****

**大学生创新训练计划**

**项目申报书**

**项目名称： 新闻网站错别字检查软件**

**项目负责人： 段士郁**

**所在学院： 计算机学院**

**专业年级： 2018级**

**学 号： 2018141461333**

**手 机： 18315173077**

**电子邮箱： 2268922282@qq.com**

**指导教师： 王运锋**

**项目起止年月： 2020年1月至2020年10月**

**项目参与学生人数： 5人**

**四川大学教务处制**

年 月

填写说明

一、凡申报**四川大学“大学生创新训练计划”**必须填写本申报书。创新训练计划项目是本科生个人或团队，在导师指导下，自主完成创新性研究项目设计、研究条件准备和项目实施、研究报告撰写、成果（学术）交流等工作。

**二、“项目所属一级学科和代码”**参考《普通高等学校本科专业目录和专业介绍（2012年）》。

三、**“项目开展支撑平台”**指支撑本项目开展的国家级和省部级重点实验室（中心、平台等）、国家双创示范基地平台、教学实验中心（实验室）、企业、事业或其他单位等，表中填写平台名称，可以多个。

四、**“项目组成员”**人数原则上不超过五人，应排序。

五、本书应该填写完整、内容详实、表达准确，数字一律填写阿拉伯数字。

六、报送申报书的电子文档至负责人所在学院。

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 项目名称 | 新闻网站错别字检查软件 | | | | | | | |
| 申请类别 | ☑科学探索与工程技术类  □人文艺术与社会科学类 □交叉学科类 | | | | | | | |
| 申请经费 | 3000 元 | | 起止时间 | | 2020年1月至2020年10月 | | | |
| 项目所属 一级学科和代码 | 计算机科学与技术 0812 | | | | | | | |
| 项目开展 支撑平台 | 大学生创新训练计划 | | | | | | | |
| 项目来源  （可多选） | ☑进课题组、进实验室、进科研团队参与的科研项目  □国家级和省部级重点实验室（中心、平台等）、国家双创示范基地平台支持申报项目  □基于学术型社团的创新创业项目  □“青年红色筑梦之旅”计划项目  □交叉学科创新项目  □基于前期创新创业成果、进一步拓展的创业项目  ☑其他 指导老师项目 | | | | | | | |
| 负责人之前参与大创项目情况 | 无 | | | | | | | |
| 项目成员之前参与大创项目情况 | |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | | 负责人/团队成员 | 项目编号 | 项目名称 | 立项级别 | 项目类别 | 立项年份 | 结题成绩 | | 王月（负责人） | 201910610957 | 一种高性能聚酮阴离子交换膜的制备及表征 | 校级大创 | 创新训练 | 2018 | 无 | | 林怡瑞(负责人)/汤怡心 | C2019106057 | 制革工业无铬鞣制及其配套染整过程的生命周期评价 | 未立项 | 创新训练 | 无 | 无 | | | | | | | | |
| 项目负责人基本信息 | | | | | | | | |
| 姓名 | 学号 | | | 专业年级 | | | 所在学院 | |
| 段士郁 | 2018141461333 | | | 2018级 | | | 计算机学院 | |
| 性别 | 手机 | | | 电子邮箱 | | | 身份证号 | |
| 男 | 18315173077 | | | 2268922282@qq.com | | | 500102200002281735 | |
| 项目组成员基本信息 | | | | | | | | |
| 序号（含排序） | 1 | 2 | | | | 3 | | 4 |
| 姓名/性别 | 郭素雲/女 | 王月/女 | | | | 桂江龙/男 | | 汤怡心/女 |
| 学号 | 2018141461053 | 2018141424166 | | | | 2018141511031, | | 2018141041098 |
| 专业年级 | 2018级 | 2018级 | | | | 2018级 | | 2018级 |
| 所在学院 | 计算机学院 | 材料科学与工程学院 | | | | 空天科学与工程学院 | | 文学与新闻学院 |
| 手机 | 18212755540 | 13568633345 | | | | 19983474712 | | 15680950790 |
| 电子邮箱 | 1876355309@qq.com | yueWang@163.com | | | | 343348459@qq.com | | 2578078812@qq.com |
| 身份证号 | 522426199906208220 | 510503200107067022 | | | | 36068120000510903X | | 510704200010180042 |
| 签名 |  |  | | | |  | |  |
| 指导教师1 基本信息  （非交叉学科类项目只允许一位指导老师） | | | | | | | | |
| 姓名 | 所在学院或单位 | | | 研究方向 | | | 职称/职务 | |
| 王运锋 | 计算机学院 | | | 多源信息融合、空管监视技术、无源定位、航空电子技术 | | | 教授 | |
| 性别/年龄 | 手机 | | | 电子邮箱 | | | 签名 | |

