   1. KNIME   KNIME是一个开源的数据分析平台，提供了丰富的文本处理和文本分析工具。用户可以通过拖放节点来构建数据处理和分析流程，包括文本预处理、特征工程、模型训练和评估等。 2.1.2 库（Library）  1. Jieba   jieba 是一个常用的中文分词库，用于将中文文本切分成词语。它是基于统计方法和词典匹配的算法实现的，简单易用，适用于许多中文文本处理任务。   1. NLTK   NLTK（Natural Language Toolkit）是一个Python库，提供了丰富的自然语言处理功能，包括文本预处理、分词、词性标注、命名实体识别等，适用于文本分析和分类任务。   1. scikit-learn   scikit-learn是一个Python机器学习库，提供了各种经典的机器学习算法，包括支持向量机、朴素贝叶斯、决策树等，可用于文本分类任务。   1. SCWS   SCWS （Small Chinese Word Segmentation）是Hightman开发的一套基于词频词典的机械中文分词引擎，它能将一整段的汉字基本正确的切分成词，采用的是采集的词频词典，并辅以一定的专有名称，人名，地名，数字年代等规则识别来达到基本分词。SCWS不仅可以处理普通的中文文本，还可以处理特殊领域或方言的文本，因为它的词典可以根据需要进行定制和扩展。此外，SCWS还提供了多种分词模式和参数设置，可以根据实际需求进行调整，以达到更好的分词效果。 2.1.3 工具包（Toolkit）  1. HanLP   HanLP 是一款面向生产环境的自然语言处理工具包，提供了中文分词、词性标注、命名实体识别等功能，具有一定的准确率和稳定性。   1. THULAC   THULAC是清华大学自然语言处理与社会人文计算实验室开发的中文词法分析工具包，具有较高的分词准确率和速度，适用于各种中文文本处理任务。   1. Stanford CoreNLP   Stanford CoreNLP是斯坦福大学开发的自然语言处理工具包，提供了中文分词、文本分类、命名实体识别等功能，具有良好的准确性和性能。   1. pkuseg   pkuseg是一款由北京大学开发的具有高准确率和高速度的中文分词工具包，支持领域自适应和用户词典扩展，适用于各种中文文本处理任务。   1. ICTCLAS   ICTCLAS 是由中国科学院计算技术研究所开发的最受欢迎的中文分词系统（中文分词工具包），是最早的中文开源分词项目之一。ICTCLAS 的功能模块包括中文分词、词性标注、命名实体识别、未录词检测和用户自定义词典。   1. FastText   FastText 是Facebook开源的文本分类工具，主要用于文本分类和文本表示学习。它通过将单词表示为子词的集合，并使用连续词袋模型（CBOW）来学习单词向量。FastText可以高效地处理大型文本语料库，并且通常比传统的词嵌入方法（如Word2Vec）更快速且表现更好。FastText还提供了文本分类功能，可以快速训练并用于对文本进行分类任务。 2.1.4 框架（Framework）  1. PyTorch   PyTorch是一个深度学习框架，提供了动态图机制和丰富的模型库，适用于文本分类和其他自然语言处理任务。PyTorch提供了Embedding层、RNN、LSTM、CNN、Transformer等模块用于处理文本数据，同时支持交叉熵损失函数和多种优化器。此外，PyTorch还包含了许多预训练的文本分类模型，如BERT、GPT等。   1. TensorFlow   TensorFlow是一个由Google开发的开源机器学习框架，用于构建和训练机器学习模型，尤其擅长深度学习任务。它提供了一个灵活的、可扩展的平台，可用于各种机器学习任务，包括自然语言处理。TensorFlow提供了文本预处理工具、神经网络模块、优化器和损失函数等功能，可用于构建、训练和评估文本分类模型。 2.2 本实验使用的工具2.2.12.2.22.2.3 *3. 利用这些工具和程序代码以及已学过的相关专业知识，完成数字资源标引与分类的实验。* 3实验操作：数字资源标引与分类 数字资源标引的过程包括文本预处理（主要为中文分词）和文本特征选取，将在3.1和3.2具体展开。数字资源分类的过程包括分类模型构建、文本相似度计算和预测类号。 3.1文本预处理 中文分词  去停用词  词性标注 3.2文本特征提取 计算特征词权重并提取高频特征词 3.3分类模型构建 结合高频特征词与训练集训练分类模型 3.4 文本相似度计算3.5 预测类号 使用分类模型对测试集进行分类  *4. 设计UI界面，显示数字资源标引与分类的实验过程中所涉及的流程、中间结果以及最终的实验结果。* 4 UI设计4.1 制作可视化页面展示实验流程、中间结果及最终分类结果 | |
| **三、实验效果分析（包括仪器设备等使用效果、实验完成情况）** | |
| **教师评语** |  |