内容审核

快速入门

文档版本 01

发布日期 2025-04-29

华为技术有限公司

商标声明


和其他华为商标均为华为技术有限公司的商标。
本文档提及的其他所有商标或注册商标，由各自的所有人拥有。

注意

您购买的产品、服务或特性等应受华为公司商业合同和条款的约束，本文档中描述的全部或部分产品、服务或特性可能不在您的购买或使用范围之内。除非合同另有约定，华为公司对本文档内容不做任何明示或暗示的声明或保证。

由于产品版本升级或其他原因，本文档内容会不定期进行更新。除非另有约定，本文档仅作为使用指导，本文档中的所有陈述、信息和建议不构成任何明示或暗示的担保。

华为技术有限公司

地址： 深圳市龙岗区坂田华为总部办公楼 邮编：518129

网址： https://www.huawei.com

客户服务邮箱： support@huawei.com

客户服务电话： 4008302118

安全声明

漏洞处理流程

华为公司对产品漏洞管理的规定以“漏洞处理流程”为准，该流程的详细内容请参见如下网址：
https://www.huawei.com/cn/psirt/vul-response-process
如企业客户须获取漏洞信息，请参见如下网址：
https://securitybulletin.huawei.com/enterprise/cn/security-advisory

目 录

1 内容审核服务使用简介

内容审核（Content Moderation），是基于图像、文本、音视频的检测技术，可自动检测涉黄图文违规等内容，对用户上传的图片、文字、音视频进行内容审核，以满足上传要求，帮助客户降低业务违规风险。

内容审核以开放API（Application Programming Interface，应用程序编程接口）的方式提供给用户，用户通过调用API获取推理结果，帮助用户打造智能化业务系统，提升业务效率。目前内容审核包括内容审核-图像、内容审核-文本。

您可以根据以下介绍选择合适的使用方式：

- 通过可视化工具（如curl、Postman）发送请求调用内容审核服务API。

  如果您是开发工程师，熟悉代码编写，熟悉HTTP请求与API调用，您可以通过postman调用、调试API。使用方法请参见**如何使用Postman调用华为云Moderation服务**。
- 通过软件开发工具包（SDK）调用内容审核服务API。

  如果您是开发工程师，熟悉代码编写，内容审核服务为您提供Java、Python、.NET、GO等版本的SDK，方便您快速集成。

2 开通服务步骤说明

说明

本服务暂时仅面向企业用户开放。

内容审核服务申请开通您可以按照如下步骤操作：

1. 已注册华为账号，并完成实名认证。
2. 登录内容审核管理控制台，控制台左上角默认显示服务部署区域，请您根据业务需要选择对应区域，服务部署的区域具体请参见**终端节点**。
3. 在左侧导航栏选择“总览”，进入总览页面，进行以下步骤操作：
   a. 单击“申请开通服务”按钮，进入到新建工单页面。

