阿里云
内容安全

用户指南

文档版本：20190920

法律声明

阿里云提醒您在阅读或使用本文档之前仔细阅读、充分理解本法律声明各条款的内容。如果您阅读或使用本文档，您的阅读或使用行为将被视为对本声明全部内容的认可。

1. 您应当通过阿里云网站或阿里云提供的其他授权通道下载、获取本文档，且仅能用于自身的合法合规的业务活动。本文档的内容视为阿里云的保密信息，您应当严格遵守保密义务；未经阿里云事先书面同意，您不得向任何第三方披露本手册内容或提供给任何第三方使用。
2. 未经阿里云事先书面许可，任何单位、公司或个人不得擅自摘抄、翻译、复制本文档内容的部分或全部，不得以任何方式或途径进行传播和宣传。
3. 由于产品版本升级、调整或其他原因，本文档内容有可能变更。阿里云保留在没有任何通知或者提示下对本文档的内容进行修改的权利，并在阿里云授权通道中不时发布更新后的用户文档。您应当实时关注用户文档的版本变更并通过阿里云授权渠道下载、获取最新版的用户文档。
4. 本文档仅作为用户使用阿里云产品及服务的参考性指引，阿里云以产品及服务的”现状“、“有缺陷”和“当前功能”的状态提供本文档。阿里云在现有技术的基础上尽最大努力提供相应的介绍及操作指引，但阿里云在此明确声明对本文档内容的准确性、完整性、适用性、可靠性等不作任何明示或暗示的保证。任何单位、公司或个人因为下载、使用或信赖本文档而发生任何差错或经济损失的，阿里云不承担任何法律责任。在任何情况下，阿里云均不对任何间接性、后果性、惩戒性、偶然性、特殊性或刑罚性的损害，包括用户使用或信赖本文档而遭受的利润损失，承担责任（即使阿里云已被告知该等损失的可能性）。
5. 阿里云网站上所有内容，包括但不限于著作、产品、图片、档案、资讯、资料、网站架构、网站画面的安排、网页设计，均由阿里云和/或其关联公司依法拥有其知识产权，包括但不限于商标权、专利权、著作权、商业秘密等。非经阿里云和/或其关联公司书面同意，任何人不得擅自使用、修改、复制、公开传播、改变、散布、发行或公开发表阿里云网站、产品程序或内容。此外，未经阿里云事先书面同意，任何人不得为了任何营销、广告、促销或其他目的使用、公布或复制阿里云的名称（包括但不限于单独为或以组合形式包含”阿里云”、Aliyun”、“万网”等阿里云和/或其关联公司品牌，上述品牌的附属标志及图案或任何类似公司名称、商号、商标、产品或服务名称、域名、图案标示、标志、标识或通过特定描述使第三方能够识别阿里云和/或其关联公司）。
6. 如若发现本文档存在任何错误，请与阿里云取得直接联系。

通用约定

格式	说明	样例
	该类警示信息将导致系统重大变更甚至故障，或者导致人身伤害等结果。	 **禁止:** 重置操作将丢失用户配置数据。
	该类警示信息可能导致系统重大变更甚至故障，或者导致人身伤害等结果。	 **警告:** 重启操作将导致业务中断，恢复业务所需时间约10分钟。
	用于补充说明、最佳实践、窍门等，不是用户必须了解的内容。	 **说明:** 您也可以通过按Ctrl + A选中全部文件。
>	多级菜单递进。	设置 > 网络 > 设置网络类型
粗体	表示按键、菜单、页面名称等UI元素。	单击 确定。
`courier` 字体	命令。	执行 `cd /d C:/windows` 命令，进入Windows系统文件夹。
*## *	表示参数、变量。	`bae log list --instanceid` *Instance_ID*
[]或者[a|b]	表示可选项，至多选择一个。	`ipconfig` *[-all|-t]*
{}或者{a|b}	表示必选项，至多选择一个。	`swich` *{stand | slave}*

目录

1 OSS违规检测

1.1 扫描设置

OSS违规检测使用人工智能技术帮助您智能检测存储在阿里云对象存储服务OSS（Object Storage Service）中的图片、视频是否包含有色情、涉政等风险内容，并支持自动冻结检测出的违规内容。您可以通过内容安全控制台抽查机器审核的结果。

背景信息

- OSS违规检测只向开通了阿里云OSS的用户提供服务。使用OSS违规检测前，您需要在内容安全控制台通过RAM授权内容安全服务读取OSS Bucket的权限。

  说明:
  您在首次登录内容安全控制台并访问OSS违规检测页面时，可以一键完成该操作。

- OSS违规检测支持检测OSS Bucket中的增量内容和存量内容。
  - 增量内容（图片、视频）支持自动检测。增量检测一次配置即可长期生效。
  - 存量内容（图片、视频）不支持自动检测。如果要对存量图片、视频进行检测，您需要手动设置要检测的Bucket和时间范围，并执行存量扫描。
- OSS违规检测支持检测特定格式后缀的文件。
  - 支持检测的图片格式如下：jpg、png、jpeg、gif、bmp
  - 支持检测的视频格式如下：avi、mp4、3gp、mkv、mpg、mpeg、ts、rmvb、wmv、flv、mov
- OSS违规检测支持扫描以下endpoint中的OSS Bucket：
  - oss-cn-hangzhou.aliyuncs.com（杭州）
  - oss-cn-shanghai.aliyuncs.com（上海）
  - oss-cn-qingdao.aliyuncs.com（青岛）
  - oss-cn-beijing.aliyuncs.com（北京）
  - oss-cn-shenzhen.aliyuncs.com（深圳）
  - oss-cn-zhangjiakou.aliyuncs.com（张家口）
  - oss-cn-huhehaote.aliyuncs.com（呼和浩特）

使用授权

OSS违规检测只向开通了阿里云OSS的用户提供服务，在使用OSS违规检测前请确保您已开通OSS。如果您还未开通OSS，请前往OSS控制台开通OSS。

> 说明:
> 如果您开启了OSS防盗链，则您必须在OSS控制台将`https://yundunnext.console.aliyun.com`增加至Refer白名单。

您只有通过RAM授权云盾内容安全读取OSS Bucket的权限后，才能使用OSS违规检测。授权步骤如下：

1. 登录云盾内容安全控制台。
2. 在左侧导航栏，单击OSS违规检测。
3. 根据页面提示，单击授权，授权云盾内容安全检测您的OSS图片。

> 说明:
> 如果您已授权内容安全检测 OSS，则此处没有提示。

4. 在云资源访问授权页面，单击同意授权，完成访问授权。

设置增量扫描

增量扫描只需配置一次即可，配置完成后，每当您的Bucket中有新增的内容，系统会自动对其进行检测。

1. 登录云盾内容安全控制台。
2. 前往设置 > OSS违规检测页面，并打开增量扫描页签。

3. 完成下表中描述的增量扫描配置。

功能	配置	说明
-	Bucket设置	从左侧待选择框中勾选需要检测的Bucket，将其添加到右侧的已选择框中。 **说明:** · 左侧待选择框中罗列了当前阿里云账号在OSS中的可检测的Bucket，上面的数字表示可选择的Bucket的数量（例如下图中的“15”）。 · 右侧已选择框中罗列了已选择的Bucket，上面的数字表示已选择的Bucket的数量（例如下图中的“9”）。  单击已选择的Bucket右侧的漏斗图标设置过滤规则，仅扫描指定前缀（或目录下）的文件。例如，添加`img/test_`，表示只扫描该Bucket中以*`img/test_`*为前缀的图片件。  如果要扫描的图片文件在特定目录下，您可以在文件名前加上目录路径，以整体作为前缀。例如，您要扫描的文件在*`img/201805`*目录下，且前缀为*`test_`*，您可以添加`img/201805/test_`；如果您只想

功能	配置	说明
-	冻结方式	对于检测出来的违规文件，选择一种冻结方式： · 修改文件访问权限：将您的Bucket中public权限的违规文件设置为private访问权限。 · 移动违规文件：将您的Bucket中违规的文件移动到您Bucket中的备份目录下，并删除原路径下的文件。请慎重选择。
图片检测	检测类型	勾选涉黄、涉政。
	自动冻结	开启自动冻结后，当检测分值高于指定分值时，自动冻结图片。开启后，需要分别设置涉黄自动冻结阈值和涉政自动冻结阈值。 说明: 请慎重修改冻结阈值。默认冻结分数是100分，正常情况下不建议设置到99分以下，分数过低可能会导致正常图片被冻结，尤其是开启存量扫描时。冻结后的图片前台不可访问，帮助您防止风险外露。对于被冻结的照片，您可在OSS违规检测页面将其删除或者解冻。
	高级设置	设置每日扫描量上限，单位是万张。 说明: 默认没有上限。如果设置了扫描上限，扫描数量超出限制后将会停止扫描，存在巨大违规图片外露的风险。常规情况下，不建议您设置扫描上限。
视频检测	检测类型	勾选涉黄、涉政。
	视频截帧设置	· 截帧频率：每多少秒截取一帧。最慢60秒1帧，最块1秒1帧。截帧频率越高，识别准确率越高。 · 单视频最大截帧数：单个视频的最大截帧数量，取值范围是100～9999。 · 视频大小上限：检测的视频大小上限，取值范围是1～500MB，单个视频超过该上限不会被扫描。

功能	配置	说明
	自动冻结	勾选涉黄视频、涉政视频。开启后，违规视频将被自动冻结。

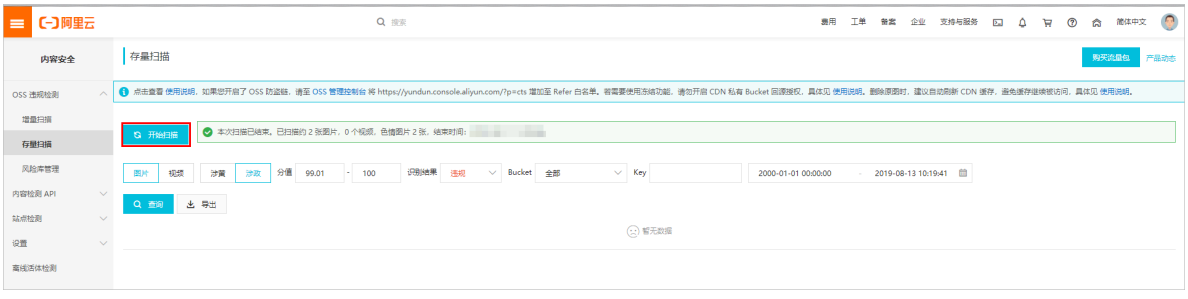
4. 勾选我已同意《OSS 违规检测服务条款》，并单击保存。

   保存后的设置即时生效。系统会按照配置，自动对已选择的Bucket进行图片/视频检测。

执行存量扫描

存量扫描为一次性的扫描动作，需要您手动配置并启动扫描。

1. 登录云盾内容安全控制台。
2. 前往OSS违规检测 > 存量扫描页面。
3. 单击开始扫描，并在提示对话框中单击确认。

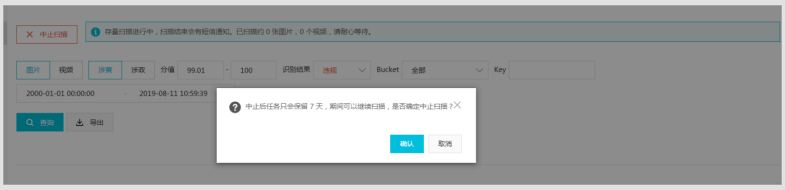


4. 在存量扫描设置侧边页完成存量扫描配置。配置内容与增量扫描一致，具体请参见增量扫描的配置说明。


存量扫描设置
Bucket 设置
待选择
已选择
移动全部
移动全部
时间范围
2000-01-01 00:00:00
2019-08-13 10:19:41
文件类型
图片（ 元 / 千张 ）
视频（ 元 / 千张 ）
检测类型
涉黄
涉政
冻结方式
修改文件访问权限
移动违规文件
视频截帧设置
截帧频率
1
单视频最大截帧数
100
视频大小上限
100
MB
图片自动冻结
当检测分值高于指定分值时自动冻结图片
开启自动冻结
涉黄自动冻结阈值
99.01
涉政自动冻结阈值
99.01
视频自动冻结
开启后，违规视频将被自动冻结
涉黄视频
涉政视频
图片 元 / 千张，视频帧 元 / 千张，涉黄与涉政检测分开收费，支持 流量包 优惠，具体可查看 详细价格
注：截帧数 = min[下取整(视频时长/截帧频率)+1, 最大截帧数]
开始扫描
取消

