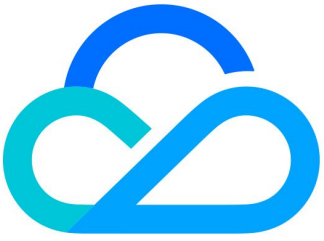
文本内容安全

操作指南



腾讯云

版权所有：腾讯云计算（北京）有限责任公司

【版权声明】

©2013-2025 腾讯云版权所有

本文档（含所有文字、数据、图片等内容）完整的著作权归腾讯云计算（北京）有限责任公司单独所有，未经腾讯云事先明确书面许可，任何主体不得以任何形式复制、修改、使用、抄袭、传播本文档全部或部分内容。前述行为构成对腾讯云著作权的侵犯，腾讯云将依法采取措施追究法律责任。

【商标声明】

腾讯云

及其它腾讯云服务相关的商标均为腾讯云计算（北京）有限责任公司及其关联公司所有。本文档涉及的第三方主体的商标，依法由权利人所有。未经腾讯云及有关权利人书面许可，任何主体不得以任何方式对前述商标进行使用、复制、修改、传播、抄录等行为，否则将构成对腾讯云及有关权利人商标权的侵犯，腾讯云将依法采取措施追究法律责任。

【服务声明】

本文档意在向您介绍腾讯云全部或部分产品、服务的当时的相关概况，部分产品、服务的内容可能不时有所调整。您所购买的腾讯云产品、服务的种类、服务标准等应由您与腾讯云之间的商业合同约定，除非双方另有约定，否则，腾讯云对本文档内容不做任何明示或默示的承诺或保证。

【联系我们】

我们致力于为您提供个性化的售前购买咨询服务，及相应的技术售后服务，任何问题请联系 4009100100或95716。

文档目录

操作指南
敏感词排行榜查询指引
纠错操作指引
纠错统计
审核结果纠错
纠错名单

版权所有：腾讯云计算（北京）有限责任公司

操作指南
敏感词排行榜查询指引

最近更新时间：2025-02-05 16:02:22

功能说明

用户可以使用敏感词排行榜功能对天御文本内容安全机审模型的审核结果中出现的**敏感关键词命中数据**进行查询。本文将介绍如何查看敏感关键词排行榜。

操作步骤

1. 登录 内容安全控制台 ，在左侧导航栏中，单击**数据统计 > 排行榜**。

> **说明：**
> 子账号使用敏感词排行榜查询如未被主账号授予查询权限，需要按照 子账号授权指引 完成 ListUsers 权限授权后，才可查询敏感词数据。

2. 在排行榜页面，根据业务需求，筛选字段应用名、场景名、内容类型（文本）、识别类型、处理建议、账号、时间范围、敏感关键词即可查看统计数据。

> **说明：**
> 时间范围字段筛选时长按照天进行筛选，筛选周期为1-30天。

3. 在排行榜页面，根据业务需求筛选对应字段后，即可查看纠错统计数据。

> **说明：**
> - 敏感词排行可显示排名前1000的敏感词。
> - 敏感词数据支持导出，单击下载按钮即可。

字段	说明
排行名次	按照敏感关键词命中次数正序排名
关键词	模型识别到的违规关键词
处置建议	模型识别到的违规关键词处理建议，包含违规、疑似
识别类型	敏感关键词命中违规的类型
命中次数	敏感关键词对应筛选条件下命中总次数
命中词库类型	命中词库包含系统默认库和自定义词库
操作	• 命中详情：单击**命中详情**，可前往 机审明细页面 查看对应命中的敏感关键词的汇总明细。 • 关键词详情： ○ 当关键词命中词库为系统默认库时，“关键词详情”将置灰，不支持查看系统默认词库明细。 ○ 当关键词命中词库为自定义词库时，单击**关键词详情**可查看当前命中词库，前往关键词命中详情页面。

4. 在排行榜页面，当关键词命中词库为**自定义词库**时，单击**关键词详情**可查看当前命中词库，前往关键词命中详情页面，并可以进行以下操作：

- **管理：**单击**管理**，将跳转至 关键词名单页面 ，管理该关键词所属词库。
- **删除：**单击**删除**，该命中关键词将会从命中的自定义词库名单中删除。
- **批量删除：**如果关键词同时命中多个自定义词库，勾选对应词库后单击**批量删除**，该关键词将从所选自定义词库中删除。

