成员:RiCKy

红日学院

社区风控与发布内容自动化审核系统建设分享

目录

业务价值

以UGC（用户生成内容）为主的社区：

1. 社区风控系统的价值：

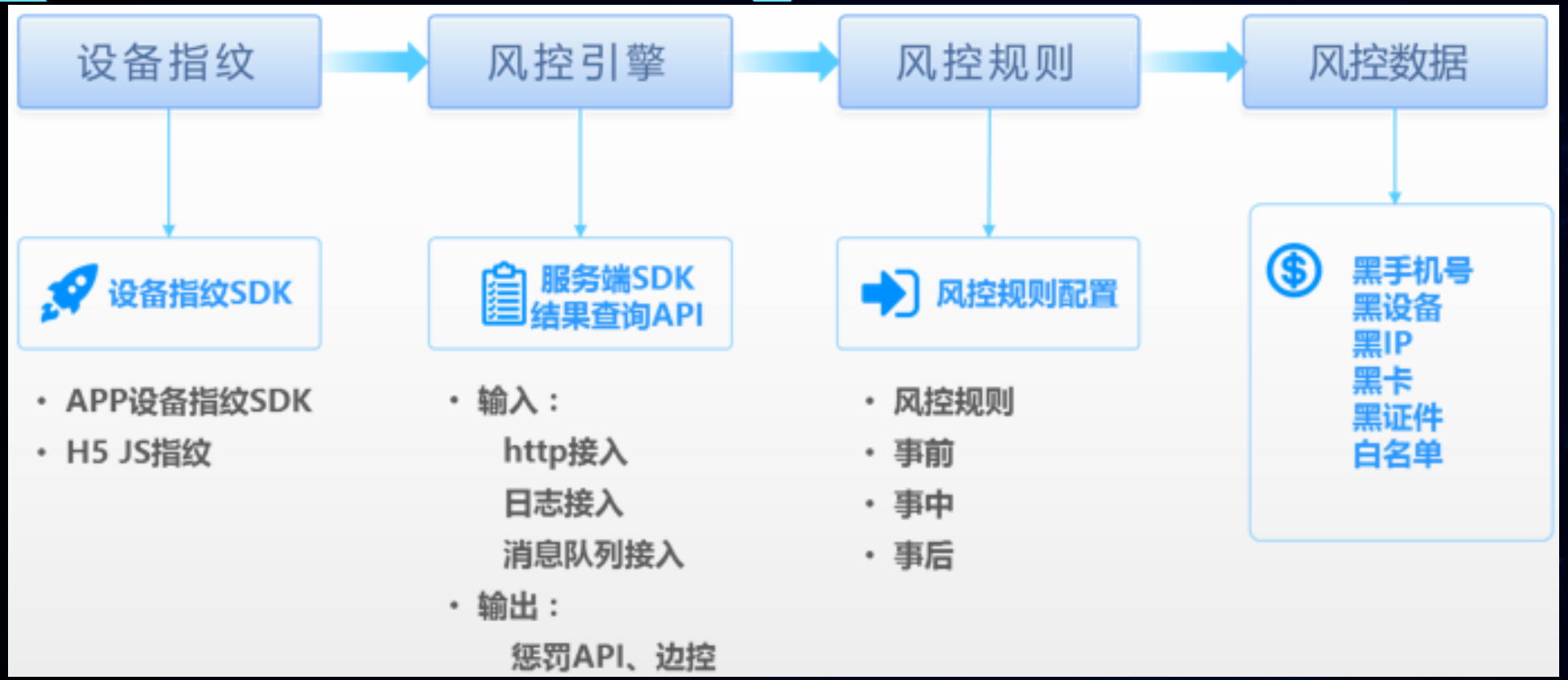
用户行为验证：注册/登录/领取试用/运营活动奖励等
防薅羊毛：免费试用和社区活动奖励
防止机器刷帖、防刷点赞、评论、人气