5. 单击开始扫描，并等待扫描完成。扫描结束后您会收到短信通知。

> **说明:**
> 在存量扫描过程中，您可以随时中止扫描。手动中止扫描后，扫描任务及配置会为您保留7天。您可以在合适的时间选择继续扫描。



1.2 自助审核

OSS内容检测任务执行完成后，您可以在控制台查询检测结果，并根据检测结果执行人工抽检和审核。本文以增量检测结果页面为例介绍了结果查询和自助审核的具体操作，存量检测结果的相关操作与之类似。

操作步骤

1. 登录内容安全控制台。
2. 前往OSS违规检测 > 增量扫描页面，并打开扫描结果页签。

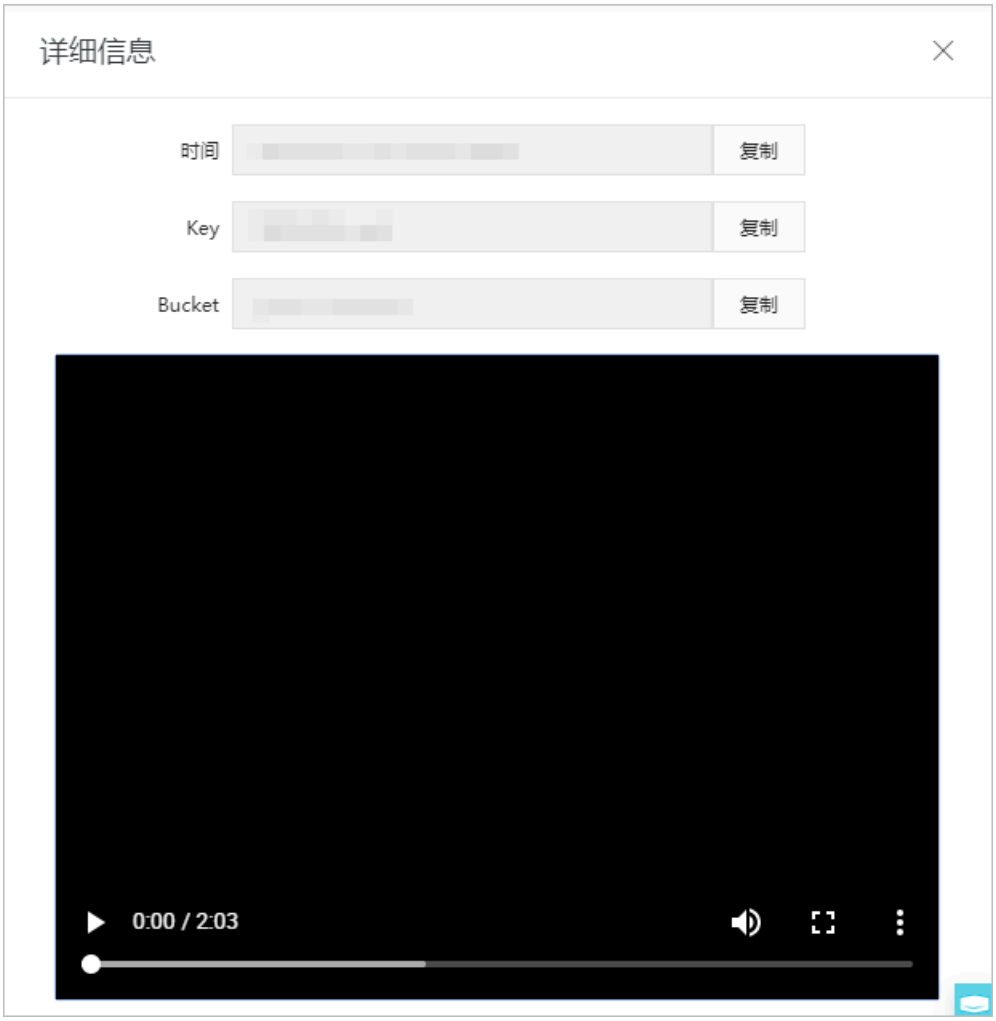
3. 设置查询条件，查询指定结果。

您可以通过以下筛选项过滤结果数据，直接查看您希望看到的结果：图片/视频、涉黄/涉政、分值范围、识别结果、Bucket、Key值、上传的时间范围。其中，

- 分值范围：每一张图片下方会有违规检测服务检测出的风险分值，分值范围是0~100，分值越高，代表文件违规的概率越大。
- 识别结果：违规、疑似、正常。
- Bucket：筛选扫描的Bucket。
- Key：Bucket中文件（或文件夹）的完整路径。

4. 检测内容的下方会显示与该内容对应的操作项。您可以进行以下操作：

- 单击违规并删除，可将图片/视频从内容安全控制台和OSS Bucket中一并删除。
- 单击正常并忽略，则忽略该检测结果。忽略后该图片/视频将不再在控制台展示，并不影响存储在OSS Bucket中的图片/视频。
- 选中一个或多个图片/视频后，单击违规并删除可将选中的图片/视频全部删除；单击正常并忽略可将选中的图片/视频全部忽略。
- 若您设置了自动冻结功能，则还可以在选中图片/视频后单击正常并解冻，将已冻结的图片/视频解冻。
- 单击图片/视频，可以查看其详细信息，包括文件创建时间、Key值、所在Bucket。



1.3 自定义图库

OSS违规检测的风险库通过收集自助审核的回流信息，搭建系统审核白名单、黑名单，为相同文件的检测提供依据。您也可以搭建自定义风险库用于风险防控。本文介绍了使用自定义风险图库的具体操作。

背景信息

目前，风险库管理只支持图库管理。OSS违规检测的图片结果经自助审核后自动回流到系统图库（即系统黑白名单）。审核违规的图片回流到对应检测场景（鉴黄、暴恐、广告）的系统黑名单，审核正常的图片回流到对应检测场景（鉴黄、暴恐、广告）的系统白名单。下图显示了对应检测场景的系统黑名单和白名单。

名称	Code	使用场景	识别结果	数量	最近修改时间	BizType	状态	操作
系统 AD_FEEDBACK_BLACK		广告	黑名单			--	已启用	管理
系统 PORN_FEEDBACK_BLACK		鉴黄	黑名单			--	已启用	管理
系统 ILLEGAL_FEEDBACK_BLACK		暴恐	黑名单			--	已启用	管理
系统 PORN_FEEDBACK_WHITE		鉴黄	白名单			--	已启用	管理
系统 AD_FEEDBACK_WHITE		广告	白名单			--	已启用	管理
系统 ILLEGAL_FEEDBACK_WHITE		暴恐	白名单			--	已启用	管理

系统黑白名单在下次检测时生效。系统黑名单中的图片在下次检测时被判为违规；系统白名单中的图片在下次检测时被判为正常。您可以管理系统黑白名单的文件，如添加或删除图片，用于风险防控。修改后的黑白名单在15分钟内生效。

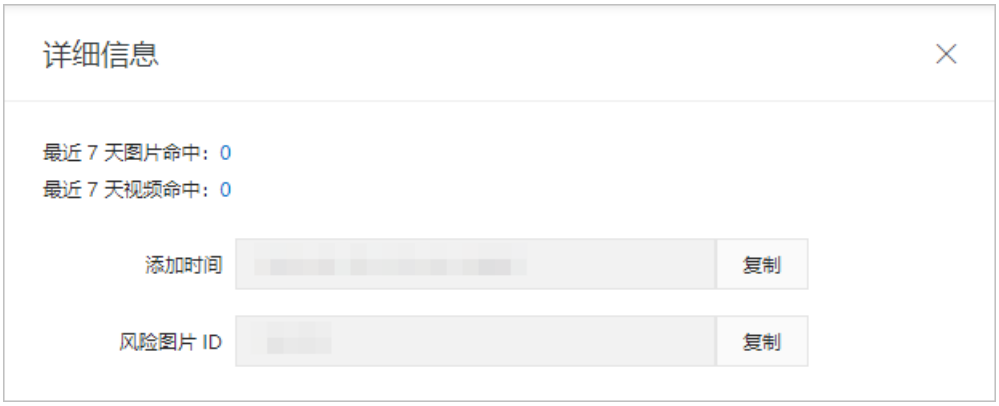
除了系统黑白名单外，您还可以创建最多10个自定义风险图库。自定义图库支持创建对应检测场景（鉴黄、暴恐、广告）的疑似名单，引导OSS违规检测对名单中图片判定为疑似。

管理系统图库

1. 登录云盾内容安全控制台。
2. 前往OSS违规检测 > 风险库管理页面。
3. 在自定义图库页签下，定位到要操作的系统图库，单击其操作列下的管理。

4. 在图库管理页面，根据需要执行以下操作：

   - 根据风险图片ID、添加时间范围查询图片。
   - 单击图片，打开图片的详细信息页面，查看近期命中数量、添加时间、风险图片ID等信息。



   - 单击选择文件，上传图片到图库。

> **说明：**
> 支持上传.png、.jpg、.jepg、.bmp格式的图片文件。单个图片文件的大小需小于5M。一次最多上传20张图片。

   - 单击图片下的删除，删除图片；选中多张图片后，单击批量删除，批量删除图片。

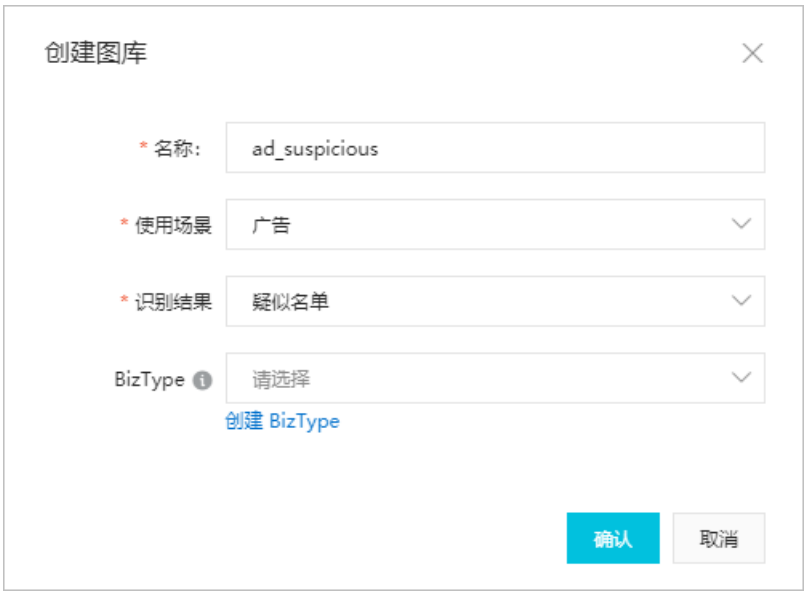
创建自定义图库

1. 登录云盾内容安全控制台。
2. 前往OSS违规检测 > 风险库管理页面。
3. 在自定义图库页签下，单击创建图库。

4. 在创建图库对话框中，完成以下配置，并单击确认。

配置项	说明
名称	为图库命名。
使用场景	选择图库的使用场景，可选值： · 鉴黄 · 暴恐 · 广告
识别结果	选择图库的用途，可选值： · 黑名单 · 疑似名单 · 白名单

配置项	说明
BizType	选择图库应用的业务场景。 **说明:** 只有没有Biztype属性的风险图库会在OSS违规检测场景中生效。



成功创建图库。您可以在图库列表中看到新建的图库。新建图库即刻生效。您可以在图库列表中对新建图库执行以下操作：

- 停用/启用图库
- 修改图库信息
- 删除图库
- 管理图库内容

5. 定位到新建的图库，单击其操作列下的管理。

6. 在图库管理页面，单击选择文件，在图库中添加图片。

1.4 回调通知

回调通知指内容安全服务通过异步消息通知的方式向您发送机器内容检测或者您自助审核的结果。OSS违规检测和内容检测API均支持回调通知。本文介绍了配置OSS违规检测回调通知的方法。