|  |  |  |  |
| --- | --- | --- | --- |
| 男/44 | 13088029093 | yfwang@scu.edu.cn |  |

|  |  |
| --- | --- |
| **项目内容概述(限200字以内)** | |
| 本项目设计的网站错别字识别系统将对新闻网站进行数据分析文本提取。采用利用开源的逆向最大匹配算法进行改进制作成分词器,参考stanford-parser开源代码进行依存句法分析，并对一些散串优化处理，通过以上方法对网站句子进行分词和语义分析。以开源n-gram为语言模型优化该模型并利用互联网中的语料库进行大量训练，精准识别分词后句子的搭配错误及逻辑错误。最终生成检查报告， 旨在对新闻网站错别字检测精度和速度等多方面优化和完善**。** | |
| **项目特色创新点概述（限100字以内）** | |
| 1. **针对句子不是机械式的局部的分词和查错，本项目通过优化逆向匹配算法合语料库和依存树进行句子的语义分析，精确的达到分词目的。** 2. **优化开源n-gram模型，进行机器学习精确进行错误检测。** 3. **对整个网站进行检测，生成报告。** | |
|  |  |
| **项目组成员分工** | |
| **姓名** | **承担工作内容** |
| **段士郁** | **错别字检测算法设计、模型训练** |
| **郭素雲** | **句子分词算法设计、模型训练** |
| **王月** | **句子语义算法设计、模型训练** |
| **桂江龙** | **网站文本的爬取及整理，** |
| **汤怡心** | **语料库的收集整理，文案** |