2. 发布内容自动化审核系统的价值：

用户发布的内容验证
拦截发布违法有害文字、图片、视频
防灌水、拦截广告、拦截无意义的重复内容

1 社区风控系统

1 风控系统组成

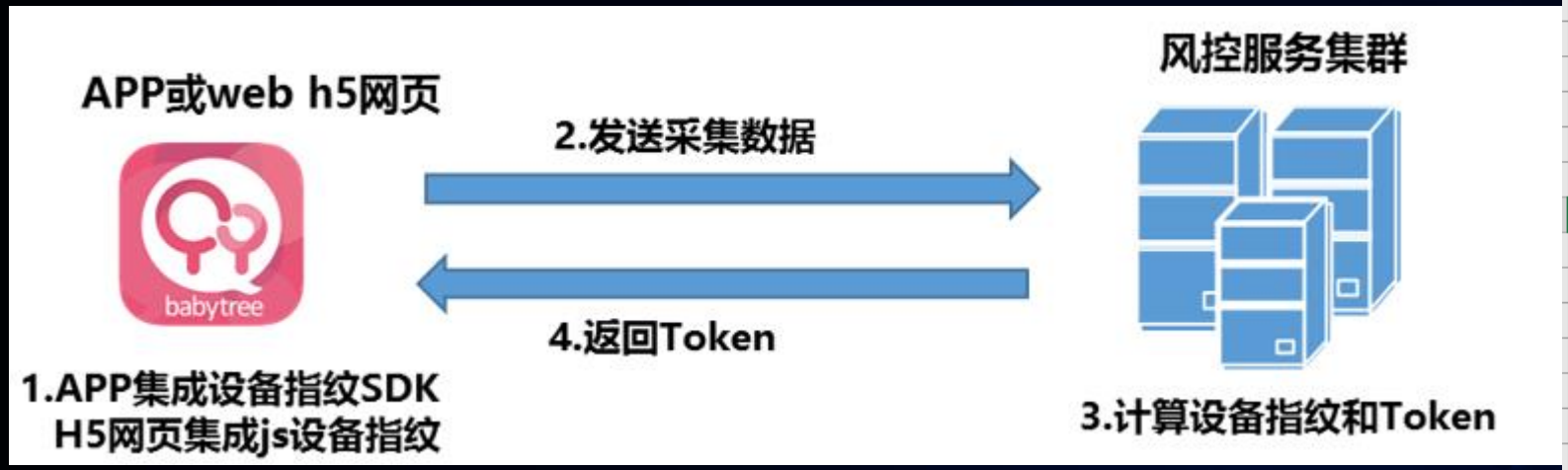


1、设备指纹：采集手机上几十个软硬件信息，识别模拟器、root越狱、改机/注入，唯一标识设备；Web网页指纹：js采集浏览器信息

2、风控引擎：实时/离线规则计算引擎

3、风控数据：手机号黑卡、数据卡（不能接打电话、只能收发短信）

2 设备指纹



Android风控SDK采集基础数据项

字段	字段解释
osver	Android操作系统版本
time	系统时间
cpuCount	手机CPU核数量
model	手机型号
screen	屏幕分辨率
cpuModel	CPU型号
btmac	蓝牙mac
boot	手机启动时间
appver	APP版本
appname	app名称
appdisplay	app
emu	模拟
ssid	wif
wifiip	wif
operator	运营
network	手机
mem	总内
sensor	传感
cpuFreq	CPU

IOS风控SDK采集基础数据项

字段	字段解释
os	手机系统类型:ios
osver	手机系统版本号
time	系统当前时间戳
appname	app名称
appver	app版本
apputm	app渠道
idfa	IDFA
idfv	IDFV
model	手机型号
carrier	所属运营商
mnc	移动网络号码
mcc	移动信号国家码
bssid	wifi地址，移动网是为"null"
ssid	wifi名称，移动网是为"null"
freeSpace	剩余空间
battery	电池电量
root	是否越狱
brightness	手机屏幕亮度
languages	系统语言
totalSpace	存储容量
boot	启动时间
dns	DNS
networkType	网络连接方式
countryIso	所属国家
sysname	系统名称
memory	手机内存
name	用户自定义名称
riskapp	高危软件列表(越狱采集)
riskdir	高危目录列表(越狱采集)
orientation	手机倾斜角度

模拟器

AMIDuOS
Andy
Bluestacks
Genymotion
KOPLAYER
MEmu
NoxPlayer
Remix OS Player
Windroy
YouWave
Android Studio内置的安卓模拟器
Visual Studio内置的安卓模拟器
Xamarin内置的安卓模拟器
ARChon（Chrome扩展，可运行Android app）

夜神模拟器
逍遥安卓
天天模拟器
游信模拟器
雷电模拟器
海马玩模拟器
itools模拟器
靠谱助手
51模拟器
新浪手游助手
腾讯手游助手
MuMu模拟器

手机+改机工具

改机工具通过劫持系统函数，伪造模拟指定手机（模拟器）的设备信息（包括型号、串码、IMEI、定位、MAC地址、无线名称、手机号等）的app，能够欺骗厂商在设备维度的检测，或是提供虚假数据以达到特定的盈利目的。

改机工具会从系统层面劫持获取设备基本信息的接口，app只能得到伪造的假数据。Android和iOS都有很多相应的改机工具，Android改机大部分都基于Xposed框架，需要root，iOS大多基于Cydia框架，需要越狱。