背景信息

在使用回调通知功能前，请通过下表了解相关概念。

名称	说明
扫描结果回调消息	OSS违规检测完成后，由内容安全服务端将机器扫描的结果以POST请求的方式，发送到您设置的HTTP回调通知地址。
审核结果回调消息	您通过云盾内容安全控制台或者以调用接口的方式对机器检测的结果进行修改后，由内容安全服务端将审核结果以POST请求的方式，发送到您设置的HTTP回调通知地址。
Seed	Seed值用于校验发送到您设置的HTTP回调通知地址的请求是否来自内容安全服务端。
回调地址	回调地址是您在内容安全控制台配置的服务端地址，通常是您自己的业务服务器的公网地址。回调地址需要满足以下要求： · 应为HTTP/HTTPS协议接口的公网可访问的URL。 · 支持POST方法。 · 支持传输数据编码采用UTF-8。 · 支持数据接收格式为`application/x-www-form-urlencoded`。 · 支持表单参数`checksum`和`content`。
回调次数	您的服务端接收到内容安全推送的回调结果后，若返回的HTTP状态码为200，表示接收成功；若返回其他的HTTP状态码，均视为您接收失败，我们将重复推送最多16次。
回调数据	回调数据是内容安全服务端向您设置的回调通知地址返回的数据内容。回调数据的结构描述见下表。

表 1-1: 回调通知表单数据

名称	类型	描述
checksum	String	由<用户uid> + <seed> + <content>拼成字符串，通过SHA256算法生产。用户UID即阿里云账号ID，您可在阿里云控制台上查询。 说明: 为防篡改，您可以在获取到推送结果时，按此算法生成字符串，与checksum做一次校验。
content	String	字符串格式保存的JSON对象，请自行解析反转成JSON对象。根据不同的检测功能（OSS违规检测、内容检测API），content解析成JSON后的格式有所差异，详见下文Content说明。

扫描&审核回调通知设置

1. 登录云盾内容安全控制台。
2. 前往设置 > OSS违规检测页面，并打开消息通知页签。
3. 在OSS回调通知下，单击设置。

说明:
如果您设置过回调地址，则此处显示查看，您可以单击查看。

4. 在OSS回调通知设置侧边页，完成以下配置，并单击设置。

配置项	说明
回调地址	填入回调通知地址。

配置项	说明
生效模块	勾选要应用回调设置的功能模块，取值： · 增量扫描 · 存量扫描
开启审核回调	是否开启审核回调。
开启扫描回调	是否开启扫描回调。
扫描回调类型	开启扫描回调后，选择对哪种类型的扫描结果进行回调通知，取值： · 扫描结果违规 · 扫描结果疑似 · 扫描结果正常

设置完成后，系统会自动生成一个Seed。Seed值用于校验您的回调接口收到的请求来自阿里云。请保存自动生成的Seed，并在调用相关接口时根据需要传入该参数。

回调数据说明

启用回调通知后，内容安全将按照回调配置发送OSS违规检测的回调通知。回调通知中包含content表单数据。下表描述了content表单字段的结构。

表 1-2: content表单字段结构描述

名称	类型	是否必须	说明
bucket	字符串	是	OSS bucket的名称。
object	字符串	是	OSS文件名。
stock	布尔	是	是否是存量对象，true表示是存量，false表示是增量。
region	字符串	是	OSS文件所在地域。
freezed	布尔	是	对象是否被冻结，true表示被冻结，false表示未被冻结。

名称	类型	是否必须	说明
scanResult	JSON对象	否	扫描结果。根据不同的检测对象（图片、视频），结构有差异。 · 针对图片对象，结构同图片同步检测中的results返回参数。 · 针对视频对象，结构同视频异步检测中的results返回参数。
auditResult	JSON对象	否	审核结果。审核操作时才会有该字段，具体结构描述见表 1-3: auditResult。 说明: 如果只推送扫描结果，则没有该字段。

表 1-3: auditResult

名称	类型	是否必须	说明
suggestion	字符串	是	审核时给的建议值。取值： · block：审核时设置违规 · pass：审核时设置正常
resourceStatus	整型	是	审核后，object的状态。取值： · 0：已删除 · 1：已冻结 · 2：可用可访问

content示例

```
{
    "bucket": "xxxxb",
    "freezed": false,
    "object": "xxx.jpg",
    "region": "cn-hangzhou",
    "scanResult": {
        "code": 200,
        "dataId": "5C919E2FBD6CD6940D4A8B46",
        "msg": "OK",
        "results": [
            {
                "label": "porn",
                "rate": 100.0,
                "scene": "porn",
                "suggestion": "block"
            },
            {
                "label": "normal",
                "rate": 99.6,
```

```
                "scene": "terrorism",
                "suggestion": "pass"
            }
        ],
        "taskId": "img1ZxzVQUuAz36scZEjyGXzx-1qpzhY"
    },
    "auditResult": {
        "resoureStatus": 2,
        "suggestion": "pass"
    },
    "stock": false
}
```

2 内容检测API

2.1 自定义机审标准

内容安全采用阿里云默认的机器审核标准为您提供内容检测服务。如果您在测试过程中发现默认的审核标准相对您的业务需求过于严格或宽松，您可以通过内容安全的审核标准模板搭建并应用自定义机审标准。本文介绍了搭建自定义机审标准的具体方法。

背景信息

自定义审核标准模板目前只支持配置图片色情、图片涉政暴恐两个场景的机审标准。

在自定义机审标准前，请先熟悉以下概念：

- bizType（或BizType）：业务场景。审核标准基于bizType搭建，每个bizType对应一套审核标准；未配置自定义审核标准时，统一使用默认的bizType以及对应的默认审核标准。配置自定义bizType后，您必须在内容检测API的接口中传递自定义bizType，检测才会按照自定义bizType的标准进行。
- 准确率：机审判定违规且人审确认违规的检测数量 / 机审判定违规的检测数量。
- 召回率：机审判定违规且人审确认违规的检测数量 / 人审确认违规的检测数量。
- 审核比：机审判定疑似的检测数量 / 机审结果的总数。

操作步骤

1. 登录云盾内容安全控制台。
2. 前往设置 > 内容检测API页面，并打开机审标准页签。

3. 单击创建BizType图标（首次创建bizType）或者新增BizType（已创建过bizType）。

4. 在新增BizType对话框中，完成bizType配置，并单击确认。bizType的配置描述见下表。

配置项	说明
BizType名称	为bizType命名。支持使用数字、英文字符、下划线（_），且不超过32个字符。
行业分类	业务所属行业分类，非必选项。若传入行业分类，我们能够更好地帮助您调整策略配置。
从现有导入	如果您已经创建过bizType，则可以选择直接导入已创建的bizType的配置。

成功新建bizType，您可以在左侧bizType列表中看到新建的bizType。

5. 单击新建的bizType，编辑其审核标准。

   不同检测场景的审核标准定义不同，具体以控制台页面显示为准。 目前仅支持配置图片色情、图片涉政暴恐检测的审核标准。审核标准定义的描述见下表。

检测对象	检测场景	审核标准	说明
图片	色情	选择一种审核策略。支持的审核策略包括： · 高召回率策略 · 高准确率策略 · 严格色情忽略性感策略	在调用接口检测图片涉黄风险时（参见#unique_13），请求参数scene需要传递porn，标准才会生效。
图片	涉政暴恐	选择一种或多种管控场景。支持的管控场景包括： · 涉政人物 · 特殊标识 · 特殊着装 · 枪支 · 利器 · 军警徽 · 钱币 · 血腥场景	在调用接口检测图片涉政暴恐风险时（参见#unique_13），请求参数scene需要传递terrorism，标准才会生效。

6. 修改配置后单击保存。

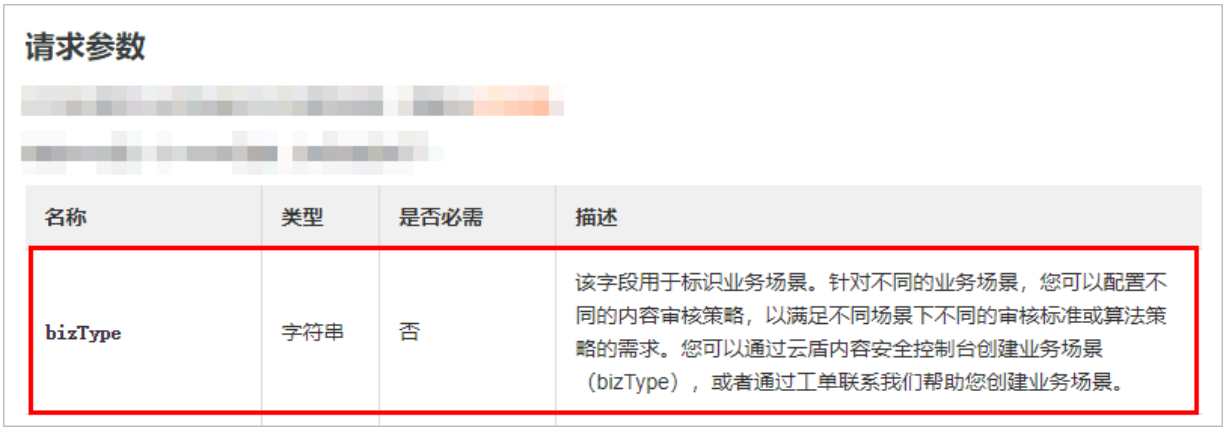
预期结果

完成自定义审核标准的定义。

后续步骤

接下来，您只需将要应用的审核标准的bizType名称告知开发人员；开发人员在使用开发文档进行服务接入时，传入对应的bizType参数即可。这样，机审过程将采用bizType配置的自定义审核标准。

以图片审核接口为例，若您在请求参数bizType中传入您在控制台创建的bizType的名称，检测就会按照自定义机审标准进行。



请求参数

名称	类型	是否必需	描述
bizType	字符串	否	该字段用于标识业务场景。针对不同的业务场景，您可以配置不同的内容审核策略，以满足不同场景下不同的审核标准或算法策略的需求。您可以通过云盾内容安全控制台创建业务场景（bizType），或者通过工单联系我们帮助您创建业务场景。

2.2 自定义文本库

自定义文本库允许您将已知的风险或安全文本内容添加到黑名单或白名单，并在您调用内容检测API进行图片广告检测（ad）、文本反垃圾（antispam）、语音反垃圾（antispam）时，自动匹配文本库中的内容，满足突发性或个性化的管控需求，达到紧急止血的目的。本文介绍了使用自定义文本库的具体操作。

背景信息

自定义文本库支持两种文本类型：关键词和相似文本。您可以创建关键词文本库，用来管理关键词黑名单、忽略名单；或者创建相似文本文本库，用来管理相似文本黑名单、白名单及疑似名单。调用检测服务中，系统会根据黑名单、忽略名单、白名单及疑似名单的命中情况，返回相应的suggestion。

说明:
如果您不清楚如何使用该功能，请通过工单咨询我们。不建议您随意添加关键词，因为可能导致误抓，使检测效果无法得到保障。

进行具体操作前，请先熟悉以下概念：

- 关键词

  关键词是针对短小词语进行防控的一种方式。您可以将其理解为：一句话或者一段文本里面是否包含某个既定词语；当包含该词语时，则表明命中该关键词。不同的业务场景支持配置不同的关键词。

  在内容安全的识别中，关键词技术可以被应用到图片广告、文本反垃圾、语音反垃圾场景中，具体配置见对应场景中的使用描述（配置参数可能略有出入）。

- 相似文本

  相似文本是针对句子或者段落式文本进行相似性判断的一种方式。您可以将其理解为：两句话或者两段文本，从句意上具有非常强的相似性，但又不是百分百一样；局部可能有变化，但是整体却具有相同的意思或者在描述同一件事情。通过既定或者参照的文本样本，可以判断要识别的文本是否与样本具有强相似性。当相似性的概率在一定程度上时，则表明命中样本。

  相似文本文本库适用于文本反垃圾的检测场景。通过定义自己业务的相似文本库黑名单、白名单、疑似名单（疑似名单是指业务上需要识别出来，且需要人工审核），并在相似文本库里面维护与您业务相关的文本样本，从而指导文本反垃圾识别去过滤命中相似文本样本的内容。

使用限制

类型	项目	限制
文本库	库个数	不超过10个。
文本库	库名长度	不超过20个字符。
关键词	关键词类型	仅支持中文关键词，支持用字母和数字作为关键词；暂不支持英文关键词。 说明: 检测时，字母和数字会被当作整体进行分词。
关键词	单个文本库中关键词个数	不超过10000个。
关键词	关键词最大长度	50个字符（包括符号）。
关键词	中文关键词编码类型	UTF-8
关键词	关键词格式	不允许包含以下特殊字符（包括全角）：@ # $ % ^ * ( ) < > / ?, . ; _ + - = ‘ “ 空格 tab键
相似文本	相似文本长度	10~4000个字符。 说明: 如果添加的文本过长，容易引起文本误抓。建议文本长度不要超过200个字符，具体情况可提工单咨询。
相似文本	单个文本库中相似文本个数	不超过10000个。
相似文本	文本编码格式	UTF-8