版权所有：腾讯云计算（北京）有限责任公司

关键词详情
关键词
我
处理建议
违规
识别类型
色情
自定义名单
批量删除
词库名称
处理建议
匹配模式
最近修改时间
操作
ly测试词库
违规
模糊匹配
2024-01-04 16:27:37
管理
删除

纠错操作指引
纠错统计

最近更新时间：2025-02-05 16:02:22

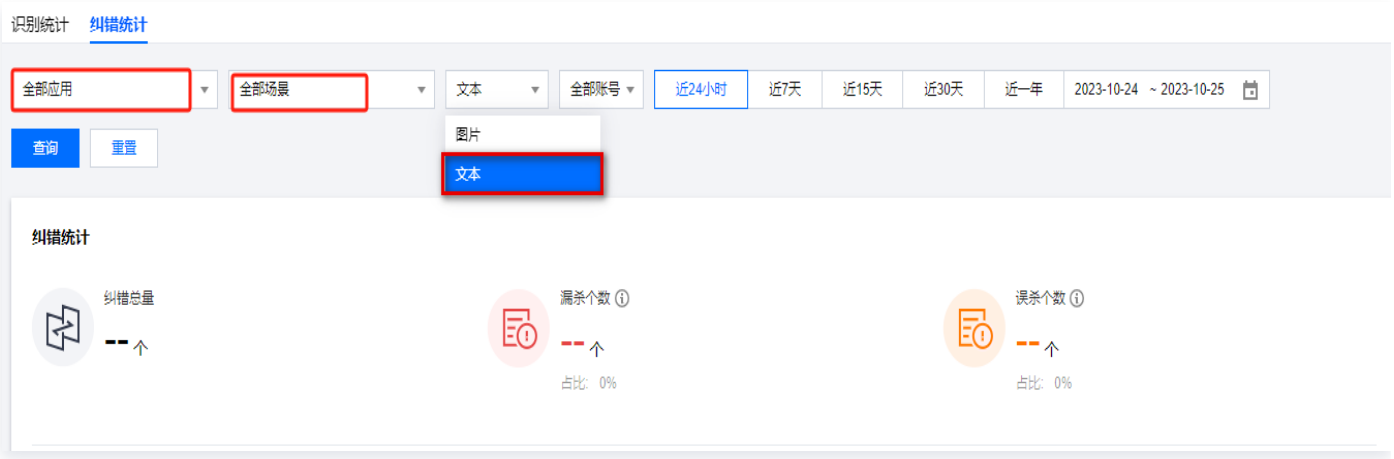
功能说明

用户可以使用 纠错功能 对天御文本内容安全机审模型的审核结果进行修改，以保障业务需求。
本文将介绍如何查看已完成纠错的案例的数据统计。

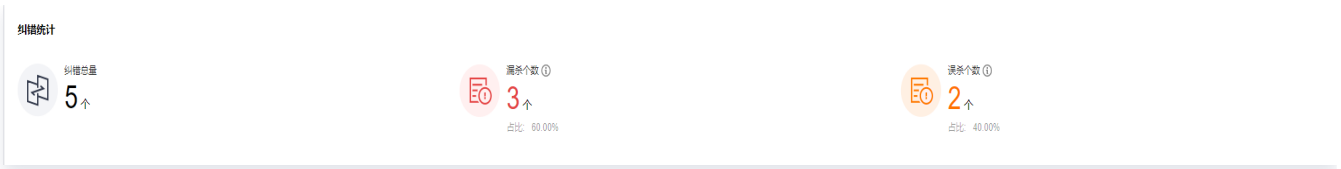
使用指引

1. 登录 内容安全控制台，在左侧导航栏中，单击**数据统计 > 纠错统计**。

2. 根据业务需求，筛选字段应用、场景、服务类型（文本）、统计周期后查看统计数据。

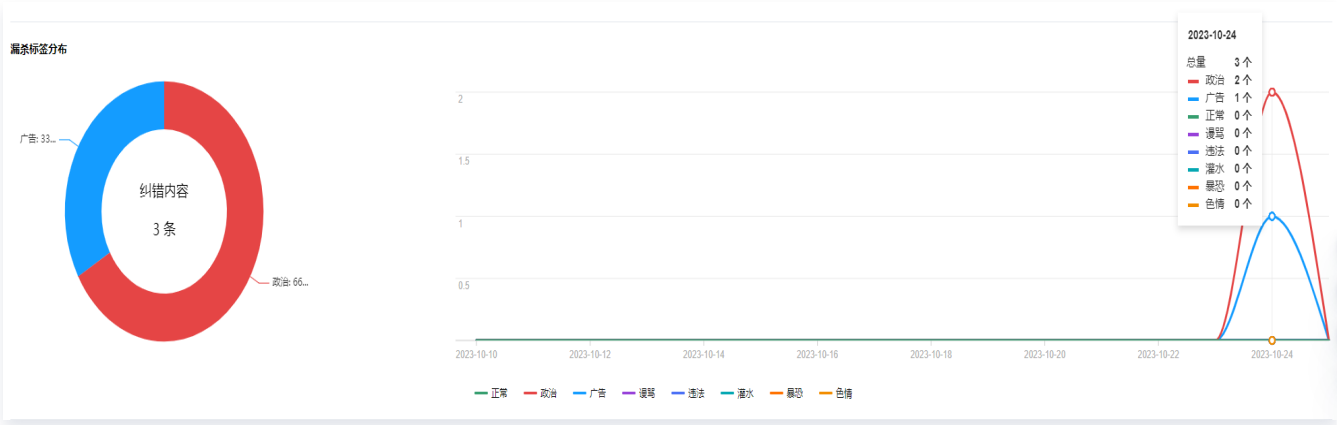


3. 根据业务需求筛选对应字段后，即可查看纠错统计数据。
   - **纠错总量**：进行了纠错操作的案例总量统计
   - **漏杀个数**：机审模型审核识别为“正常”，用户纠正其为风险标签的数据。
   - **误杀个数**：机审识别为“风险标签”,客户纠正为“正常”或“其他风险标签”。