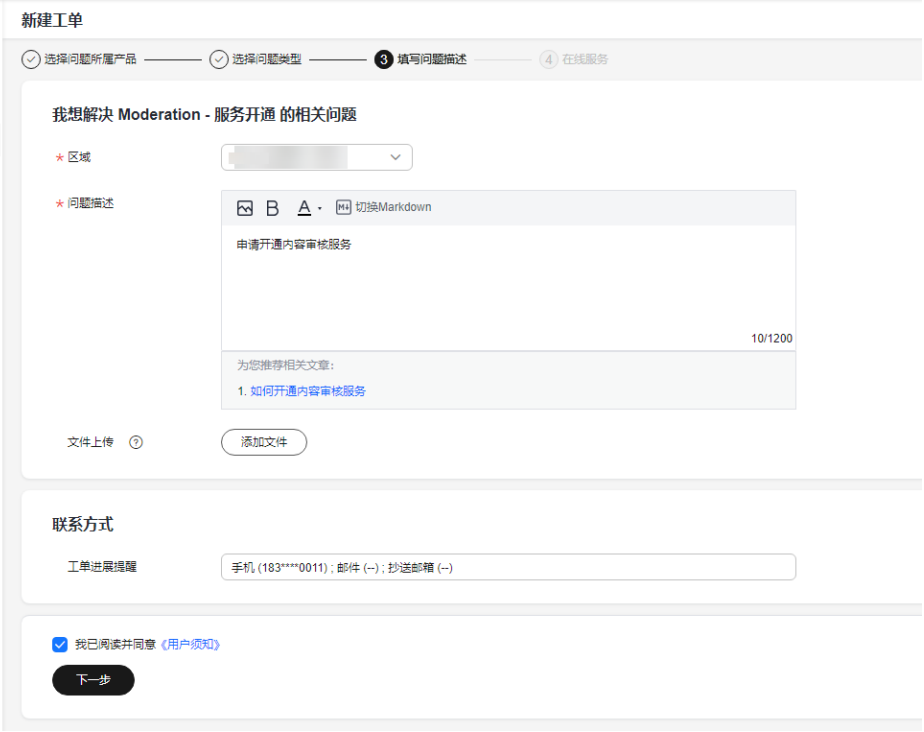
图 2-1 总览页面

   b. 在“我在Moderation遇到问题类型”分类中选择“服务开通”，进入到智能客服对话框中。

图 2-2 服务开通

   c. 在对话框中输入“申请开通内容审核服务”，单击“下一步”。

图 2-3 转人工



d. 单击“未解决，提交工单”在对话框中智能客服将为您创建工单。

e. 等待客服审核完成后帮您开通本服务。

说明

- 服务只需要开通一次即可，后面使用时无需再申请。

4. 商用服务申请成功后，在“总览”页面中显示已经申请开通成功的服务，此时，您可以通过调用API的方式使用内容审核服务。
5. 开通服务后，单击右上角的“预付套餐包”按钮，进入到本服务套餐包购买页面，按需选择想要购买的功能类型和规格，选择完成后单击“立即购买”，确认购买信息无误后完成付款即可开始使用本服务。

说明

目前内容审核服务提供两种计费模式供您选择：按需计费和预付套餐包计费。具体介绍请参见**计费说明**。如果您想使用按需计费的方式，详细费用价格请参见**内容审核价格详情**。

3 调用 API 实现内容审核功能

3.1 内容审核-文本

本章节提供了通过Postman调用“内容审核-文本”的样例，帮助您快速体验并熟悉使用本服务，具体步骤如下：

步骤1 **开通服务**，用户在内容审核控制台，申请开通内容审核-文本服务。

步骤2 **配置自定义词库**，用户可配置自定义白名单词库和自定义黑名单词库。

步骤3 **配置环境**，把准备的配置文件导入到开发环境中。

步骤4 **Token认证鉴权**用户调用API接口时，需要使用Token进行鉴权。

步骤5 **调用服务**，调用API接口使用服务，使用过程中可以随时查看状态码与错误码。

步骤6 **查看调用次数（可选）**，在控制台查看调用详情和调用次数统计。

----结束

开通服务

内容审核-文本服务开通步骤请参考**开通服务步骤说明**。

配置自定义词库

文本内容审核服务可支持用户配置自定义白名单词库和自定义黑名单词库进行文本审核。

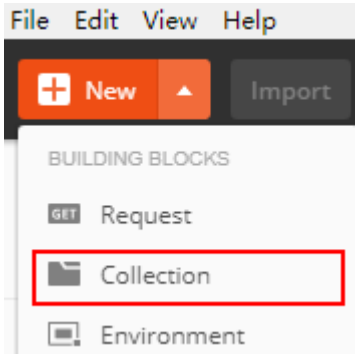
具体步骤请参考**创建自定义词库**。

配置环境

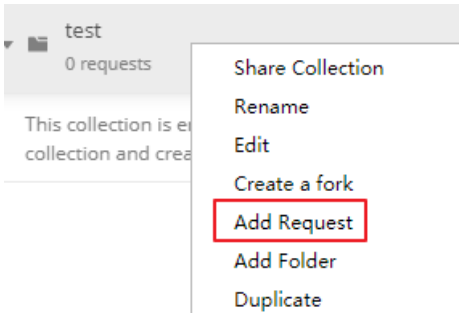
下载并安装Postman。Postman建议使用7.24.0版本。

Token 认证鉴权

1. 在Postman界面，选择“New > Collection”，设置相应的名称并单击“Create”完成创建。



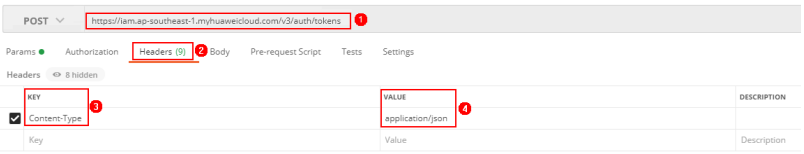
2. 选择创建的Collection，单击鼠标右键，选择“Add Request”，设置Request name并单击“Save”。