类型	项目	限制
相似文本	相似文本内容	文本样本需要包含明确的可提取的中文语义特征。如果经过引擎分析特征数太少，该文本样本将不会生效，引擎将其直接忽略。 说明: 如果一段样本都是无意义的字母数字，或各种表情符等，则可能被忽略。

关键词高级特性

中文关键词支持“与（&）”、“非（~）”的逻辑判断属性。例如：

- 定义“你&我”，则只有在句子中同时出现“你”和“我”时，才会命中。
- 定义“你~我”，则只有在句子中只出现“你”且不出现“我”时才会命中，同时出现“你”和“我”则不会命中。

说明:
与（&）必须在非（~）之前。例如，您可以设置“你好&再见~他们”作为关键词，但不能设置“你好~他们&再见”作为关键词。

操作步骤

1. 登录云盾内容安全控制台。
2. 前往内容检测API > 风险库管理页面。
3. 打开自定义文本库页签，并单击创建文本库。

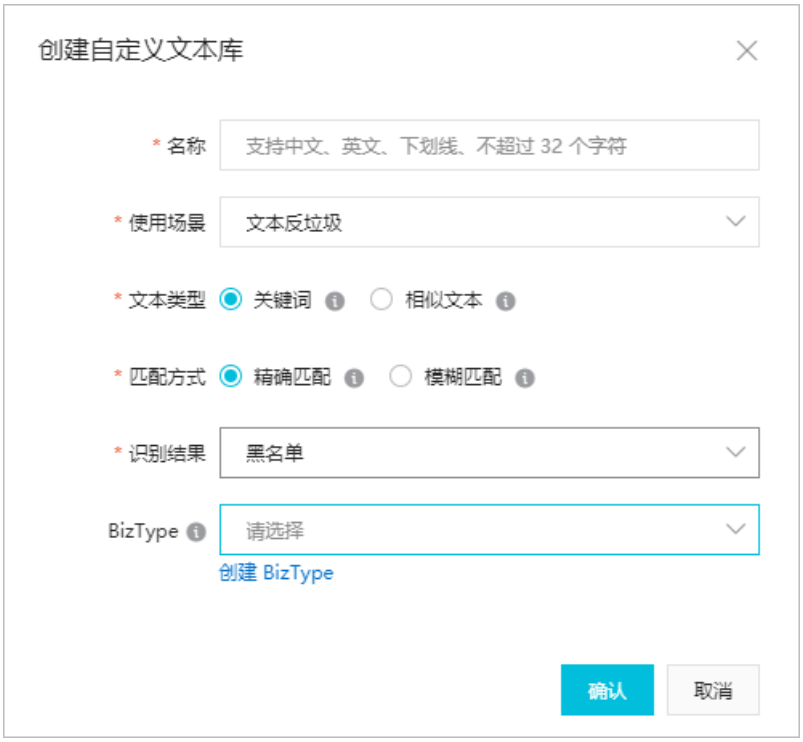
4. 在创建自定义文本库对话框中，完成文本库配置，并单击确认。文本库的配置描述见下表。

配置项	描述
名称	为文本库命名。文本库名称允许重复，但建议您在业务中将其设置为唯一。

配置项	描述
使用场景	选择文本库的使用场景，取值： · 文本反垃圾 · 语音反垃圾 · 图片广告 说明： 该场景参数对应于API调用时通过scenes传入的参数。例如，假如该文本库适用于文本反垃圾，则使用场景选择文本反垃圾；那么，在您调用文本反垃圾或文件反垃圾检测时，都会默认匹配该文本库。 文本库支持启用/停用操作，如果文本库被停用，则文本库在检测时不会被使用。
文本类型	选择文本库的文本类型，取值： · 关键词：使用关键词匹配，只要包含关键词就会命中，覆盖面大。 · 相似文本：使用文本相似度匹配，只有整段文本相似才会命中，精确度高。
匹配方式	文本类型为关键词时，选择文本库的匹配方式，取值： · 精确匹配：待检测文本中包括与库中的词完全一样的内容时才命中。 · 模糊匹配：待检测文本以及关键词都会经过预处理，预处理后进行匹配。预处理的逻辑如下： - 字母大写统一转换为小写。例如，输入检测文本“bitCoin”，会命中关键词“bitcoin”。 - 繁体中文统一转换为简体。例如，输入检测文本“中國”，会命中关键词“中国”。 - 相似字转换。例如，输入检测文本“②”，会命中关键词“2”。

配置项	描述
识别结果	选择命中后的处理方式。 · 文本类型为关键词时，取值： - 黑名单：当检测文本命中黑名单中的样本时，API检测请求返回的suggestion为block（拒绝）。 - 忽略名单：当检测文本中包含了忽略名单中的样本时，该关键词即会被替换为空字符串，然后再进行检测（并非直接返回suggestion为pass）。 · 文本类型为相似文本时，取值： - 黑名单：当检测文本与黑名单中的文本内容相似时，API检测请求返回的suggestion为block（拒绝）。 - 疑似名单：当检测文本与疑似名单中的文本内容相似时，API检测请求返回的suggestion为review（审核）。 - 白名单：当检测的文本与白名单中的文本内容相似时，API检测请求返回的suggestion为pass（通过）。

配置项	描述
BizType	BizType属于高级功能，目的是能够根据不同的业务需求配置不同的文本库，请根据需要进行设置（建议通过工单联系我们指导配置）。BizType生效逻辑如下： · 文本库设置BizType为“A”，且API检测请求中传递了biztype为“A”，则检测文本只会使用biztype为“A”的文本库（前提是库已开启）。 · 其他情况下，检测文本均会使用所有已开启的文本库。



成功创建文本库。您可以在文本库列表中看到新建的文本库。

5. （可选）若新建的文本库的文本类型是关键词，参照以下步骤管理关键词。

   a) 定位到目标（关键词）文本库，单击其操作列下的管理。

   b) 在文本库管理页面，维护文本库内的关键词：

   · 单击新增关键词或导入，按照页面提示在文本库中增加关键词。

   > 说明：
   > 已添加的关键词，您可以查看其最近7天命中次数（不包括当天的命中数据）。

   · 勾选不需要的关键词，单击批量删除，删除关键词；也可以单击不需要的关键词下的删除，单独将其删除。

   在文本库新增、删除关键词后，系统大约在10分钟左右生效。

6. （可选） 若新建的文本库的文本类型是相似文本，参照以下步骤管理相似文本。

   a) 定位到目标（相似文本）文本库，单击其操作列下的管理。

   b) 在文本库管理页面，维护文本库内的相似文本：

   - 单击新增文本或导入，按照页面提示在文本库中增加相似文本。

     说明:
     已添加的相似文本，您可以查看其最近7天命中次数（不包括当天的命中数据）。

   - 勾选不需要的相似文本，单击批量删除，删除相似文本；也可以单击不需要的相似文本下的删除，单独将其删除。

   在文本库新增、删除相似文本后，系统大约在1分钟左右生效。

7. 启用与停用文本库。回到文本库列表，选择对应文本库，单击其操作列下的启用或停用，可以根据实际需求启用或者停用文本库。
8. 删除与修改文本库。回到文本库列表，选择对应文本库，单击其操作列下的删除和修改，可以分别删除目标文本库和修改目标文本库的配置。

自定义文本库API与SDK

我们同时提供了操作自定义词库（关键词库/相似文本库）的API接口与部分语言的SDK，供您直接将文本库管理功能集成到自己的业务平台中。具体包含以下接口：

- 获取文本反垃圾关键词库列表：DescribeKeywordLib
- 创建文本反垃圾关键词库：CreateKeywordLib
- 修改文本反垃圾关键词库：UpdateKeywordLib
- 删除文本反垃圾关键词库：DeleteKeywordLib
- 查找文本反垃圾关键词：DescribeKeyword
- 添加文本反垃圾关键词：CreateKeyword
- 删除文本反垃圾关键词：DeleteKeyword

2.3 自定义图库

使用自定义图库，您可以为具体的检测场景定义要直接拦截或放行的图片。本文介绍了在云盾内容安全控制台管理自定义图库的具体操作。

背景信息

自定义图库分为黑名单、白名单、疑似名单三种类型，您可以将已知的要拦截的、要直接放行的、要人工审核的图片样本添加到相应图库中，系统在检测的时候会匹配对应场景的图库，应对突发性

的管控需求。自定义图库针对具体的检测场景生效，目前仅适用于智能鉴黄、暴恐涉政、图片广告三种场景。

在使用自定义图库时，您需要创建一个图库，指定其类型和适用的检测场景，然后管理其中的图片样本。自定义图库生效后，当您调用具体场景的检测服务时，

- 若系统检测到黑名单库中的样本或类似样本，则suggestion返回block。
- 若系统检测到白名单库中的样本或相似样本，则suggestion 返回pass。
- 若系统检测到疑似名单库中的样本或相似样本，则suggestion 返回review。

除了在控制台管理自定义图库外，您还可以通过API接口或SDK完成相关操作：

- 使用API管理自定义图库
- 使用Java SDK管理自定义图库

操作步骤

1. 登录云盾内容安全控制台。
2. 前往内容检测API > 风险库管理页面。
3. 打开自定义图库页签，并单击创建图库。

4. 在创建图库对话框中，完成图库配置，并单击确认。图库的配置描述见下表。

配置项	说明
名称	根据您的业务为图库命名。建议您设置可读性较强的中文名称，且不超过64个字符。

配置项	说明
使用场景	图库的使用场景，取值： · 鉴黄 · 暴恐 · 广告 说明： 该场景值对应API调用时的scenes参数。假设，该图库适用于鉴黄场景，则使用场景选择鉴黄；那么，在调用图片或者视频鉴黄服务时，默认启用该场景的图库。
识别结果	图库的类型，取值： · 黑名单 · 疑似名单 · 白名单
BizType	BizType属于高级功能，目的是能够根据不同的业务需求配置不同的图库，请根据需要进行设置（建议通过工单联系我们指导配置）。BizType生效逻辑如下： · 图库设置BizType为“A”，检测时API中传递了BizType为“A”，则检测只会使用biztype为“A”的图库（前提是库已开启）。 · 其他情况下，检测均会使用所有已开启的所有图库。

成功创建图库，已创建图库默认启用。

5. 管理图库。

a) 回到图库列表中，选择目标图库，单击其操作列下的管理。

b) 在图库管理页面，维护图库内的图片：

- 单击选择文件，上传本地图片到图库。

> 说明：
>
> - 支持上传png、jpg、jpeg、bmp格式的图片文件，且单个文件需要小于50M，一次最多上传20张图片。
> - 一个图库下最多可以添加10000张样本图片。

- 勾选不需要的图片，单击批量删除，删除图片。

新增、删除图片样本后，约需3分钟才会生效。

6. 删除、修改、停用图库。回到图库列表中，选择目标图库，单击其操作列下的删除、修改、停用可以分别删除目标图库、修改目标图库的配置、停用目标图库。

2.4 自助审核

内容安全控制台中呈现了内容检测API检测出的数据结果。针对您的业务场景，您可以对机器的检测结果进行二次人工审核；人工审核后，下次同样的检测内容识别出的结果会与您设置的结果保持一致。本文介绍了使用自助审核的具体操作。

背景信息

- 自助审核默认只展示机器审核结果为疑似（review）或者违规（block）的数据。如需展示机审结果正常（pass）的数据，请在控制台上进行设置，具体请参见设置入审数据类型。
- 图像、视频、语音、文本均可以进行人工审核，但只有图像、文本的自助审核结果会自动回流入风险样本库。
- 机器的检测数据只保留最多7天，请及时处理。