|  |
| --- |
| **一、项目简介（研究内容、目的意义、具体目标、国内外研究现状分析及评价等）** |
| 1. **研究内容**   本项目是针对新闻网站的错别字检查软件，帮助网站建设者和广大民众更好的管理和识别网站的错误内容，提高网站内容的质量以及用语的规范化。用户可以通过输入网址来链接到目标网站，并识别出目标网站的错别字内容并将检查结果打印出来。   * 1. 新闻网站数据爬取   通过网页的URL得到这个网页的源代码，编写特定的java爬虫来爬取新闻网页的相关数据，并从源代码筛选出需要的数据。将爬取下来的数据中的html文件加以甄别，并将数据以保存到系统中。   * 1. 语料库的收集及分类整合   运用机器学习的原理，实时收集，整理，更新语料库。保证语料库能够识别网络中的新潮词语。   * 1. 自然语言分词和语言分析   对保存的本地的文本进行分词处理，得到几个字的词语，对词语进行判别使得甄别错别字这一功能实现起来更加方便，并通过语意分析，再次判别相关词语是否符合作者所要表达的意思，使得程序的精度得到进一步地提高。   * 1. n-GRAM模型/开源程序   通过丰富的网络资源，适度调用n-GRAM模型来完善程序，并参考网上已有的相应程序的开源代码。   * 1. 识别软件开发   确定需求后，利用Pycharm、Eclipse等开发平台设计建立相应的软件系统的体系结构，并将整个系统分解成若干个子系统或模块，定义子系统或模块间的接口关系，对各子系统进行具体设计定义，并实现各个功能模块，最后完成测试。最终将程序做成软件形式，进一步方便用户的使用与程序的管理及更新。   1. **目的意义**   在目前互联网迅猛发展的背景下，网络文字影响力日益增大，而多音字、多义词等的存在使网络错别字现象十分普遍，在一定程度上导致了网络错别字的增长。作为网站运营的人员更应该注意避免输入错别字。 随着拼音输入法越来越智能， 越来越多的人开始依赖输入法， 在打出的字中往往因为追求速度而忘记校正或选择正确的汉字， 尤其是输入法的联想输入功能， 一方面确实提高了输入效率， 另一方面同声异义词却成为了联想输入出现错别字的“罪魁祸首”。用拼音输入法打字如果一味求快，那么打出来的字更可笑，比如“你嚎（你好） ”、 “兽击号（手机号） ”等。  错字忽略不计，在传播展示时会造成可大可小的影响。一个看似很小的错误，有时候却能影响甚远，例如国务院关于“严重错别字”指标的说明：  示例：  1. 将“中华人民共和国”错写成“中华人名共和国” ，将“人大常委会”错写成“大人常委会”等；  2. 将“中国共产党”错写成“中国共产常”等；  3. 将“祖国万岁”错写成“祖国万死”等；  4. 将“科学发展观”错写成“科学发展现”等；  5. 将“社会主义”错写成“杜会主义”等；  6. 将“反腐倡廉”错写成“友腐倡廉”等；  7. 将“扶贫”错写成“扶贪”等；  8. 将“严禁公款吃喝”错写成“严谨公款吃喝”等；  9. 将“政务公开领导小组”错写成“政务公开领导小姐”等。  也许只是网站维护人员的一个无心之失， 把国名、 国家机构名称， 以及党和国家领导人姓名写错，但是在传播影响上极有可能被冠以“背离了社会主义核心价值观”的恶劣头衔，而产生难以想象的影响。  减少网络错别字，不仅对人们、特别是文化基础不够牢固的人群起着良好的引导作用，而且对中华文化的传承和发展有着不可忽视的影响。  作为网站的运营人员应该严把审核网站中是否存在严重错别字，对网站信息内容进行检测，在细节之处确保网站运营良好，但是人工抽查等方法不仅容易忽略错别字，而且十分低效。  因此开发一款网络错别字检查软件，为大众提供一个方便有效、符合社会发展趋势、提高全民文化素养并且兼具实用性、普遍性与简便性的检查工具，显得尤为重要。为此， 本项目旨在为大家提供了精准而快速的网站错别字检查软件， 帮助快速分析、 定位网站错别字词问题。   1. **具体目标** 2. 通过输入网站方式对网站内每个网页中的文字逐次进行快速抓取网页数据并精准提取数据中的自然语言部分。为错别字的甄别提供文本信息。 3. 运用机器学习的原理，实时收集，整理，更新语料库。保证语料库能够识别网络中的新潮词语。给予程序以生命力，让软件可以随着时代的潮流不断发展并紧跟时代，随着语言的发展软件的甄别水平与精度也会得到相应的提高。 4. 通过逆向匹配，依存树等分词技术，对提取到的自然语言进行基于语义进行精准分词，并解析句子中的语句结构。为下一步采用数据库中的词语进行组合做基础，得到众多的词语排列组合。 5. 通过精确分词后的词字数组，导入优化n-gram模型结合常用语库对词字数组进行快速检查并精准的判断其文字错误及逻辑错误。通过上一步得到的词语的排列组合与相应语句中的相比对，得到相应的概率通过与系统中的阈值相比对，从而查询出错别字。 6. 利用Pycharm、Eclipse等开发平台设计软件的各个功能，设计UI，并进行测试。最终设计出完整的软件。 7. **国内外研究现状分析及评价** 8. **研究现状：**   自然语言处理，是指用计算机对自然语言的形、音、 义等信息进行处理，即对字、词、句、篇章的输入、输出、识别、分析、理解、生成等的操作和加工。实现人机间的信息交流，是人工智能界、计算机科学和语言学界所共同关注的重要问题。自然语言处理的具体表现形式包括机器翻译、文本摘要、文本分类、文本校对、信息抽取、语音合成、语音识别等。可以说，自然语言处理就是要计算机理解自然语言，自然语言处理机制涉及两个流程，包括自然语言理解和自然语言生成。近年来，自然语言处理处于快速发展阶段。各种词表、语义语法词典、语料库等数据资源的日益丰富，词语切分、词性标注、句法分析等技术的快速进步，各种新理论、新方法、新模型的出现推动了自然语言处理研究的繁荣。互联网与移动互联网和世界经济社会一体化的潮流对自然语言处理技术的迫切需求，为自然语言处理研究发展提供了强大的市场动力。  http://5b0988e595225.cdn.sohucs.com/images/20180820/d08df901bfab48cba80c4b7c21bf5c63.jpeg  我国直到上世纪80年代中期才开始较大规模和较系统的自然语言处理研究，尽管较国际水平尚有较大差距，但已经有了比较稳定的研究内容，包括语料库、知识库等数据资源建设，词语切分、句法分析等基础技术，以及信息检索、机器翻译等应用技术。  当前国内外出现了一批基于 NLP 技术的应用系统，例如 IBM 的 Watson 在电视问答节目中战胜人类冠军；苹果公司的 Siri 个人助理被大众广为测试；谷歌、微软、百度等公司纷纷发布个人智能助理；科大讯飞牵头研发高考机器人……以阿里巴巴为例，阿里自然语言处理为其产品服务，在电商平台中构建知识图谱实现智能导购，同时进行全网用户兴趣挖掘，在客服场景中也运用自然语言处理技术打造机器人客服，例如蚂蚁金融智能小宝、淘宝卖家的辅助工具千牛插件等，同时进行语音识别以及后续分析。阿里的机器翻译主要与其国家化电商的规划相联系，可以进行商品信息翻译、广告关键词翻译、买家采 购需求以及即时通信翻译等，语种覆盖中文、荷兰语、希伯来语等语种，2017 年初阿里正式 上线了自主开发的神经网络翻译系统，进一步提升了其翻译质量。