3. 请求方式修改为"POST"，输入URL。

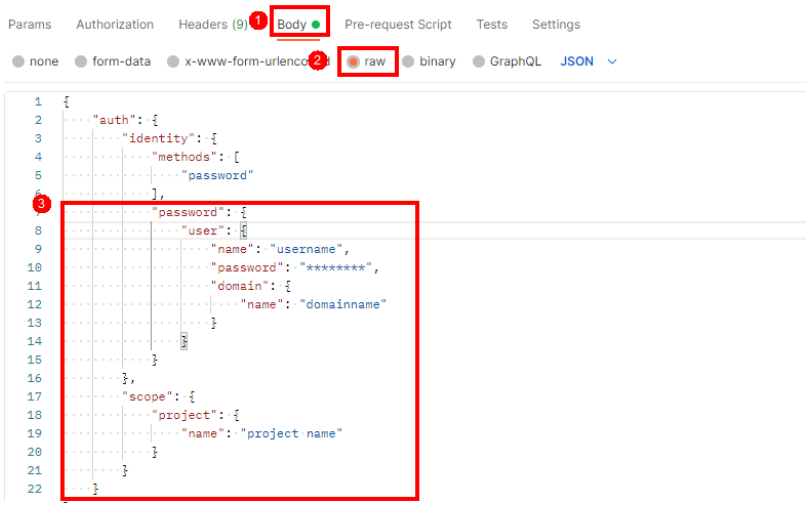
   例如，以ap-southeast-1为例，URL为“https://iam.ap-southeast-1.myhuaweicloud.com/v3/auth/tokens”。

4. 在“**Headers**”列表中添加“KEY”为“Content-Type”，“VALUE”为“application/json”。

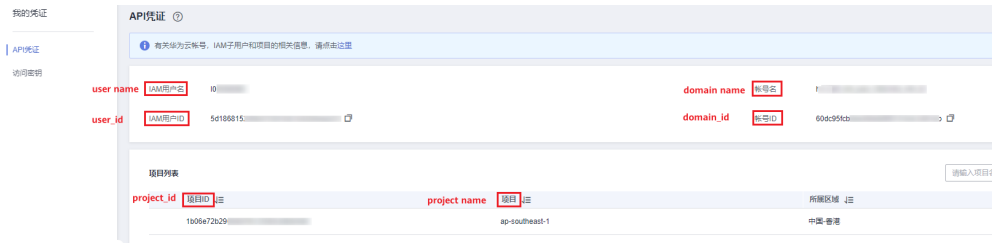


5. 选中“**Body**”的配置项，选中“**raw**”，在空白处添加以下代码。

图 3-1 Token 认证鉴权



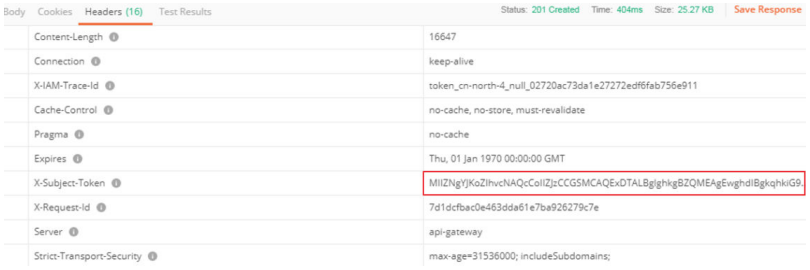
图中代码红框内加粗的蓝色字段需要根据实际值填写，其中username为用户名，domainname为用户所属的账号名称，********为用户登录密码，project name为服务的部署区域，获取方法请登录**我的凭证**获取。