操作步骤

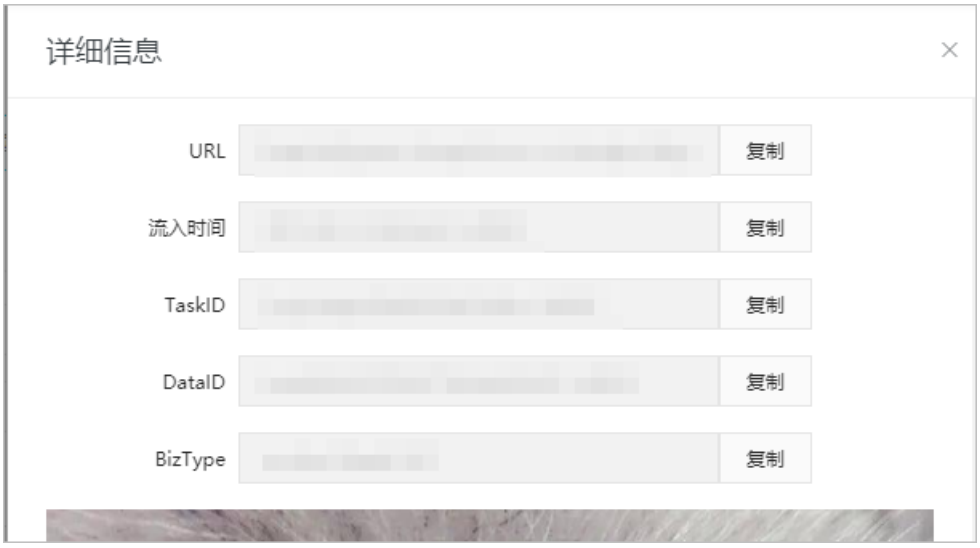
1. 登录云盾内容安全控制台。
2. 前往内容检测API > 自助审核页面。
3. 通过页签选择要审核的内容的类型（图片、视频、文本、语音），单击进入审核页面。
4. 按照以下方式进行标记。

说明:

仅以图片标记为例，其他类型内容的标记方法与之类似。

- **对于您认为正常，却被识别为违规（block）或者疑似（review）的样本，将其标记为正常。**
- **对于您认为需要管控，却被识别为疑似（review）或者正常（pass）的样本，将其标记为违规，并选择违规原因：涉黄、暴恐涉政、图文广告、不良场景。**

- **支持勾选多张图片后进行批量处理，如批量标记正常、批量标记违规。**
- **单击样本图片，查看其详细信息。**



被标记样本以及类似样本的检测结果将会按照您的标记结果实时纠正，且自动回流入对应的样本库中。

如下图所示带有红色系统标识的图库即为系统回流图库（文本库与之类似）。

设置入审数据类型

1. 登录云盾内容安全控制台。
2. 前往设置 > 内容检测API页面。
3. 打开入审数据页签，并勾选流入审核页面的数据类型：
   - 确认违规数据（默认勾选）
   - 疑似违规数据（默认勾选）
   - 正常数据

修改设置后，即时生效。

2.5 回调通知

回调通知指内容安全服务通过异步消息通知的方式向您发送机器内容检测或者人工审核的结果。OSS违规检测和内容检测API均支持回调通知。本文介绍了配置内容检测API回调通知的具体操作。

背景信息

在使用回调通知功能前，请通过下表了解相关概念。

名称	说明
扫描结果回调消息	内容检测API完成您的检测请求后，由内容安全服务端将机器扫描的结果以POST请求的方式，发送到您设置的HTTP回调通知地址。
审核结果回调消息	您通过云盾内容安全控制台或者以调用接口的方式对机器检测的结果进行修改后，由内容安全服务端将审核结果以POST请求的方式，发送到您设置的HTTP回调通知地址。
Seed	Seed值用于校验发送到您设置的HTTP回调通知地址的请求是否来自内容安全服务端。
回调地址	回调地址是您在内容安全控制台配置的服务端地址，通常是您自己的业务服务器的公网地址。回调地址需要满足以下要求： · 应为HTTP/HTTPS协议接口的公网可访问的URL。 · 支持POST方法。 · 支持传输数据编码采用UTF-8。 · 支持数据接收格式为`application/x-www-form-urlencoded`。 · 支持表单参数`checksum`和`content`。
回调次数	您的服务端接收到内容安全推送的回调结果后，若返回的HTTP状态码为200，表示接收成功；若返回其他的HTTP状态码，均视为您接收失败，我们将重复推送最多16次。
回调数据	回调数据是内容安全服务端向您设置的回调通知地址返回的数据内容。回调数据的结构描述见下表。

表 2-1: 回调通知表单数据

名称	类型	描述
`checksum`	String	由`<用户uid> + <seed> + <content>`拼成字符串，通过SHA256算法生产。用户UID即阿里云账号ID，您可在阿里云控制台上查询。  说明: 为防篡改，您可以在获取到推送结果时，按此算法生成字符串，与`checksum`做一次校验。
`content`	String	字符串格式保存的JSON对象，请自行解析反转成JSON对象。根据不同的检测功能（OSS违规检测、内容检测API），`content`解析成JSON后的格式有所差异，详见下文Content说明。

扫描结果回调设置

1. 自行准备好接收请求的HTTP回调地址以及Seed参数。

2. 调用内容检测异步API接口时，传递相应的callback和seed请求参数，具体请参见API接口描述中的参数说明。

请求参数

关于在请求中必须包含的公共请求参数，请参考公共参数。

请求body是一个JSON对象，字段说明如下：

名称	类型	是否必需	描述
callback	字符串	否	异步检测结果回调通知您的URL，支持HTTP/HTTPS。该字段为空时，您必须定时检索检测结果。
seed	字符串	否	随机字符串，该值用于回调通知请求中的签名。当使用callback时，该字段必须提供。
tasks	JSON数组	是	检测对象，JSON数组中的每个元素是一个图片检测任务结构体（image表）。每个元素的具体结构描述见task。

审核结果回调设置

1. 登录云盾内容安全控制台。
2. 前往设置 > 内容检测API页面，并打开消息通知页签。
3. 在API审核回调通知下，单击设置。

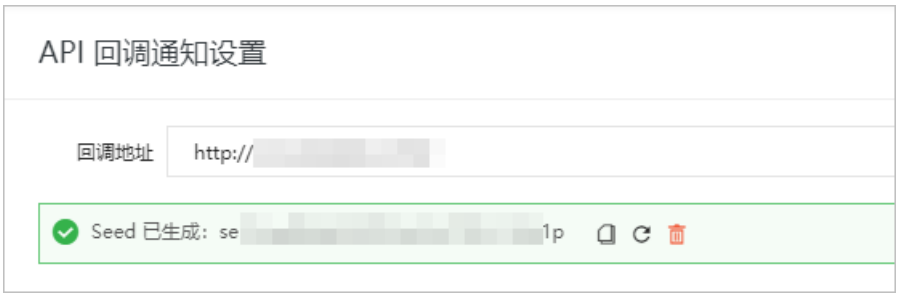
说明:

如果您设置过回调地址，则此处显示查看，您可以单击查看。

4. 在API回调通知设置侧边页，完成以下配置，并单击设置。

配置项	说明
API回调通知设置	填入回调通知地址。

设置完成后，系统会自动生成一个Seed。Seed值用于校验您的回调接口收到的请求来自阿里云。请保存自动生成的Seed，并在调用相关接口时根据需要传入该参数。



回调数据说明

启用回调通知后，内容安全将按照回调配置发送内容检测API回调通知。回调通知中包含content表单数据。下表描述了content表单字段的结构。

表 2-2: content表单字段结构描述

名称	类型	是否必须	说明
scanResult	JSON对象	否	扫描结果。根据不同的检测对象（图片、视频），结构有差异。 · 针对图片对象，结构同图片同步检测中的results返回参数。 · 针对视频对象，结构同视频异步检测中的results返回参数。
auditResult	JSON对象	否	您在云盾控制台审核页面上的审核结果。审核操作时才会有该字段，具体结构描述见表 2-3: auditResult。 说明: 如果只推送扫描结果，则没有该字段。
humanAuditResult	JSON对象	否	阿里云的人工审核结果。如果您购买了阿里云的人工审核服务，人工审核后结果会在该字段中。具体结构描述见表 2-4: humanAuditResult。

表 2-3: auditResult

名称	类型	是否必须	说明
suggestion	字符串	是	审核时给的建议值。取值： · block：审核时设置违规 · pass：审核时设置正常
labels	JSON数组	是	人工审核给内容打的标签，取值可以为以下可选值中的一个或者多个： · porn：鉴黄 · terrorism：暴恐涉政 · ad：图文广告

表 2-4: humanAuditResult

名称	类型	是否必须	说明
suggestion	字符串	是	阿里云人工审核的建议，取值： · block：阿里云人工审核结果为违规 · pass：阿里云人工审核结果为正常
taskId	字符串	是	检测任务的ID。通过任务ID可以关联到对应内容的机器审核的结果。
dataId	字符串	是	您通过接口请求中传递的用于标识您检测内容的Id。

content示例

```
{
    "scanResult": {
        "code": 200,
        "msg": "OK",
        "taskId": "fdd25f95-4892-4d6b-aca9-7939bc6e9baa-1486198766695
",
        "url": "http://1.jpg",
        "results": [
            {
                "rate": 100,
                "scene": "porn",
                "suggestion": "block",
                "label": "porn"
            }
        ]
    },
    "auditReult": {
        "suggestion": "block",
        "labels": [
            "porn",
            "ad",
            "terrorism"
```

```
            ]
        },
        "humanAuditResult": {
            "suggestion": "pass",
            "dataId":"yyyy",
            "taskId": "xxxxxx"
         }
}
```

2.6 样本反馈

在使用内容检测API过程中，如果您发现内容安全的算法结果在您的业务中被认为是不符合预期的，您可以通过反馈接口将样本回流给我们。

如果您有自己的审核平台，您可以直接对接反馈接口，将审核后认为识别有误的样本回流给我们。在收到您的反馈后，我们会在下个版本的模型迭代中将您的反馈数据加入训练。训练后的模型对您的业务场景更具有适应性。

样本反馈目前支持图片样本反馈，视频样本反馈，和文本样本反馈，具体使用方式请参考文档说明。

自动加入自定义图库

模型训练需要积累足够的样本，因此可能无法立即生效。如果有需要，您可以开启自动加入自定义图库功能，实时纠正结果。

针对图片样本的反馈，您可以在内容安全控制台直接管理回流图像库。操作方法类似其他自定义图库，但是不支持创建与删除回流图像库。关于自定义图库的操作方法，请参考自定义图库。

在反馈接口的label字段中，

- 对于您认为正常的样本，传入normal，可以将该样本加入白名单。
- 对于您认为违规的样本，传入任意字段（建议您使用porn、ad、terrorism等风险字段），可以将该样本加入黑名单。

2.7 数据统计

您可以在内容安全控制台查看内容检测API的调用统计数据。

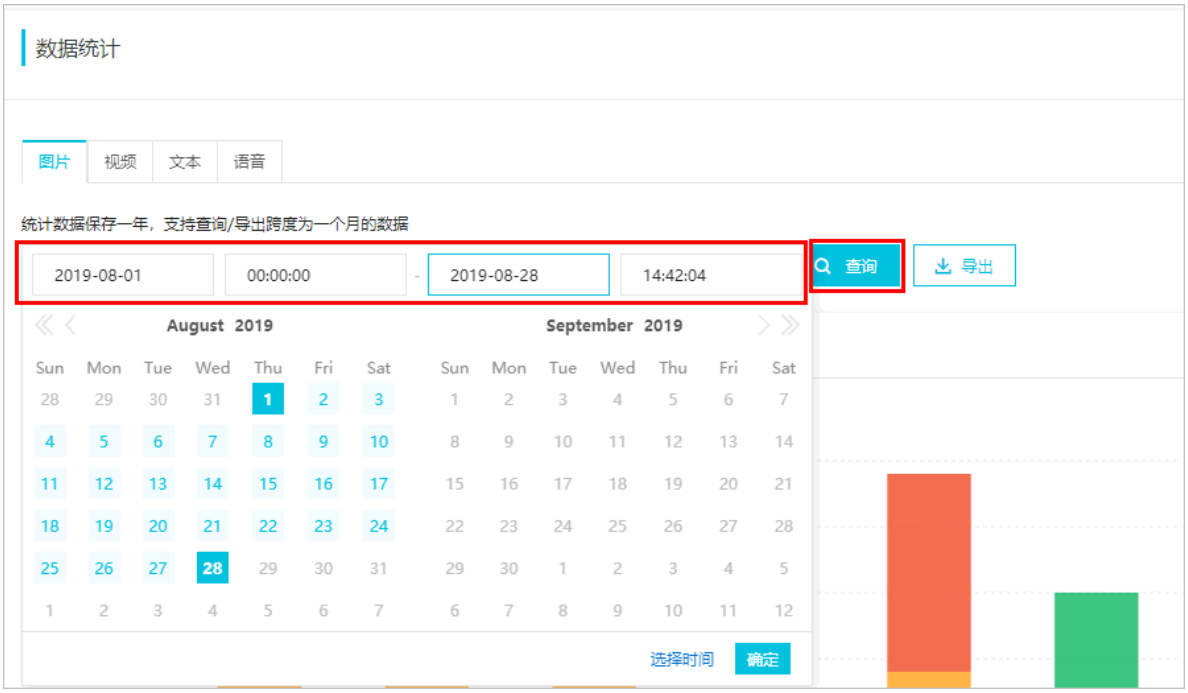
背景信息

内容安全控制台汇总了内容检测API的调用统计数据，支持查询最近1年内图片、视频、文本、语音检测接口的总调用次数，以及不同检测场景下检测结果（确认违规量、疑似违规量、正常量）的分布信息。

