### 基于浏览器环境的违禁词识别与渲染的设计与实现 Design and Implementation of Forbidden Word Recognition and Rendering with Browser Environment

学生姓名：唐晨宇 指导教师：李小薪

违禁词检测指筛查并标记文本中的违禁词的一种自动化技术。违禁词检测是一种重要的信息安全措施，常见于媒体平台。目前，金融企业数字化转型不断升级，线上平台的用户数进一步增长，对于发布内容的风险管理需求日益增长。因此，建立企业合规性检测平台对降低企业内容发布风险至关重要。

本文设计并实现了一个基于浏览器环境的违禁词检测系统，用于在发布内容过程中进行自动化违禁词检测。该系统由文件类型解析、内容检测和浏览器文件渲染等三个个模块构成：文件类型解析用于解析用户上传的文件类型，结合文件头十六进制数据与MIME类型进行类型检测，并通过配置文件确定与限制文件类型；内容检测用于提取文件中所需被检测的内容，基于云服务接口从输入文本中生成违禁词列表；浏览器文件渲染则将检测出的敏感文本进行定位并加上不同的颜色底纹进行高亮展示。此外，还探讨了文件违禁词识别中二进制文件的网络传输数据量较大的问题。由于二进制文件通常包含了富文本信息，文件通常较大，本文以文本内容传输取代二进制文件进行网络传输。首先对文件类型进行解析，根据不同的文件类型，采用不同的方式提取文本信息，进行网络传输。本系统能够对PDF文件、Office系列文件等主流的文件类型进行文本内容提取，极大提高了大文件的违禁词检测速度。

系统基于B/S结构，提供了一个交互式图形界面，提升了系统的易用性。用户可以上传需要检测的文件，系统能够快速确认文件内容的合规性，为后续进行相应的人工审核与追责建立了可视化手段。

关键词：违禁词检测，违禁词渲染，文件解析，Node.js