内容审核服务部署的区域必须与调用的服务所在区域一致，本示例中为ap-southeast-1。

```
{
    "auth": {
        "identity": {
            "methods": [
                "password"
            ],
            "password": {
                "user": {
                    "name": "username",
                    "password": "**********",
                    "domain": {
                        "name": "domainname"
                    }
                }
            }
        },
        "scope": {
            "project": {
                "name": "ap-southeast-1"
            }
        }
    }
}
```

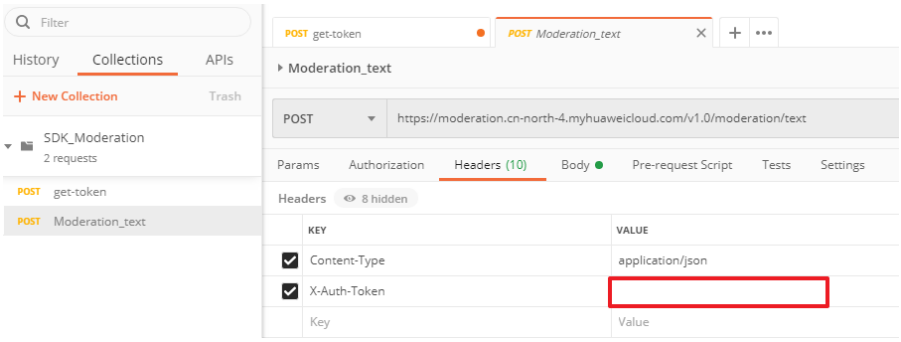
6. 单击右上角“**Send**”按钮发送请求。获取返回结果中的Token值（Token有效期为24小时）。



调用服务

1. 在Postman左侧导航栏的“Collections”目录下单击“Moderation_text”配置文件。
2. 单击“Headers”配置项，将获取的Token复制到“X-Auth-Token”值中。

图 3-2 填入 Token



3. 单击“Body”配置项，将待检测的文本填入到“text”参数中。
   - 关于body体中其他参数说明，请参考**文本内容审核API**。

图 3-3 修改参数

4. 单击“Send”，发送请求，获取调用结果。

```
{
   "request_id": "065561ff4bba1af6dd63a27c5c1371de",
   "result": {
      "details": [],
      "label": "normal",
      "suggestion": "pass"
   }
}
```

查看调用次数（可选）

步骤1 登录内容审核服务管理控制台。

步骤2 在左侧导航栏中选择“内容审核 V3>识别统计>文本审核”，可以查看识别统计详情。您可以设置时间范围，策略（事件类型）来观察这段时间内的调用次数变化情况。

- 识别结果统计：显示一段时间范围，内容审核的调用总数，拒绝数，疑似数和通过数，帮助您更好了解服务的调用情况和审核情况。
  - 总数：指的是审核调用总次数。
  - 拒绝数：指的是block总数，即文本中包含敏感信息，审核不通过的次数。
  - 疑似数：指的是review总数，即人工复查审核的次数。
  - 通过数：指的是pass总数，即通过审核的次数。
- 数据趋势：显示您设置的这段时间范围内，总数，拒绝数，疑似数和通过数的变化趋势。
- 拒绝数据原因分布：显示您设置的这段时间范围内，审核不通过的检测场景占比数。