操作步骤

1. 登录云盾内容安全控制台。
2. 在左侧导航栏，单击内容检测API > 数据统计。
3. 通过页签选择要查询的检测接口类型：图片、视频、文本、语音。
4. 在数据统计页面，选择查询时间，并单击查询。

   支持查询的时间段为最近1年内。支持设置的时间跨度为1个月。



5. 分检测场景查看调用统计报表，并根据需要导出报表。

   - 报表说明
   - 导出说明

报表说明

以图片鉴黄调用量报表为例，报表展示了每日调用图片鉴黄检测接口（scene=porn）的次数，并统计了不同检测结果的数量和分布。图例说明如下：

- 确认违规量：返回suggestion=block的检测请求的数量。
- 疑似违规量：返回suggestion=review的检测请求的数量。
- 正常量：返回suggestion=pass的检测请求的数量。

不同检测对象的调用量单位不同，具体说明如下：

- 图片检测：调用量单位是图片的数量（张）。
- 视频检测：调用量单位有两种，一种是视频的截帧数（张），一种是视频的时长（分钟）。
- 文本检测：调用量单位是文本行数（条）。
- 语音检测：调用量单位是语音的时长（分钟）。

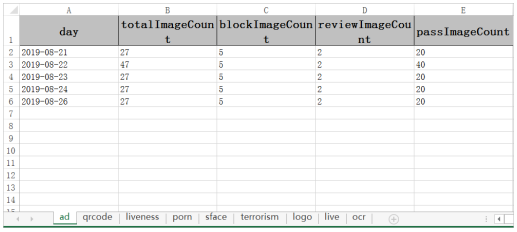
导出说明

导出的报表是Excel格式。导出数据的时间范围与您设置的查询条件一致。导出的Excel只包含有调用量的检测场景（对应API接口调用时传递的scene参数值），每个场景对应一张表单，表单中按天记录调用量。

表单中出现的行头字段说明见下表。

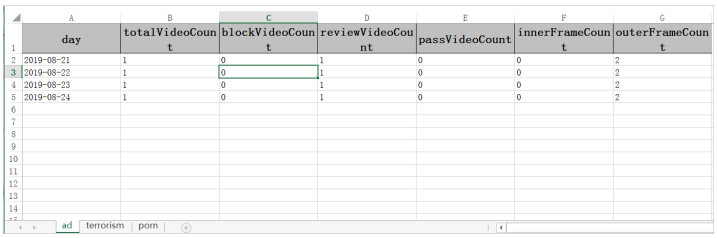
名称	含义	单位
day	调用日期	-
totalImageCount	检测图片总量	张
blockImageCount	违规图片量	张
reviewImageCount	疑似违规图片量	张
passImageCount	正常图片量	张
totalVideoCount	检测视频总量	个
blockVideoCount	违规视频量	个
reviewVideoCount	疑似违规视频量	个
passVideoCount	正常视频量	个
innerFrameCount	视频的系统截帧总量	张
outerFrameCount	视频的用户截帧总量	张
totalTextCount	检测文本总量	条
blockTextCount	违规文本量	条
reviewTextCount	疑似违规文本量	条

名称	含义	单位
passTextCount	正常文本量	条
totalVoiceDuration	检测语音总量	分钟
blockVoiceDuration	违规语音量	分钟
reviewVoiceDuration	违规语音量	分钟
passVoiceDuration	违规语音量	分钟



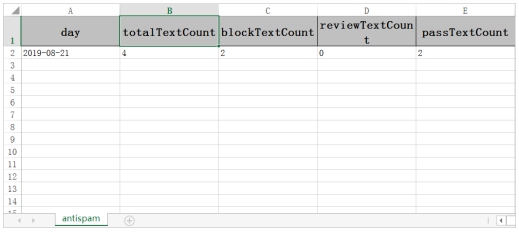
	A	B	C	D	E
1	day	totalImageCount	blockImageCount	reviewImageCount	passImageCount
2	2019-08-21	27	5	2	20
3	2019-08-22	47	5	2	40
4	2019-08-23	27	5	2	20
5	2019-08-24	27	5	2	20
6	2019-08-26	27	5	2	20

ad | qrcode | liveness | porn | sface | terrorism | logo | live | ocr



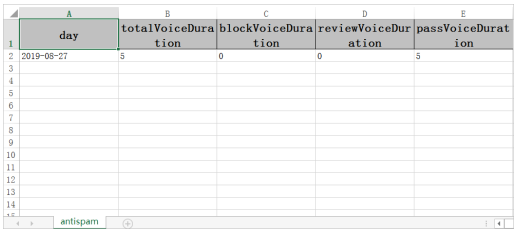
	A	B	C	D	E	F	G
1	day	totalVideoCount	blockVideoCount	reviewVideoCount	passVideoCount	innerFrameCount	outerFrameCount
2	2019-08-21	1	0	1	0	0	2
3	2019-08-22	1	0	1	0	0	2
4	2019-08-23	1	0	1	0	0	2
5	2019-08-24	1	0	1	0	0	2

ad | terrorism | porn



	A	B	C	D	E
1	day	totalTextCount	blockTextCount	reviewTextCount	passTextCount
2	2019-08-21	4	2	0	2

antispam



	A	B	C	D	E
1	day	totalVoiceDuration	blockVoiceDuration	reviewVoiceDuration	passVoiceDuration
2	2019-08-27	5	0	0	5

antispam

2.8 授权访问MTS服务

在提交视频内容检测任务时，如果您选择通过OSS地址（oss://xxxx）上传视频URL的方式，则云盾内容安全对上传的OSS视频自动截帧。内容安全调用阿里云媒体处理服务（MTS）进行视频截帧，避免公网访问用户数据，最大限度降低流量费用。您必须授权MTS服务以内容安全的身份递交视频截帧任务。该操作通过阿里云访问控制中的角色管理功能实现，本文介绍了您需要完成的步骤。

背景信息

您只能在提交视频异步检测任务时，选择上传视频URL的方式进行配置。

关于访问控制服务的角色功能，请参见角色。

通过完成操作步骤，您将实现以下目的：

- 在您的阿里云账号下创建MTS服务角色，并指定由内容安全的阿里云账号扮演使用该角色。
- 授权所创建的MTS服务角色只读访问您的OSS空间。
- 在通过OSS地址上传视频URL时，按照格式要求拼接生成URL并上传。