但相比于性能趋于饱和的计算机视觉和语音识别技术，自然语言处理因技术难度太大、应用场景太复杂，研究成果还未达到足够的高度。  http://5b0988e595225.cdn.sohucs.com/images/20180820/ec4efbac011544609d41fb8e625ab71a.jpeg  **图二 语法错误检查发展**  近几年主要是利用机器学习模型来实现语法错误的检测，例如n-gram模型  虽然此种方法在英语学习领域取得了不错的效果，但是并没有许多的研究人员把其应用到中文辅助学习的领域。  另外就是深度学习的方法，其中利用较为广泛的模型为递归神经(Recurrent neural network，RNN)。其用于处理和预测序列数据。RNN 可以更好地利用传统的神经网络无法建立的模型信息，但这也带来更大的挑战——长期依赖的问题。  对于中文错别字订正任务，现在由 The SIGHAN Bake-offs 会议4进行组织，其提供一个供研究学者交流比较自动中文错别字订正策略的平台。然而，与英语或其他字母语言不同的是，汉语是一种以音调性的音节和特点为特征的语言，其中每个字母都被发音为音调性音节。在中文里，没有单词分隔符或单词之间的界限，每个中文“单词”的长度很短，大多数情况下可能只有两个或三个字符。此外，中文字为表意的象形文字，许多中文字符的形状类似，一些词语在形状和发音方面甚至相似，所以错别字的类型比其他语言更多。到目前为止，已经有很多的研究正在进行。最近一次会议统计的方法也基本上都是基于规则的方法，如使用 n-gram 模型， 自定义的规则，翻译模型，线性回归等等。   1. **当前研究存在的问题**   对于中文检查错误来说，虽然在机器学习的方法下，中文辅助学习工具已经取得了不错的发展，但是由于机器学习本身的一些缺点使得这些辅助学习系统还是没有达到预期的效果，其缺点主要表现在数据量过少，非常容易过拟合，为了避免过拟合会使用剪枝处理，但是剪枝处理又会丢失一些细微的特征，在这些特征中有许多非常重要的信息，会使得某些错误无法被分类。如朴素贝叶斯生成的模型的缺点就是需要独立假设，这样会牺牲一定的准确性，使得分类性能不高。  此外机器学习处理的文本数量较少，没有很强的自适应性，在特定的环境中会发挥很好的效果，但是由于中文词汇量的巨大，常用字的繁多，所以只能做一些已经标定好的特定中文信息，使得其不能全面覆盖所有的中文特征，并且分类模型不利于保存，这就使得用机器学习模型做出的计算机中文辅助学习策略不能完全适用于各种外国学习者的中文文本分析，所以其准确率会发生降低，得不到想要的效果。这就使得越来越多的学者把更多的目光放在深度学习模型身上。 对于中文错别字订正任务，前文所总结的这些方法都有其自身的局限性，那就是很难适应新的语言环境或者是出现新的词语后更新模型较为困难，泛化能力不够好的问题。现今网络已经进入到千家万户，利用网络的便利性很容易获得更加全面的词语库，并及时更新词语的特点，对于错别字的辨析将有很大的帮助。  总体来说，计算机在自然语言处理方向主要存在的难点有以下几点：   1. 分词过程中对词语边界的界定。我们让计算机来界定词语边界往往是根据词义特点或者词语组成法则，但中文尤其是在口语语言中，许多词存在边界模糊的现象。 2. 词语歧义的处理。人们在生活中辨别歧义句主要是结合语境和上下文，在生活场景中的交流本身就对随后可能发生的语义有一定的预判。而计算机因缺乏语境处理能力，因此对歧义句处理仍有一定困难。 3. 对模糊句法的处理。在自然语言中所存在的语法规则是人们根据大量的语言所归纳出的一个规律，这种规律能帮助计算机更好的理解自然语言。但是，许多时候语言中，尤其是口语中的句法不是一个很明确的死规矩，而是存在一定的“错误”，计算机在这种偏离既定规则的句法识别过程中容易差错。 4. 语言行为的识别。在生活中不少这样的例子，“你能告诉我叫什么名字吗？”一个好的否定回答不是告诉对方“不能”，也不是告诉对方“我不能告诉你我叫汤怡心”，而是解释理由。这种基于生活经验的语言行为在计算机应用上仍需提升。 5. **参考文献** 6. 梁桢. 基于尾字词典的逆向回朔中文分词技术研究. 武汉工业学校硕士学位论文，2010，32(4)：101-102 7. 熊艳秋 严碧波.基于jsoup爬取图书网页信息的网络爬虫技术.长江大学电子信息学院，2019，27（4）：61-63 8. 张劲松，袁健.回溯正向匹配中文分词算法[J].计算机工程与应用 2009， 5(22)： 32-134. 9. Mays, Eric, Damerau. Context-based spelling correction. Information Processing and Management. vol.27, no.5, 1991,517 – 522. 10. 罗桂琼，费洪晓，戴弋.基于反序词典的中文分词技术研究.计算机技术与发 展，2008，18(1)：80-83. 11. 郑豪,何彦雨.基于Java平台的分布式网络爬虫系统研究[J]. 科技创新与应用. 2017(01) 12. 马金山,刘挺,李生.基于n-gram及依存分析的中文自动查错方法[C].//Chinese Information Processing Society of China%Chinese Languages Computer Society, USA%Northeastern University%Tsinghua University.Advances in Computation of Oriental Languages-Proceedings of the 20th International Conference on Computer Processing of Oriental Languages.2003. 13. 姚文琳.汉语依存句法分析方法的研究与实现[D].山东:中国海洋大学,2009. DOI:10.7666/d.y1828769. 14. IGERT J．Probabilistic Detection of Context-Sensitive Spelling Errors[C]／／Proceedings of LREC，2004：1-4． 15. 王璐.中文文本真词错误自动校对算法研究[D].浙江:浙江工商大学,2018. 16. 沈涛.结合N-gram模型与句法分析的语法纠错[D].江苏:东南大学,2017 17. 张松磊.中文拼写检错和纠错算法的优化及实现[D].湖北:华中科技大学,2018. 18. 陈智鹏,吕玉琴,刘华生, 等.基于N-gram统计模型的搜索引擎中文纠错[J].中国电子科学研究院学报,2009,4(3):323-326. DOI:10.3969/j.issn.1673-5692.2009.03.020. 19. KALCHBＲENNEＲ N， GＲEFENSTETTE E， BLUNSOM P．A convolutional neural network for modelling sentences［C］ / /Proceedings of the 52nd annual meeting of the association for computational．Baltimore， USA: ［s．n．］， 2014: 655－665． |

