**软件学院创新项目实训**

**申请表**

**（2021版）**

项目编号 （申请人不必填写）

项目名称 基于文本分析的批量代码语法检测和实验报告查重系统

团队负责人 李成

手机号码 15688826021

项目时间 2021年3月 5日 至 2021年 6月13日

山东大学软件学院

**填报说明**

1. **项目目标**请简短说明项目的研发定位。
2. **指导教师**请联系在院内承担本科教学任务的教师。
3. **技术点**请列出项目中使用的关键技术点，最多不超过5个名词。
4. **参加人员**中的第一行必须为项目负责人。团队总人数4-5人，特殊情况（项目非常大）7-8人。所有成员必须为2018级软件学院大三在校本科生，其他参与人员请在其他人员中列出。
5. **项目资料地址**必须可公开访问，项目组每个成员都需要维护各自的博客，以便督导老师随时可以抽查项目实施过程。必须选用新浪、CSDN等知名服务提供商，不支持自己搭建的服务器。
6. **项目介绍**从项目背景、技术创新、工作内容、技术路线、实施方案等方面，详细说明项目的具体工作。请详细说明，可以添加附件，或者音视频等多媒体材料的网址。
7. **实施计划**按照时间节点，分阶段完成预定目标。
8. 申请承诺需要团队所有成员签字。

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **项目名称** | 基于文本分析的批量代码语法检测和实验报告查重系统 | | | | | |
| **实施时间** | 2021年3月5日至2021年6月13日 | | | | | |
| **项目目标** | 实现对于学生实验报告或代码文件中代码的语法问题和简单逻辑问题的检查，进行查重，并得出检测报告。 | | | | | |
| **技术要点** | OCR、lint、apache tika、自然语言处理、TF-IDF。 | | | | | |
| **指导教师** | 林丰波 手机号码：13665318136 | | | | | |
| **团队名称** | 晨曦 | | | | | |
| **参加人员** | **姓名** | **学号** | **专业** | **手机号码** | **电子邮箱** | **项目分工** |
| 李成 | 201800301202 | 软件工程 | 15688826021 | 609172827@qq.com | 组长，负责人，开发人员 |
| 吴玮桓 | 201800301011 | 软件工程 | 18818877088 | 1031648247@qq.com | 开发人员 |
| 牛翔宇 | 201800301108 | 软件工程 | 13296447689 | 1127857501@qq.com | 开发人员 |
| 王帅 | 201800301249 | 软件工程 | 13298209930 | 382486783@qq.com | 开发人员 |
|  |  |  |  |  |  |
| 其他人员 | | | | | |
| **项目资料地址** | | | | | | |
| **项目博客地址：**  https://blog.csdn.net/m0\_46817064  **Github地址：**  https://github.com/ChenXiTeam/codeChecker  **成员个人博客地址：**  李成：https://blog.csdn.net/qq\_43533083  吴玮桓：https://blog.csdn.net/qq\_43958699  牛翔宇：https://blog.csdn.net/jas0nxy  王帅：<https://blog.csdn.net/Mr__666> | | | | | | |
| **项目介绍（可跨页，可附加多媒体网址）** | | | | | | |
| 1. 项目背景   在软件工程专业学习人数较多，需要批改的实验报告等文档的数量也相对不少的情况下，批改实验报告必然会耗费学科老师，助教等相关工作者的大量时间以及精力。结合软件工程专业固有的专业特点，我们小组决定开发一款可以自动批量识别word文档中代码的语法错误以及简单逻辑错误的软件来帮助相关工作者节约时间。   1. 项目的目的和意义 2. 首先对于我们团队而言，这个项目不仅是对我们的一次考验，也是我们提升自身实力的一个机会，通过这种形式，将学到的知识用来解决实际生活中所遇到的问题，让我们更好的贴近实践。 3. 其次这个项目的成功对于实验报告的批改人员而言，无疑可以节约大量时间以及精力并且提高工作效率，给相关工作人员带来福音。 4. 项目的可行性分析 5. 首先通过以上分析，不难得知相关需求的确是存在的，换而言之，我们的项目有一定的市场价值。 6. 其次在技术层面，一是代码的识别问题，对于这个问题就word文档而言代码部分一般有文本和图片两种形式。对于文本，可以先进行关键字或者格式来判断是否属于代码模块；对于图片，则可以对图片进行内容提取，判断其是不是代码部分。二是代码的纠错问题，不难发现我们用的一些编译器都自带有代码的代码纠错功能，所以不存在技术上达不到的问题。 7. 最后在队员方面，我们团队有较强的学习能力，对于有挑战性的项目充满热情。 8. 项目的开发平台 9. 操作系统：Windows10。 10. 开发语言：根据相关语言的实际特点进行选择，如java，c++、python等。 11. 开发工具：eclipse、Idea、PyCharm等。 12. 项目的主要功能模块 13. 代码提取模块：对批量实验报告文件（.pdf、.doc等格式）以及源代码文件（.c、.java、.py等格式）的代码进行提取。 14. 代码检查模块：进行代码的静态语法检查。 15. 正文提取模块：提取实验报告中的正文文字进行提取。 16. 查重模块：对提取到的代码和正文进行内部相似度检测，并根据相似度进行抄袭判断。 17. 生成结果模块：是对系统判断结果的处理，将判断结果输出到一个文件中，方便教师、助教查看结果。 18. 项目可能涉及的主要技术 19. 代码的提取：   方式一：直接提交代码文件，如.c、.java、.py等格式源文件。  方式二：规定学生嵌在作业中的代码前后有标记，如：  ------begin-java-code------  public class HelloWorld {  public static void main(String[] args) {  System.out.println("Hello World!");  }  }  ------end-java-code------  方式三：从作业中的代码截屏图片中提取代码。比如可以采用python中的pytesserac库，也可以调用百度AI的OCR接口对图片中的代码进行识别。   1. 检查代码：静态代码检查：比如C语言代码可以用lint，Java代码用SonarLint等工具代码进行静态语法检查。 2. 提取作业正文文字：可以使用apache tika等工具提取，支持doc、pdf等各种常见格式文本，要求学生作业必须支持tika，比如不能对pdf使用口令保护。 3. 查重：使用jieba、gensim等自然语言处理库进行语言处理，collection库下的defaultdict 来计算单词频率，使用TF-IDF计算相似度进行简单的查重。 | | | | | | |
| 实施计划 | 1. 分析需求，项目可行性。 2. 搜集相关技术的信息，判断相关技术学习的时间成本。 3. 通过上网查阅资料、咨询指导老师等途径学习相关技术。 4. 进行系统设计，对系统各模块功能对应的进行简单分析。 5. 根据系统设计编程实现系统。 6. 进行系统测试，根据测试结果找出系统存在的问题，并加以修改。 7. 发布系统。 | | | | | |
| 预期成果 | 1、可以批量的对源代码文件或实验报告文档中的代码进行语法正确性进行检查，并可以对正文和代码内容进行检查。  2、最终成品一个界面友好的Windows程序，操作简单不需要工作人员额外的时间成本。  3、可以一定程度上减少人为误判的因素，提高实验报告成绩的客观性。 | | | | | |
| 指导教师评语 | （签字） 年 月 日 | | | | | |
| 本人郑重承诺，此申请书内容真实有效。    （所有团队成员签字）  年 月 日 | | | | | | |