附件1

**浙江工业大学本科生毕业设计（论文、创作）申报表**

2022年 12月 6日

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 题 目 | Design and Implementation of Forbidden Word Recognition and Rendering with Browser Environment | | | | | | | |
| 专 业 | 软件工程(中外合作办学) | | | 学生人数 | | | 1 | |
| 指导教师 | 李小薪 | | | 职 称 | | | 副教授 | |
| 课题内容：（国内外现状简介；重点要解决的问题）  随着互联网的普及，更多的企业开始进行数字化转型，通过建立企业的数字化平台来推动企业组织转变业务模式，通过新兴媒体平台发布信息。与此同时加强信息媒体发布的风险管理需求也逐步提升，不符合规定的词汇应当在发布时被正确指出，而人工审核的开销较大，使用机器审核成为了更多数企业的首选。目前阿里云的内容安全服务已提供了一套较为完整的内容审核工具。为了解决企业对内容发布平台内容审核的需求，拟开发一个浏览器环境下的内容审查交互模块，并实现如下功能：（1）能够对发布信息中的文字与文件进行违禁词的检查：内容发布者往往会通过多种方式进行信息的发布，例如微信公众号、企业官网网站、企业APP等，针对不同平台所使用的内容与格式也会有所不同，而不同的内容也应当对应不同的检测模式；（2）对上传的不同文件进行解析：内容发布的文件附件中往往包含了很多不同种类的文件，主要包括PDF、WORD、PPT这三种文件格式，阿里云服务的文件审核平台有文件大小的限制，同时为了提高用户使用体验，降低网络传输的时间损耗，将文件解析为文本再进行传输是一种更好的选择。在文件解析的过程中，需要解决不同种文件格式与版本问题，并精确识别该文件类型，增强该系统的鲁棒性。同时在文本提取的过程中往往需要用到不同的方法，寻找设计适合该文件种类的对应解析模块也是该系统的开发难点之一，例如部分模块虽然能完成文件解析，但其目的主要是针对文件在页面上的展示，针对提取文本的工作有相对大的难度；（3）精确显示违禁词在文件中的位置：为了让发布者能更好的找出违禁词，以便能在原文当中进行修正，该模块能准确的在文档中渲染出违禁词所在的位置，并提供违禁词目录提供快速检索违禁词在文档中位置的能力。首先需要解决如何将文件利用浏览器显示的问题，在控制开发成本的前提下选取合适的技术来进行前端的文件解析渲染工作，例如canvas，pdf.js等前端技术的选择使用；其次是渲染的精度问题，由于文件的字体，中英文的字体宽度等问题，在渲染上可能会出现不精确的问题；同时为了保证浏览器的流畅工作，需要采用懒加载的渲染模式，在此渲染模式下不需要的界面会被清除，项目需要解决在该模式下如何保持文本渲染的问题。 | | | | | | | | |
| 课题的准备情况及对学生的要求：（文献、仪器设备、材料、预做、工作地点及学生须具备的技能）  课题准备情况：指导教师已经对课题的设计与开发进行了详细的规划，并准备了与课题相关的技术资料及软硬环境设备。学生须具备的技能：1.要有一定的基于JavaScript的网页设计基础；2. 具备一定的通过网络和请教别人来解决所遇到的问题的能力。 | | | | | | | | |
| 课题内容性质 | | 软件开发 | | | | | | |
| 课题来源性质 | | 企业课题 | | | | 类 型 | | 设计 |
| 实习地点及时间 | | | 永安期货 8月至12月 | | 用机时数 | 20周 | | |
| 学科负责人意见：  签名： 月 日 | | | | | | | | |
| 主管院长意见：  签名： 月 日 | | | | | | | | |