|  |
| --- |
| 1. **研究技术路线及可行性分析** |
| **技术路线**  **一. 整体框架**    **图1 整体流程图**   * 1. **具体实现**   **1.** **爬取网站中所有网页的源代码和链接**   1. 用递归遍历完成对网站每个网页内链接的获取和源码的获取 2. 剔除重复链接   **2. 爬取网页数据信息**  在爬取网页时，使用一个html页面解析jar包——jsoup。Jsoup是一款java的html解析器，可以直接解析某个URL地址、HTML文本内容。将用户输入的html文档解析转换称一个Doucument对象进行处理。在此基础上使用Dom方法遍历文档，爬取文档中的数据，通过jsoup elements对象支持的选择器语法，抽取想要的内容，进行选择。  利用java语言，导入io流、util、，通过创建maven项目，导入jsoup包，从而使用java爬虫将网页中的文本信息下载到本地并以在系统中储存起来  编写该爬虫脚本基本步骤：   1. 确定首页URL。 2. 发送POST、GET请求，获取网页相关的数据，爬取其中所有的html元素。 3. 解析网页具备的数据并将其中的文本数据挑选出来。 4. 获取到的文本数据在存储进系统中。   **3. 基于Hash结构词典逆向最大匹配算法中文分词技术**  **3.1 构建 Hash结构词典**  Hash结构词典通常由3个部分组成：   1. **首字Hash表：**以散列方式存储词汇首字，每个词汇首字含一个指针指向词汇索引表的本字词指针域首地址 2. **词汇索引表：**顺序存储各字词指针，每一个指针指向词典正文中的对应词汇首址 3. **词典正文：**存储词典中的所有词条。Hash结构词典结构如图l所示。     **图2 逆向最大匹配词典结构**  **3.2 逆向最大匹配算法**  **最大匹配法的基本思想**：根据分词词典中最长词条的字数来确定待切分材料中 所选取的汉字串的字数，即假设最长词条字数为 n，则选取待切分材料的前 n个汉字组成的汉字串与分词词典中词条进行匹配，如匹配成功则认为该 n个汉字为一个词，把其切分出来，然后选取材料中的下一个 n 个汉字进行处理；如匹配不成功，则去掉匹配汉字串的最后一个汉字，将剩下的汉字作为新的匹配字段进行匹配，依此类推，直至匹配成功为止。  逆向最大匹配算法从待切分材料的末尾开始取词来与词典中的词条进行匹配，并且每次匹配不成功时去掉的是最前面的一个汉字。  **算法过程：**   1. 设词典中最大词条所含汉字个数为n个，待处理字符串所含汉字个数为L个 2. 取待处理字符串尾部的m个字作为待匹配词，其中m=min(n，L) 3. 然后取待匹配词的第1个字作Hash运算，在首字Hash 表中取得指针 4. 顺序取得词汇索引表中的指针；依次获取词典正文中 的词汇与待匹配词进行匹配 5. 若匹配成功．则分出该词，L=L-m，否则转至(8)执行； 6. 若L>0。则转至(2)执行 7. 成功，结束分词 8. m=m-1，若m>0，则转至(3)执行 9. 分词失败     **图3 逆向最大匹配**  **4．基于 n-gram 查找局部错误**  **4.1 理论基础**  n-gram以马尔可夫模型为理论基础，对一词串W=，可认为词 （1≤i≤n）的出现与上文的前n个词相关，则词串 W 出现的概率可通过如下的方法得出：    图三 n-gram概率  在语言模型的构造中，可以字、词、词性或词义等作为 n-gram 的统计单元。通过大量的语料联系，和合理的模型构造，对对文本进行局部分析，以查出文本中的局部错误。  它基于如下假设：对正确的语言现象，词与词之间的共现概率较高，对一些不符合语法的错误语言，词与词之间的共现概率较低。  **4.2 模型分析**  当n=1时，此时的模型被称为一元文法模型(unigram model)，退化为上下文无关模型。而上下文无关模型在考虑当前符号本身的概率分布的同时并没有考虑该符号所对应的上下文环境，因而统计信息不足导致在实际应用中系统精确度往往不高  当n=2时，出现在第i位上的符号仅与它前面的一个历史符号 有关。  当n=3时，出现在第i位上的符号仅与它前面的两个历史符号 有关  一般来说，n 取2或3，n=2称为bigram，n=3称为trigram：  Bi-gram: P(T)=p(|begin)\*p(|)\*p(|)\*\*\*p(|)  Tri-gram: P(T)=p(|begin1,begin2)\*p(|,begin1)\*p(|)\*\*\*p(| ,)  其中： P(|begin) = (以为开头的句子数) / (句子总数)  **4.3 查错原理**  将词语对列表分别导入两个n-gram模型，得到两个模型的2-gram和3-gram分数，总共两个分数，若两个分数均低于某个阈值，判定该处错误。（阈值为人工选定的数字）。  **5．基于依存句法分析中文语法错误**  在中文自动校对中，n-gram 模型能够有效的查出文本中的局部错误，但对远距离搭配错误却无能为力。词与词之间的搭配是看两者之间的语义关联强度，而依存树的边正可以用来体现这种语义关联度，如果一个句子存在语法错误，那么建成依存树也应该存在一条边是不合理的，我们可以用这条边来判断是否出现了语法错误。为此，我们利用句子的依存分析结果查找文本中的全局错误。  **5.1 理论基础**  基于依存的句法分析，其任务是通过识别句子每个单词的句法中心，推导出输入句子的句法结构。句子输入可以W=表示。依存句法分析时需要利用每个词的词性信息 构成丰富的特征，句子对应的词性序列ｔ＝通常也要作为依存句法分析的输入。输出一般是一棵标记完整的句法依存树。  **5.2 模型建立**  为了确定依存搭配是否合理，必须事先获取足够的依存知识，包括下两步：   1. 建立依存树库   本文以大量字节的文本作为训练集，利用已有的依存分析器，对其进行依存分析，每个句子的分析结果为一个依存树，这样，得到一个大规模依存树库。   1. 建立依存知识库   以依存树库为知识源，从中学习依存知识。这里，我们主要是获取树库中的依存搭配对，计算其搭配强度，然后存储到知识库中。应用该依存知识库，可判断句子中远距离的依存搭配是否合理。  **5.3例子**  “在那次火灾中，他身上的皮肤很多部分都浇了。”  句中，“烧”字错成了“浇”字，如对句子进行局部的分析，能够发现，“部分都浇了”和“部分都烧了”的概率相近，在局部范围内，符合与语法规律，即这个错误属于 n-gram 模型无法查出的全局错误。  对这个句子进行依存句法分析，得到如下依存树：    图 4错误句子的依存树  由依存分析结果，得到“浇”字的远距离依存搭配对： （在 ← 浇） （皮肤 ← 浇） ， 其中，（皮肤 ← 浇）属于远距离的搭配错误，如能利用已有知识，判定该搭配不合理，则应用依存文法分析，就能够查出文本中的全局错误  交通比较发展的地方。 觉醒了中国人的心  **nsubj(发展-3, 交通-1)**  root(ROOT-0, 觉醒-1)  advmod(发展-3, 比较-2) aux:asp(觉醒-1, 了-2)  acl(地方-5, 发展-3) nmod:assmod(心-5, 中国人-3)  mark(发展-3, 的-4) case(中国人-3, 的-4)  root(ROOT-0, 地方-5)  **dobj(觉醒-1, 心-5)**  punct(地方-5, 。-6) punct(觉醒-1, 。-6)  我已经熟练了基本的驾驶技术。 他是修理脚踏车的拿手。  nsubj(熟练-3, 我-1) nsubj(拿手-6, 他-1)  advmod(熟练-3, 已经-2)  **cop(拿手-6, 是-2)**  root(ROOT-0, 熟练-3) acl(拿手-6, 修理-3)  aux:asp(熟练-3, 了-4) dobj(修理-3, 脚踏车-4)  amod(技术-8, 基本-5) mark(修理-3, 的-5)  case(基本-5, 的-6) root(ROOT-0, 拿手-6)  compound:nn(技术-8, 驾驶-7) punct(拿手-6, 。-7)  **dobj(熟练-3, 技术-8)**  punct(熟练-3, 。-9)  **6. 散串处理**  散串是指文本中出现的连续单字词。通过对真实的文本进行分析，能够发现，文本中的错误周围多出现散串，且散串的长度较大。如：  “女人 比 男人 爱 吃 零食。”  “女人 比 那 人 爱 吃 零食。”  第二个句子中“男”字错为“那” ，导致散串的长度由二变为五。  另外，散串中常包含单独成词概率小的字。汉语中，有些字不能够单独成词，如“囱”字只能和“烟囱”构成词，如果这样的字单独出现，则文本中一定有错；有些字很少单独成词，如“保”字，常和“保持” 、 “保护” 、 “保证”等词共同出现，单独成词的情况很少。  散串为确定文本中的错误提供了两个有价值的信息： **长度和单字词概率。**  充分利用这两个信息，先对句子中的散串进行分析：   1. 根据散串的长度和其中的单字成词概率， 能够找出文本中的多数可能出错的地方 2. 根据字的三元模型进行进一步的分析处理根据散串的长度和其中的单字成词概率 3. 人为设置阈值进行判定是否出现错误。   **可行性分析**   1. **相关资料：**相关论文期刊中含有大量关于自然语言关于分词处理，分析句中错误方法等的研究性学术文章供我们参考和学习，这些相关的研究资料将极大地帮助我们进行网站中自然语言的处理，对于错误检查算法的优化，语料库的使用……；在相关论坛有相应成熟的技术可供参考学习；互联网中的中文语料库数量丰富可供我们训练使用。这些都将帮助我们的项目顺利完成。 2. **成员概况：**小组成员对于网页数据分析、自然语言处理等相关领域都怀有极大的研究热情和探索欲望，并且分工安排合理有序，团队具有强大的凝集力，能够用理性的眼光和团结的行动直面项目中出现的问题。 3. **编程语言：**该项目主要运用Python 和JAVA两种面向对象的编程语言，小组成员都较好地掌握了以上两种编程语言，并能熟练地运用在项目之中。 4. **开发平台：**该项目的开发平台在Android或Windows均可，所以设备和开发环境等硬件条件均已具备。 5. **进度安排：**该项目已经通过多次讨论，最终制定出合理详尽的项目研究进度表有了详细的安排与计划。 6. **指导老师：**四川大学计算机学院的王运锋教授作为我们项目的指导老师，一直从事信号与信息处理方面的研究工作，在自然语言处理领域能够给我们技术方面的建议。 7. **相关技术**：对于网站的爬取技术实现简单；对于逆向最大匹配算法，及依存句法分析法在当下已经有相当成熟的应用可对句子进行精确的分词；对于n-gram模型在错别字识别方面的应用也十分成熟。 |