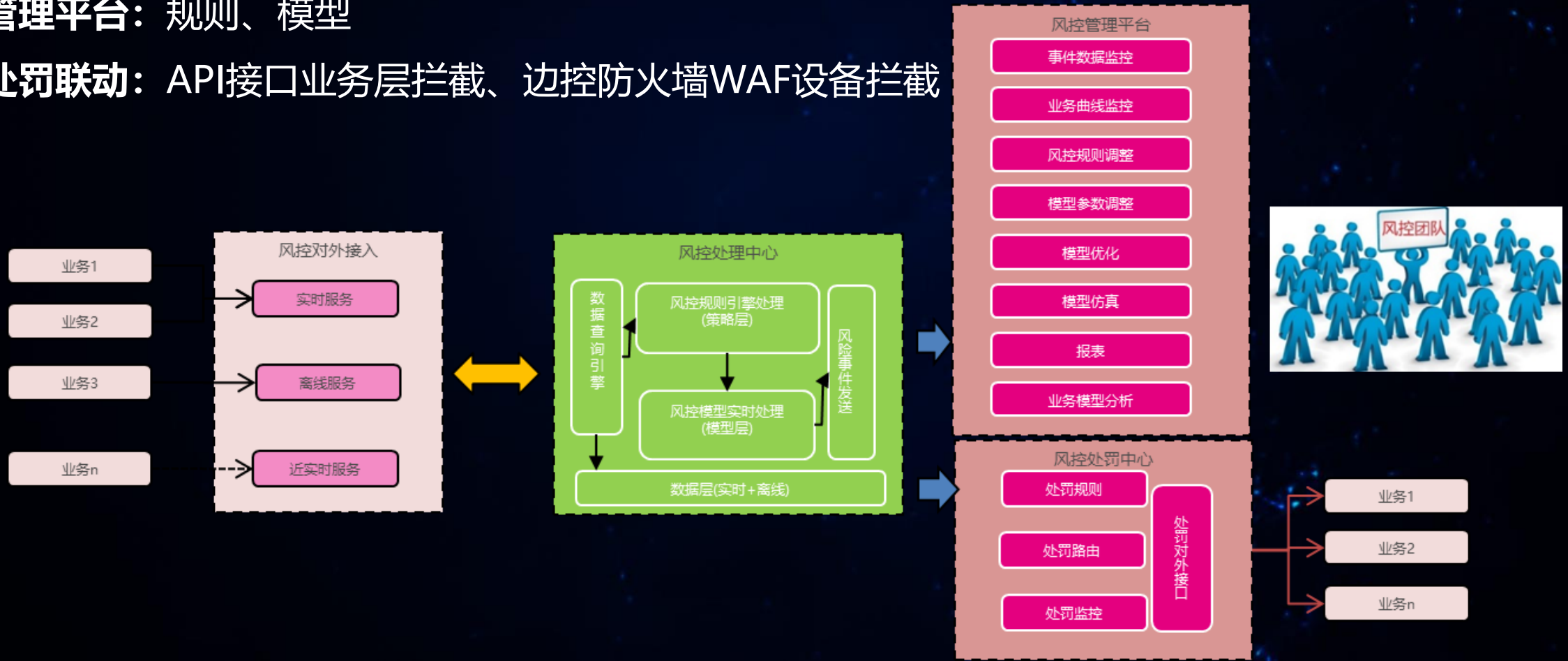
3 风控引擎

风控对外接口：http API(有代码侵入)、日志(无代码侵入，离线计算)