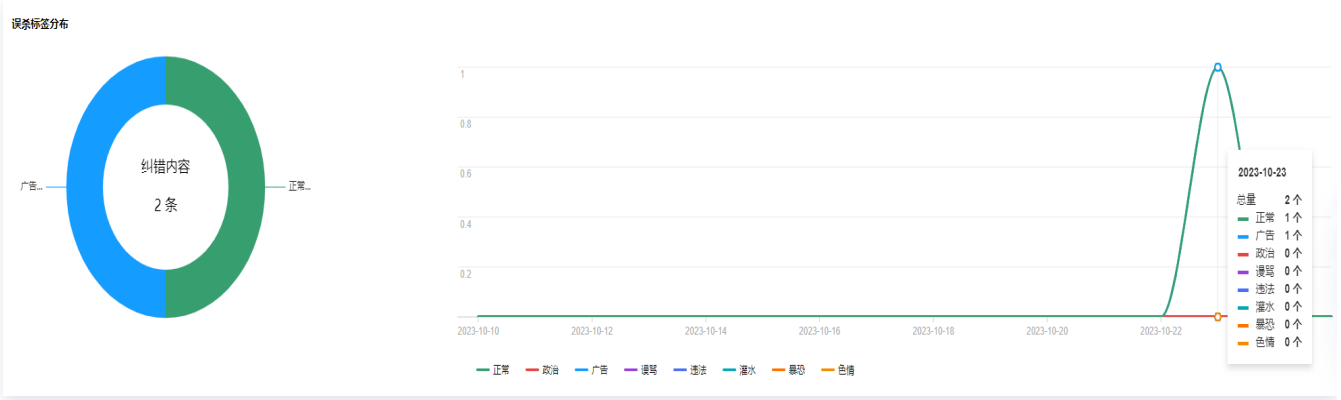


- **纠错趋势：**漏杀数和误杀数在不同时间节点的变化趋势。

- **漏杀标签分布：**不同的漏杀标签数量占比以及不同时间节点变化趋势。



- **误杀标签分布：**不同的误杀标签数量占比以及不同时间节点变化趋势。



审核结果纠错

最近更新时间：2025-02-05 16:02:22

功能说明

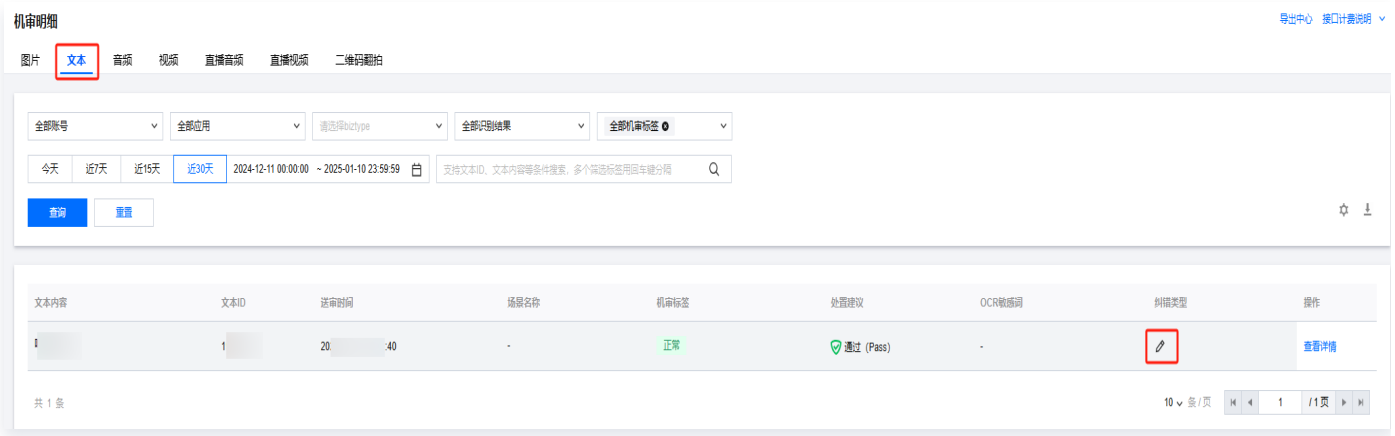
用户可以使用纠错功能对天御文本内容安全机审模型的审核结果进行修改，以保障业务需求。
本文将介绍如何进行审核结果纠错操作。

操作步骤

1. 登录 内容安全控制台 ，在左侧导航栏中，单击**明细查询** > **机审明细**。

> **说明：**
> 子账号使用审核结果纠错如未被主账号授予权限，需要按照 子账号授权指引 完成ListUsers权限及CMS策略权限授权后，才可使用审核结果纠错功能。

2. 在机审明细的文本页面，找到需要纠错案例，单击**纠错类型**列的 。



3. 在纠错选项页面，查看当前文本机审类型，并选择您认为正确的纠错类型，进行纠错操作。

> **注意：**
> - 纠错操作仅限一次，纠错成功后“纠错类型”将会显示为您所选对应标签。
> - 纠错操作包含：纠错结果回调、加入纠错名单。两类操作可全选或二选一，**不可不选**。

参数名称	参数说明
纠错结果回调	首次进行纠错操作，需配置回调地址，后续进行纠错操作将展示上次填写的回调地址。更多详情请参见 纠错回调数据结构 。 注意：回调地址可支持修改。
加入纠错名单	加入纠错名单后，md5值完全相同的数据，再次送审时将会按照纠错后的标签输出(加入名单后预计10分钟内生效)。