|  |
| --- |
| **三、研究基础（对项目的参与动机、已有知识储备、相关研究和训练基础）** |
| 1. **参与动机：** 2. 小组成员对于网页数据分析和自然语言处理有着浓厚的兴趣，这两块是我们一直都想涉足研究的领域，通过实现本项目，会给我们一个实践的机会，积累相关的经验。 3. 我们小组在互联网中调研发现，许多的查错或者纠错软件，都只保留了文本类型的错别字检测，很少有软件进行网站的错别字检测，而且在我们进行试用的过程当中我们发现绝大多数的错别字检查软件都普遍存在识别精度不高的问题，一些明显的错误在软件中并不能识别出。基于以上问题，我们小组想优化相关算法结构，来提高错别字检查软件的识别精度。 4. 我们希望能做出更加方便更加人性化的内容，就像是我们可以通过输入一个网站的网址，我们做出的软件就能告诉我们网址内容出现的全部错别字，不需要我们一个一个字的输入，可以直接调取相关的网址便能完成操作。 5. 在当下的错别字检查领域当中已有许多的成熟的理论可供我们学习，并且语料库也是日益完善，基于以上特点，我们小组希望在此基础上设计出更加完善的错别字识别软件。 6. **已有知识储备：**   C/C++、Java高级程序开发语言、Python、  NLP (Natural Language Processing)自然语言相关文献的学习、  数据结构与算法、 JavaScript脚本语言。   1. **相关研究和训练基础：**   1） 小组成员已经查阅了大量相关的自然语言处理、网页数据处理文献来了解该项目，制定了详细的项目开展的技术路线，并进行了项目的可行性分析。  2） 小组成员都具有一定的编程能力，知道如何做出相关的产品，并且相关成员对于自然语言有一些研究，能够从多角度进行错别字分析，在编写程序时思路也就会更加丰富。   1. **经费支持：** 根据下拨经费，购置相关书籍、文献、软件、硬件、服务器等，并完成相关算法的实现与软件的设计。 |
| **四、研究计划和进度（就文献查询、社会调查、方案设计、实验研究、数据处理、研制开发、撰写论文或研究报告、结题和答辩、成果推广、论文发表、专利申请等工作逐项计划时间，时间节点精确到月份）** |
| |  |  |  |  | | --- | --- | --- | --- | | **阶段** | **时间** | **具体任务** | **阶段成果** | | **项目立项** | 2019.11 | * 社会需求调查 * 分析可行性 * 查询相关参考文献 * 确定具体技术路线 * 进行技术分工 | 1. 项目申请书 2. 项目可行性分析报告 3. 项目初步开发计划 4. 项目分工计划 | | 2019.12 | * 根据分工制定相关技术的学习计划 * 购置相关学习资料 | | **项目初期** | 2020.1  2020.2 | * 建立体系结构，并分解成若干个子模块， * 细化分工模块及开发技术板块 | 1. 各板块技术学习计划 2. 完整项目体系和详细技术实现计划 | | **项目研发** | 2020.2 | * 爬取网站功能实现、文字初步整理（调整存储结构）及分析功能实现 | 1. 源程序代码   2、单元测试报告  3、初步实现网站错别字检查程序 | | 2020.3 | * 语料库的收集及分类整合 | | 2020.4 | * 自然语言分词功能实现，模型训练 | | 2020.5 | * 语言错误检查（散串分析、长短字词分析等）功能实现，模型训练 | | 2020.6 | * 程序UI设计 | 1. 设计并搭建程序UI | | **项目测试** | 2020.7  2020.8 | * 调试与整合程序 * 单元测试 * 相关问题的讨论与解决 * 综合测试 | 1. 单元测试报告 2. 总体测试报告 3. 完善的查错系统 | | **结题答辩**  **成果推广**  **论文发表**  **专利申请** | 2020.9 | * 结题答辩 * 成果推广 | 1. 结题报告 2. 软件产品 3. 成果推广 | |