风控处理中心：实时、离线计算

风控管理平台：规则、模型

风控处罚联动：API接口业务层拦截、边控防火墙WAF设备拦截



4 风控数据

1、手机号黑卡：数据卡号、未开通语音服务卡（非常多）、黑名单手机号

(1)应对手机农场（成千上万部真实手机）可部分拦截；

(2)命中率：50%已经非常优秀；

(3)误伤率：实际上10%已经优秀，仅适用于某些场景，不能只依靠黑卡解决风控问题。

2、IP黑名单：代理IP、僵尸网络IP等；IP有效期较短，某些场景可辅助使用。

3、自产数据：设备黑名单、用户黑名单

2 内容审核系统

1 内容审核系统功能

频次限制

文本识别

屏蔽词

相似性

图片识别

图片鉴黄

OCR文本

视频识别

音频识别

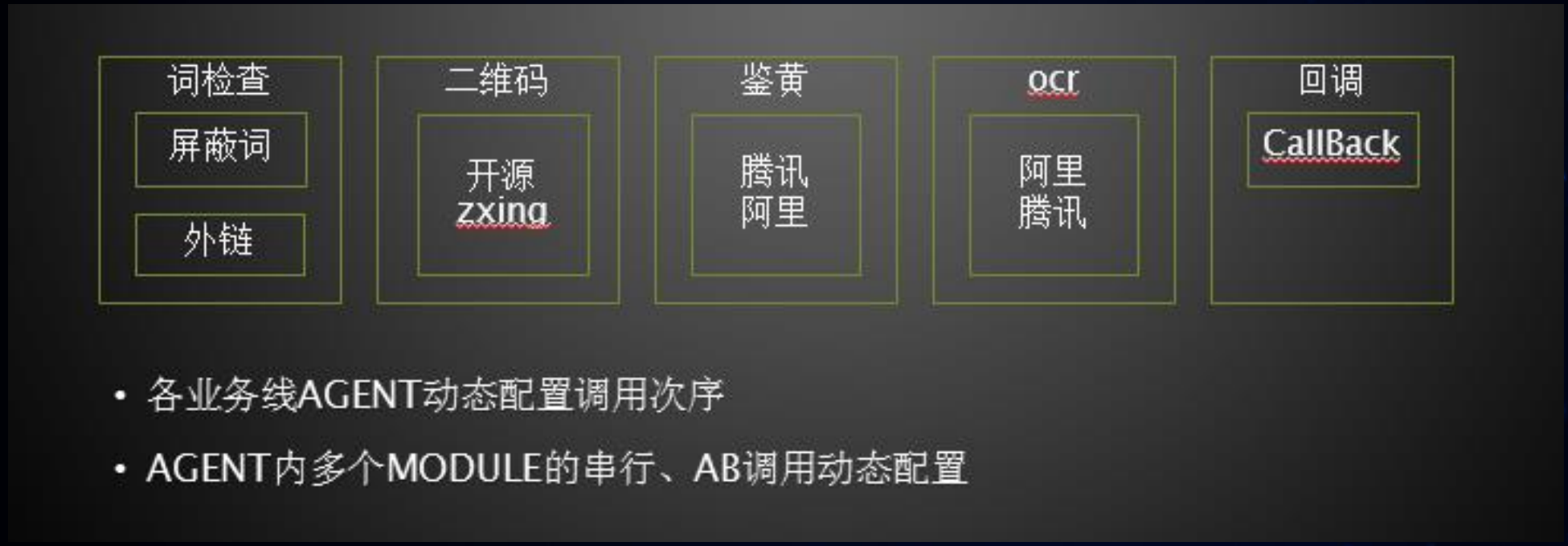
内容审核

前台抽检

内容召回

内容举报

2 代理和模块



(1) 同步检查： 文字、链接、必杀词、变义词、重复词
(2) 异步检查： 图片、视频、音频
(3) 内容召回： 复查
(4) 人工审核。
拦截动作： 禁止发布、先审核后显示、仅自己可见、删帖封号

3 文本识别

- **屏蔽词**
- 简单词
- 1：替换词
- 2：必杀词
- 3：先审后发词
- 4：先发后审词
- 组合词
- 1：香港+验血 （OR 看男女）

- **内容相似性**

文档去掉停顿词，通过开源结巴分词，计算每个词出现的频率，选择频次大于1的词建立词典，并获取词典的特征数，基于词典建立语料库，用语料库训练TF-IDF模型，计算文本的相似度。

4 内容审核后台

- 先审/先发/操作日志
- 召回
- 报表
- 配置
- 抽检
- 举报
- 白名单

3系统联动设计

黑名单共享和追杀

1.两个系统黑名单打通（设备黑名单和账号黑名单）：

(1) 被风控拉黑：不能发布内容

(2) 被审核拉黑：不能参加各种运营活动、商品试用、投票点赞等

2.追杀设计：

被任一系统拦截时，会根据用户行为划分为多个等级（标签），被拉黑的用户将被追杀到其他功能同样拦截。

4 拦截规则示例

拦截规则示例

风控规则：

1.拦截异常设备：模拟器/root越狱/改机/注入/安卓多开/无设备刷接口等

2.拦截一个设备多账号：注册、登录、促销单、抽奖、试用等活动

3.拦截设备在单位时间段内，操作频次阈值：求和、求平均、重复规律

4. 订单支付：订单付款账号相同（微信/支付宝/银行卡）视为同一人

5. 订单收货：收货地址相似、收货人手机号相同，视为同一人

6. 前置行为：是否有前置行为、是否按照前置步骤操作

7. 圈人系统配合风控系统：先从圈人系统圈出优质用户，在从风控系统过滤异常/僵尸用户，只给优质用户发奖励

发布内容审核：

1.必杀词、图片/视频识别、广告文拦截（送审）

2.重复词次数、变异字体（火星文等）

Thanks