名单生效范围	支持选择生效至不同的策略 Biztype。

4. 填写完成，单击**确定**完成纠错。

纠错回调数据结构

```
{
    "RequestId": "123",    // 机审请求ID
    "DataId": "456",       // 机审数据ID
    "ContentType": "image",   // 送审文件类型：text或image
    "Content": "http://xxx.xxx.xxx",     // 送审原始内容
    "Suggestion": "Block",    // 处理建议
    "Label": "Porn"        // 纠错标签
}
```

纠错名单

最近更新时间：2025-02-05 16:02:22

功能说明

用户可以使用 纠错功能 对天御文本内容安全机审模型的审核结果进行修改，以保障业务需求。
本文中将介绍如何查看已加入纠错名单的案例。

操作步骤

1. 登录 内容安全控制台，在左侧导航栏中，单击**名单管理 > 纠错名单**。
2. 在纠错名单的文本页面，支持查看已加入纠错名单的文本案例。

> **说明：**
> 可按照账号维度筛选并查看不同账号所操作的纠错案例。

字段名称	字段说明
文本内容	已完成纠错操作的文本内容。
文本 ID	已完成纠错操作的文本请求 ID。
生效类型	支持选择生效至不同的策略 Biztype。
识别类型	机审模型原始审核结果标签。
纠错类型	业务方所选择的结果标签。
操作	删除，业务方可将已加入纠错操作的案例进行删除。 如进行了删除操作，送审相同md5文本，机审模型不再按照纠错结果输出，将按照机审模型原始审核结果输出。

版权所有：腾讯云计算（北京）有限责任公司