**五、项目研究支撑条件（项目所依托的重点实验室（中心、平台）、双创平台、课题组等各类单位能提供的直接支持项目开展的软硬件设施和其他校内外资源）**

|  |  |
| --- | --- |
|  | |
| 1. **开发语言：java，Python** 2. **开发环境：JDK（Java Development Kit），Python解释器** 3. **开发工具：集成开发环境Eclipse，集成开发环境Pycharm** 4. **文档工具：office 2019** 5. **工作环境：学校创新实验室提供了良好的硬件设施和工作环境** | |
|  |  |
| **六、预期成果形式（可多选）** | |
| 1.□SCI论文 篇  2.□核心期刊论文 篇  3.□会议论文 篇  4.□内部编印期刊论文 篇  5.□授权发明专利 项  6.☑申请发明专利 1 项  7.□创新类竞赛获奖  8.□创业类竞赛获奖  9.☑其他 名称： 开发软件 | |
| **七、项目经费概算（按申报项目目标任务需要进行预算，经费执行情况将与结题考核成绩挂钩）** | |
| **1.经费来源（单位：元）**  申请项目专项经费 6000  **2.经费支出（项目总经费的65%，单位：元）**  （1）仪器设备费 500  （2）耗材费 800  （3）测试加工费 500  （4）国内会务及差旅费  （5）国外会务及差旅费  （6）文献/知识产权事务费 4000  （7）办公费（含文印、办公用品等） 200  （8）其他费用  **合计 6000**  3.**绩效奖励**（项目总经费的35%，单位：元）  学生不填 | |

|  |
| --- |
| **八、评审情况** |
| **指导教师意见：** |
|  |
| **指导教师（签名）： 年 月 日** |
| **学院推荐意见：** |
|  |
| **主管院长签名： 年 月 日** |
| **学校专家评审意见：** |
|  |
| **组长签名： 年 月 日** |
| **学校认定意见及批准经费：** |
|  |
| **学校负责人签名： 年 月 日** |