这样，内容安全的阿里云账号将扮演所创建的MTS服务角色，调用自身MTS服务，访问您的OSS空间，获取视频内容并对其截帧。

操作步骤

1. 创建RAM角色。

   a) 登录RAM控制台。

   b) 前往RAM角色管理页面，单击新建RAM角色。

   c) 选择可信实体类型为阿里云服务，并单击下一步。

   d) 配置角色名称，并选择受信服务为多媒体转码服务。

   e) 单击完成。

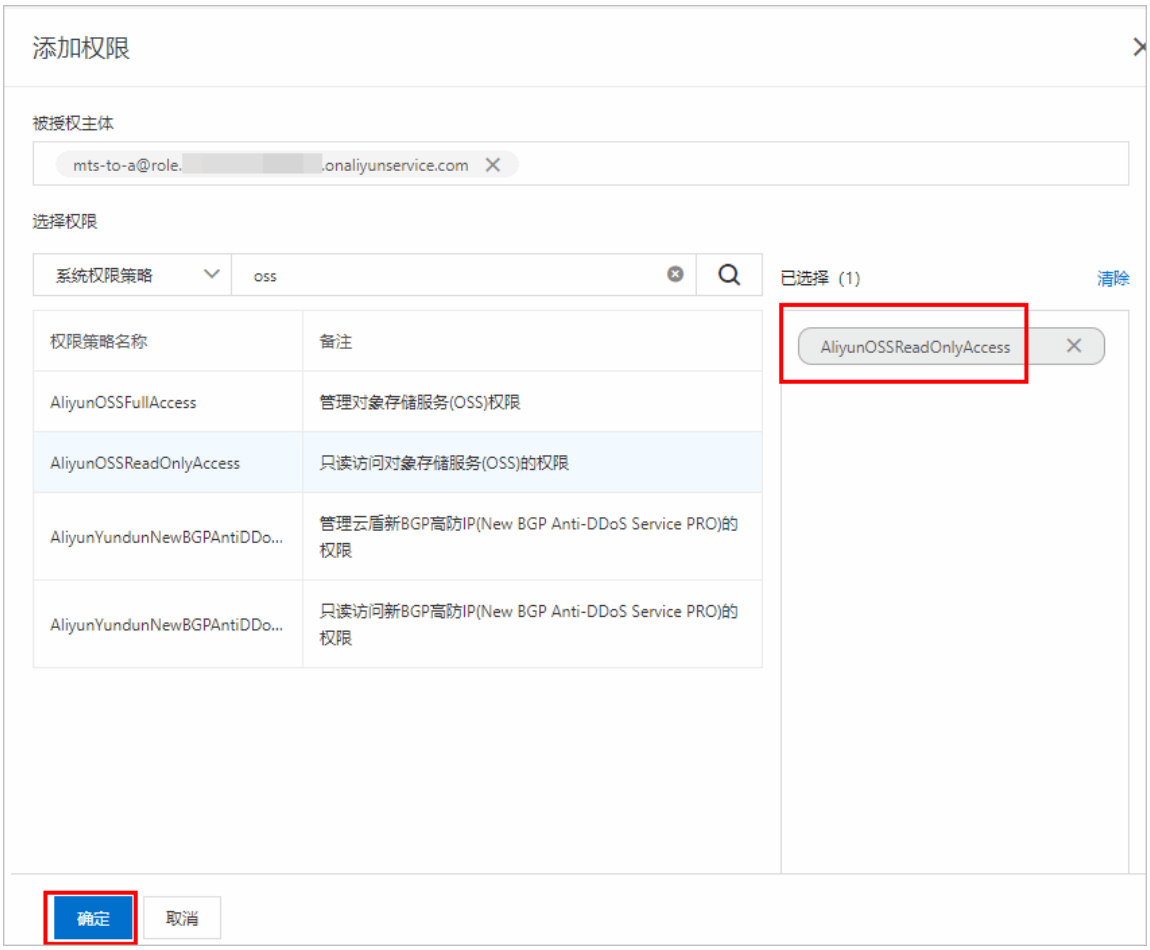
> **说明:**

该操作可能需要通过手机验证。

成功创建角色。

f) **单击为角色授权。**

g) **在添加权限页面，为角色授予AliyunOSSReadOnlyAccess系统权限策略，并单击确定。**



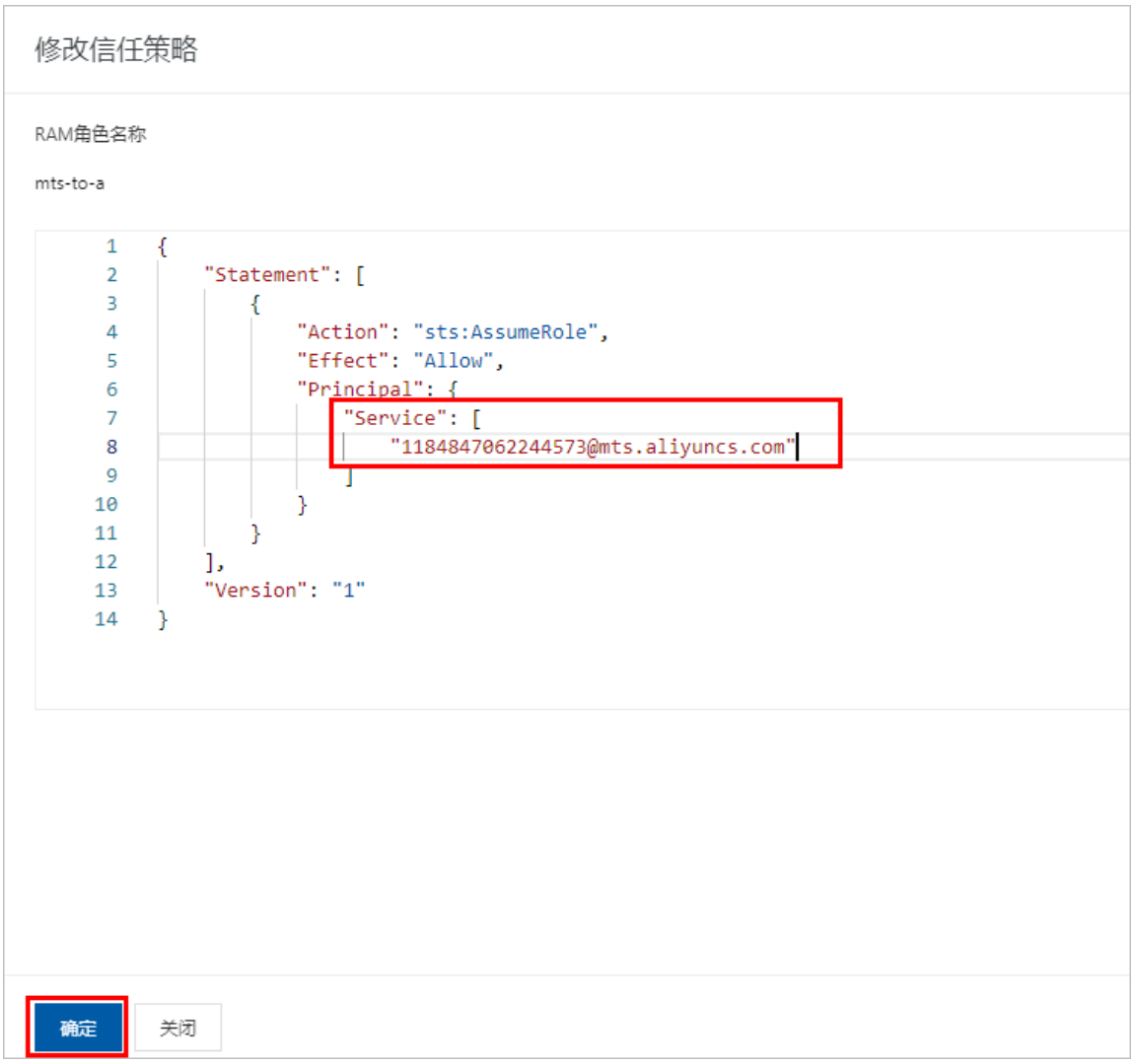
该操作授权服务角色以只读权限访问您的阿里云账号下的OSS内容。

2. 修改服务角色的信任策略。

a) 回到RAM角色管理页面，定位到新建的角色，单击其名称进入角色详情。

b) 打开信任策略管理页签，并单击修改信任策略。

c) 在修改信任策略页面，将"Service"下的内容修改为"1184847062244573@mts.aliyuncs.com"，并单击确定。



该操作指定由内容安全的阿里云账号（UID：1184847062244573）扮演所创建的服务角色，调用其MTS服务。

3. 回到角色详情，在基本信息下，查看并复制角色的ARN（Aliyun Resource Name，阿里云全局资源名称）。

4. 对要检测的OSS视频对象，按照以下格式拼接生成视频URL：`oss://arn@bucket.region/object`

   例如，假设您在深圳OSS的bucket *foo*上有视频对象*video/bar.mp4*需要检测，则拼接生成的URL为`oss://acs:ram::xxxxxxxxxxxxxxxx:role/mts-to-a@foo.cn-shenzhen/video/bar.mp4`（`xxxxxxxxxxxxxxxx`是您的16位阿里云ID。）

   > **说明:**
   > 目前支持的区域（region）包括：cn-hangzhou、cn-shanghai、cn-beijing、cn-shenzhen。

5. 提交视频检测任务时，上传拼接生成的URL作为检测对象。

2.9 自定义OCR模板

本文介绍如何创建并使用自定义OCR模板，根据需要识别的图片模板，对各种类型的票据、证件等图片进行文字识别。

功能描述

自定义OCR模板帮助您提取自定义图片中的结构化文字信息。使用过程中您需要先自定义一个图片模板，然后再调用OCR识别接口进行检测。

如果您需要识别的图片不在已有的结构化OCR支持范围内，您可以使用自定义OCR模板。

进行操作前，请先了解以下基本概念。

- 模板：为格式和包含信息完全相同的一类图片生成的一种规范版式。

  进行图片文字识别前，您需要在内容安全控制台手动创建模板。每个模板都有一个唯一的ID作为其标识；在调用OCR检测接口时，需要传入模板ID作为请求参数。

- 参考字段：用于定位模板位置的固定字段。

  参考字段的选取会影响图片的识别准确率。参考字段务必选取位置和内容都不会变化的文字内容。单个参考字段内的文字不可以换行，建议您选取4个以上的参考字段。
- 识别字段：需要识别的内容字段。

  设置识别字段时，需要给字段设置key值，最终识别结果会以key:识别内容格式返回。

操作步骤

参照以下步骤创建和使用自定义OCR模板：

1. 登录云盾内容安全控制台。
2. 在左侧导航栏单击设置。
3. 在OCR模板页签，单击创建模板按钮。

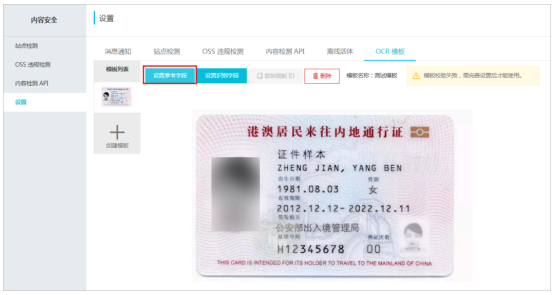
4. 在创建模板页面，输入模板名称，并单击选择文件，选择一张待识别的图片作为样本上传。

   关于作为模板的样本图片，请注意以下要求：

   - 使用.png、.jpg、.jepg、.bmp、.gif格式。
   - 大小在1KB到10M之间，分辨率在320*320像素到4096*4096像素之间。
   - 尽量摆放端正平整，不存在模糊、过度曝光、阴影等不良情况。
   - 尽量突出需要识别的部分。建议您手动剪裁掉不需要部分，以提高识别准确率。
   - 至少存在4个模板参照字段，且尽量分散在图片的边缘（越分散越好），用于准确定位模板。
   - 选取的模板参照字段、待识别字段的高度不小于20像素。

5. 设置参考字段。

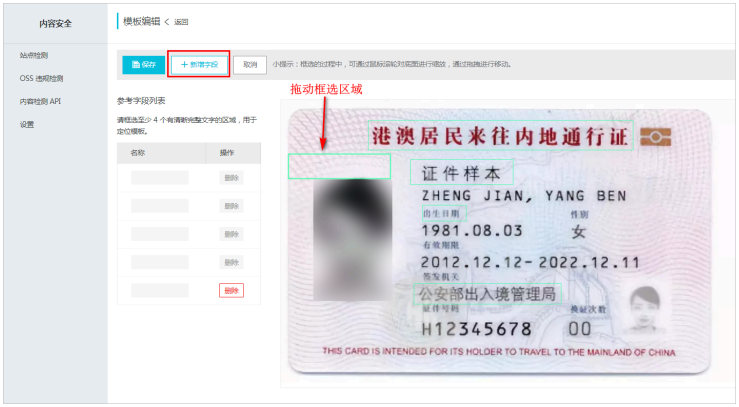
   a. 单击设置参考字段。



   b. 单击新增字段并用绿色矩形标识框框选图片上位置固定不变的单行参考字段。

   说明:

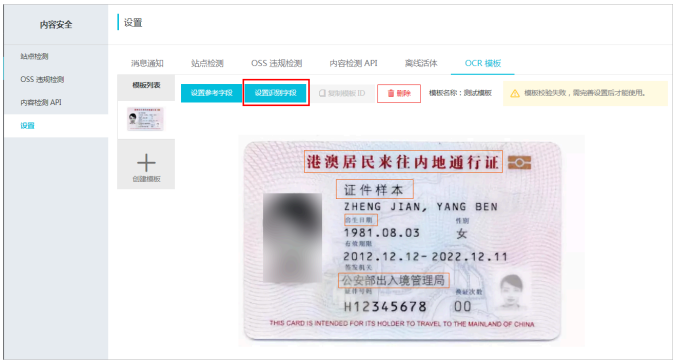
参照字段区域务必框选单行文字，且尽量将文字包裹完整。



c. 重复上述步骤，至少设置4个不同的参考文字区域后，单击保存。

6. 设置识别字段。

a. 单击设置识别字段。



b. 单击新增字段并用绿色矩形标识框框选待识别的单行文字，为框选中的内容设置一个Key值，作为识别结果的标识。

c. 重复上述步骤，添加完所有待识别内容后，单击保存。

说明:

如果要识别的字段有多行内容，建议您分别框选单行文字，并为它们设置相同的Key值。算法会将多行Key值相同的字段以框选顺序组合返回。

7. 完成模板创建后，单击选择要应用的模板，然后单击复制模板ID。

8. 参考OCR同步检测，调用检测接口进行图片OCR识别。

说明:

如果您对模板制作和OCR识别有任何疑问，请通过工单联系我们进行协助。

3 站点检测

3.1 启用站点检测

购买站点检测实例后，您可以将实例绑定到您的网站，为网站启用站点检测服务。

前提条件

已开通内容安全服务并购买站点检测实例。具体操作请参见#unique_40。

操作步骤

1. 登录云盾内容安全控制台。
2. 前往设置 > 站点检测页面，打开实例管理页签。
3. 选择一个有效的且处于未绑定状态的实例，单击其操作列下的绑定站点。

4. 在绑定站点对话框中，完成网站信息和检测频率配置，并单击下一步。配置说明如下。

配置	说明
协议	勾选HTTP或者HTTPS。 **说明:** 如果您的网站通过HTTP和HTTPS协议分别响应不同的内容，而且内容差异较大，建议您使用两个实例，分别绑定网站的HTTP协议和HTTPS协议。
域名	填写您站点的域名，填写时不要包含`http://`或者`https://`。如果您的网站有多个频道子域名，建议您在这里填写根域名。 假设您的网站有www.domain.com、news.domain.com、sports.domain.com等多个频道内容需要检测，建议您在这里输入domain.com。如果您只想检测news频道的内容news.domain.com，您可以输入news.domain.com。
默认首页地址	填写完整的站点首页网址。输入的网址必须在您要绑定的域名下。

配置	说明
首页检测间隔	设置每隔多少小时，访问您的网站首页进行一次检测。
全站检测频率	选择执行全站检测的频率，可选值：低：7天1次、高：1天1次。 **说明:** 站点检测频率越高，检测占用的带宽及产生的带宽费用也越多。如果您的网站内容比较多，且网站带宽不足的话，过高的检测频率可能影响您网站的正常访问速度。如果您不希望影响网站性能，建议您选择较低的检测频率。

5. 在验证站点对话框，选择一种验证方式，参照对话框中的验证说明完成相应操作，然后单击立即验证。通过站点验证，证明您对站点的管理权，防止未经授权的检测。如果您暂时不方便进行验证，您可以单击稍后验证，保存当前已输入的数据。

   支持的验证方式包括以下四种：

   - 阿里云账户验证：验证待检测站点（域名）是否在您当前登录的阿里云账号的资产下。

   - 主机文件验证：按要求在域名对应主机的根目录下生成相应的文件进行验证。

   - CNAME验证：按要求在待检测域名的解析记录中增加指定的CNAME记录进行验证。

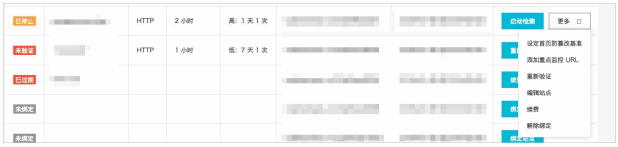
- **网站首页HTML标签验证：按要求修改网站首页HTML源文件进行验证。**

验证通过后，完成站点绑定和检测设置，目标实例自动开始检测。

6. （可选）回到实例管理页面，选择已启用的实例，在其操作列下，根据需要执行以下操作。
   - 暂停/启动检测：如果您不希望在当前时间执行检测，您可以暂停检测；已暂停的检测，通过启动检测可以恢复。
   - 设定首页防篡改基准、添加重点监控URL：具体请参见#unique_41。
   - 重新验证：如果您的验证失效或在步骤5中选择了稍后验证，您可以重新验证对站点的管理权。
   - 编辑站点：修改实例绑定的站点和检测频率信息。

     说明：
     如果您修改了站点或首页地址，需要重新验证。
   - 续费：为目标实例续费，可以延长其使用时长。
   - 解除绑定：如果您不希望继续向已绑定的站点提供检测服务，您可以解除绑定。

     说明：
     解除绑定后，已购买的实例不会释放，但是您可以将其绑定到别的站点，为别的站点提供检测服务。



后续步骤

查看检测结果。

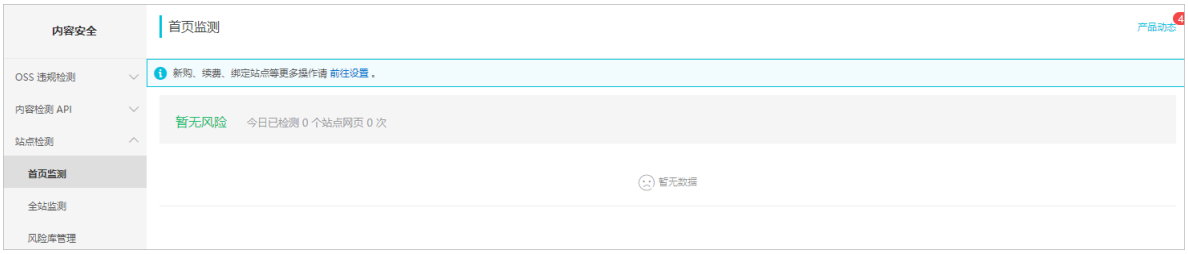
3.2 查看检测结果

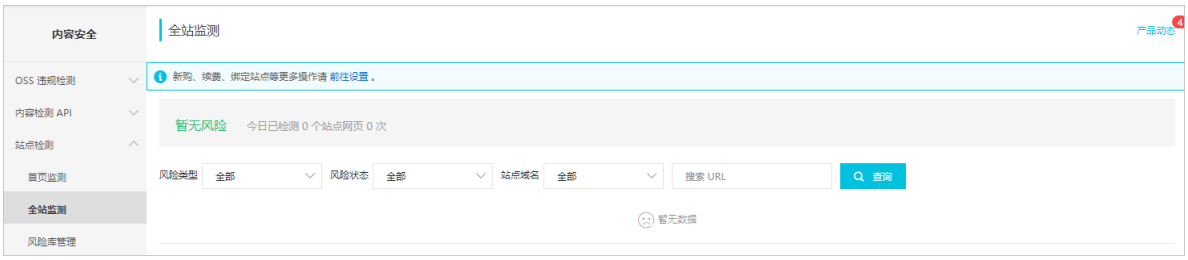
启用站点检测服务后，您可以在内容安全控制台查看站点检测的结果。

操作步骤

1. 登录云盾内容安全控制台。

2. 在左侧导航栏，选择站点检测 > 首页监测，查看最近一次首页检测结果；或者选择站点检测 > 全站监测，查看最近一次全站检测结果。





3. 单击存在风险的URL，查看并确认风险。

   - 消除风险后，单击已处理，完成处理。
   - 如果您对结果有异议，您可以单击纠错或问题反馈，通过表单将问题反馈给我们。在确认问题后，我们将在算法层面进行优化改进。

3.3 监控设置

对绑定站点开启检测时，系统会抓取当前首页作为判断首页是否被篡改的基准。若您更新过首页内容，建议您设置首页防篡改基准，让系统重新抓取当前首页。如果您的网站内容很多，您担心在检测中重要的URL会被遗漏，您可以自定义重点监控URL，系统会优先检测您添加的URL。

前提条件

站点检测实例已绑定站点，并完成验证。更多信息，请参见启用站点检测。

设置首页防篡改基准

1. 登录云盾内容安全控制台。
2. 前往设置 > 站点检测页面，打开实例管理页签。
3. 选择目标实例，在其操作选项中，选择更多 > 设定首页防篡改基准。
4. 确认当前首页基准。如果您想更换首页基准，单击重新获取当前首页，稍后即可查看到系统已重新抓取当前首页作为首页基准。

添加重点监控URL

1. 登录云盾内容安全控制台。
2. 前往设置 > 站点检测页面，打开实例管理页签。

3. 选择目标实例，在其操作选项中，选择更多 > 添加重点监控 URL。
4. 在对话框的输入区域，输入您想要添加的URL，每行一个URL，使用回车换行。

> 说明:
> 最多支持添加5000个URL。

5. 输入完成后，单击提交。

3.4 风险库管理

在站点检测时，如果您需要对特殊的词汇进行专门识别和防控，您可以自定义文本库并将特殊关键词添加进来，进行黑名单防控。在使用站点检测检查图片时，您可以将特定图片定义为白名单、黑名单图片，进行过滤、防控。

自定义文本库

1. 登录云盾内容安全控制台。
2. 前往站点检测 > 风险库管理页面，打开自定义文本库页签。
3. 单击创建文本库。

> 说明:
> 最多支持创建10个文本库。

4. 在创建自定义文本库对话框中，为文本库设置名称，并选择要应用该文本库的实例，然后单击确认。

> 说明:

在检测站点时，只有您选择的实例才会应用该文本库。

5. **选择新创建的文本库，在其操作选项中，单击管理。**

6. **在文本库管理页面，单击新增关键词。**

7. **在新增关键词侧边页，按照页面提示输入或导入关键词，并单击确定，完成添加。**

说明:

新添加的关键词在15分钟内生效。

自定义图库

1. 登录云盾内容安全控制台。
2. 前往站点检测 > 风险库管理页面，打开自定义图库页签。
3. 单击创建图库。

说明:

最多支持创建10个图库。

4. 在创建图库对话框中，完成相关配置，并单击确定。配置说明如下。

配置	说明
名称	输入一个用于识别此图库的名称。
使用场景	选择监黄、暴恐。

配置	说明
识别结果	选择黑名单、白名单。黑名单图库用于特殊防控不良图片，白名单图库会在检测中忽略并过滤您添加的图片。
实例	选择应用该图库的实例。 **说明:** 在检测站点时，只有您选择的实例才会应用该图库。

5. 选择新创建的图库，在其操作选项中，单击管理。

6. 单击选择文件，并上传本地图片至当前图库。

说明:

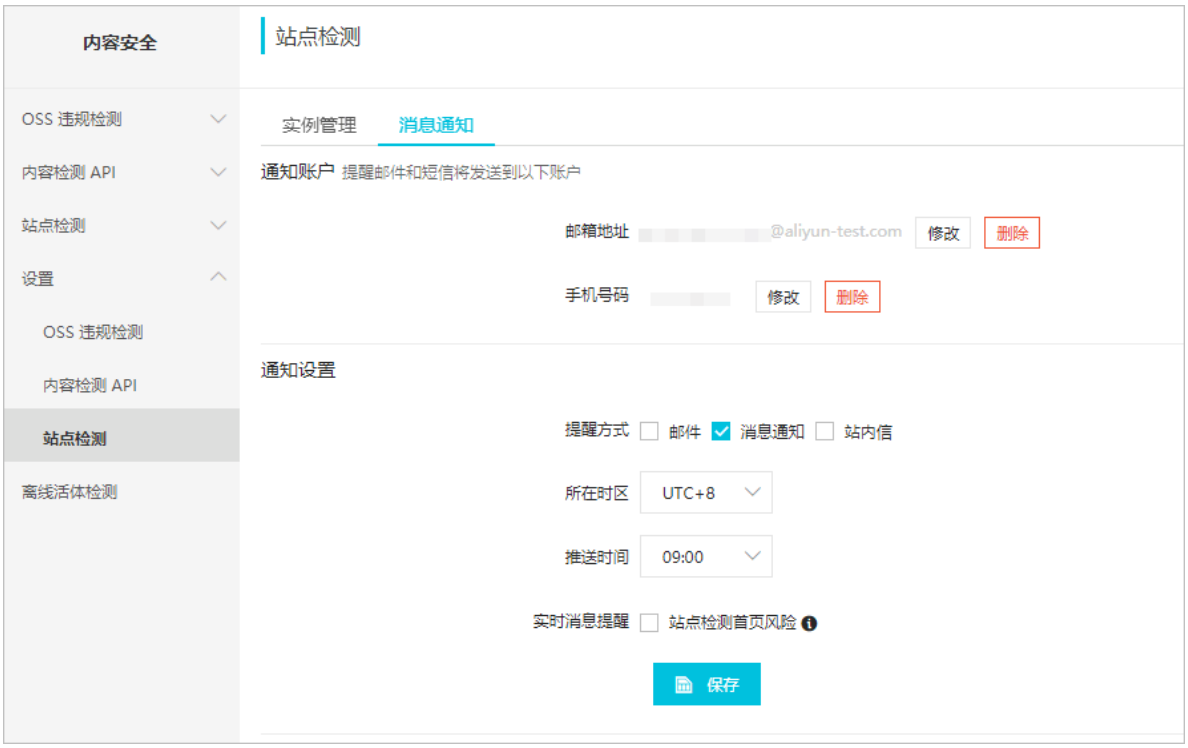
每个图库支持添加最多10000张图片，新添加的图片在15分钟内生效。

3.5 消息通知设置

云盾内容安全的默认消息推送每天触发一次。您可以设置消息接收方式、账号和接收时间，也可以开启或关闭首页风险实时通知。

操作步骤

1. 登录云盾内容安全控制台。
2. 前往设置 > 站点检测页面，打开消息通知页签。
3. 设置风险预警的通知账户（即接收邮箱地址和手机号码），勾选相应的提醒方式（邮件、消息通知、站内信），并设置所在时区和推送时间。



4. 考虑到首页风险的重要程度，您也可以勾选是否开启站点检测首页风险的实时消息提醒。
   开启后，系统一旦检测出首页存在风险，会实时发送消息给您。多次检测到风险时，为避免您被打扰，针对单个域名每天最多发送一次提醒。
5. 单击保存。

4 概述

站点检测帮助您自动检测站点上的风险内容并提供通知服务，帮助您消除站点的内容风险隐患。

简介

站点检测服务定期检查您的网站首页和全站内容，及时发现您的网站在内容安全方面可能存在的风险（例如首页篡改、挂马暗链、色情低俗、涉政暴恐等），并向您展示违规内容的具体地址，帮助您查看和修复。您可以设置消息通知，选择邮件、短信、站内信的方式，获取实时的站点首页风险提醒。

购买站点检测实例后，您需要将实例绑定到您的站点、添加要检测的网站域名和首页地址、设定首页和全站检测的频率，并完成网站鉴权。完成设置后，系统将定期按照您设定的频率对首页和全站内容（包含网页源码、文本和图片）进行检测。如果发现有风险，将按照您设定的消息接收方式通知您。您也可以登录产品控制台查看检测结果。

站点检测的对象是您的网站上的网页和图片，以URL数量进行计数。在单个网站的一个检测周期内，站点检测支持的最大检测容量为10万个URL。

功能描述

站点检测提供首页检测和全站检测功能。

- 首页检测：定期对您网站的首页进行检测，展示最近一次的检查结果。检查结果涵盖首页篡改、挂马暗链、色情低俗、涉政暴恐等风险提示，并提供源码、文本、图片三类呈现方式，供您参照和整改。
- 全站检测：定期对您网站域名下的网页进行自动化全站内容检测，展示最近一次的检查结果。检查结果涵盖挂马暗链、色情低俗、涉政暴恐等风险提示，并提供源码、文本、图片三类呈现方式，供您参照和整改。

设置描述

站点检测支持设置首页防篡改基准、添加重点监控URL以及自定义文本库和图库。

- 设置首页防篡改基准

  通过算法对比网页实时状态和您预设基准状态，判断是否为非法篡改。
- 添加重点监控URL

  添加网站重点监控URL，确保全站检测时重要页面不会遗漏。最多支持添加5000条重点监控URL。

- 自定义文本库和图库

  在使用站点检测服务时，您可以添加自定义关键词进行黑名单防控；添加的关键词在15分钟内生效，关键词只支持UTF-8格式。在使用站点检测进行鉴黄、暴恐等图像服务时，您可以添加自定义图片进行黑名单/白名单防控；添加的图片在15分钟内生效。

  关于该设置的更多介绍，请参见文本反垃圾API。

使用流程

在购买站点检测实例后，您需要将实例绑定到待检测的站点，为站点启用检测服务。然后，您可以在控制台查看首页检测和全站检测的结果。

如果您想进一步保障检测效果，建议您 监控设置，或者通过风险库管理自定义文本库和图库。

您也可以通过消息通知设置设置风险通知方式、开启或关闭首页风险实时通知。