- 疑似数据原因分布：显示您设置的这段时间范围内，需要人工复查的检测场景占比数。

----结束

3.2 内容审核-图像

本章节提供了通过Postman调用“内容审核-图像”的样例，帮助您快速体验并熟悉使用本服务，具体步骤如下：

步骤1 **开通服务**，用户在内容审核控制台，申请开通内容审核-文本服务。

步骤2 **配置自定义词库**，用户可配置自定义白名单词库和自定义黑名单词库。

步骤3 **配置环境**，把准备的配置文件导入到开发环境中。

步骤4 **Token认证鉴权**用户调用API接口时，需要使用Token进行鉴权。

步骤5 **调用服务**，调用API接口使用服务，使用过程中可以随时查看状态码与错误码。

步骤6 **查看调用次数（可选）**，在控制台查看调用详情和调用次数统计。

----结束

开通服务

内容审核-图像服务开通步骤请参考**开通服务步骤说明**。

配置自定义词库

文本内容审核服务可支持用户配置自定义白名单词库和自定义黑名单词库进行文本审核。

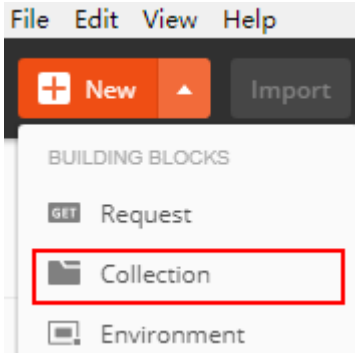
具体步骤请参考**创建自定义词库**。

配置环境

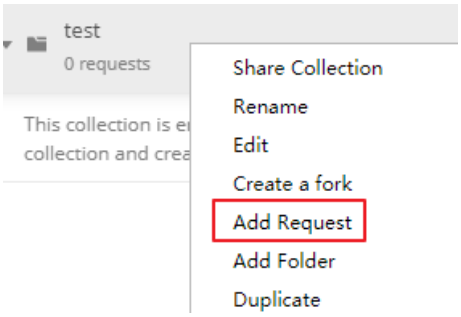
下载并安装Postman。Postman建议使用7.24.0版本。

Token 认证鉴权

1. 在Postman界面，选择“New > Collection”，设置相应的名称并单击“Create”完成创建。



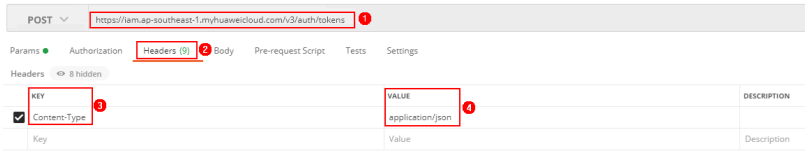
2. 选择创建的Collection，单击鼠标右键，选择“Add Request”，设置Request name并单击“Save”。



3. 请求方式修改为"POST"，输入URL。

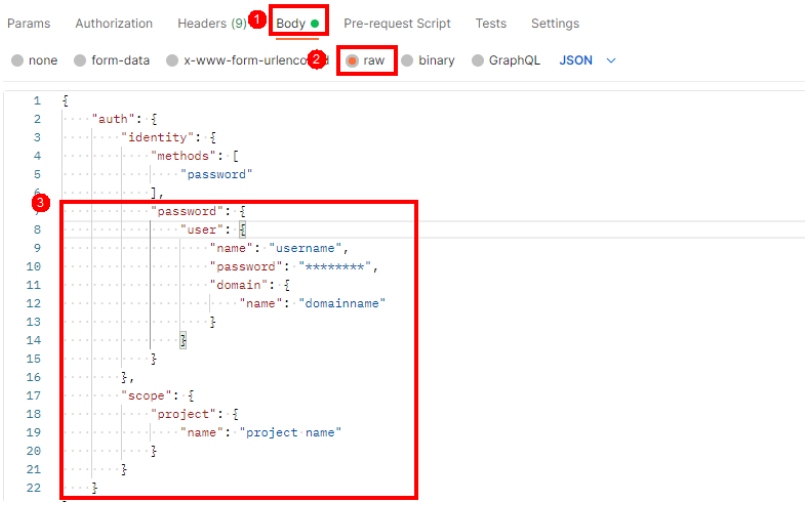
   例如，以ap-southeast-1为例，URL为“https://iam.ap-southeast-1.myhuaweicloud.com/v3/auth/tokens”。

4. 在“**Headers**”列表中添加“KEY”为“Content-Type”，“VALUE”为“application/json”。

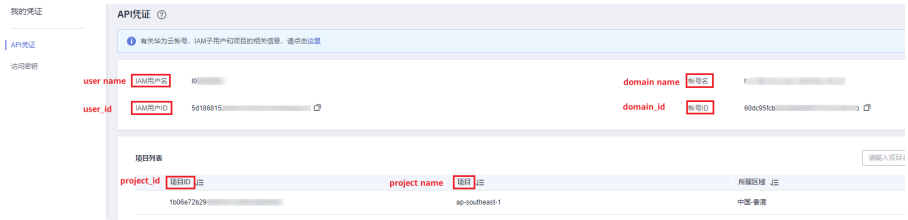


5. 选中“**Body**”的配置项，选中“**raw**”，在空白处添加以下代码。

图 3-4 Token 认证鉴权



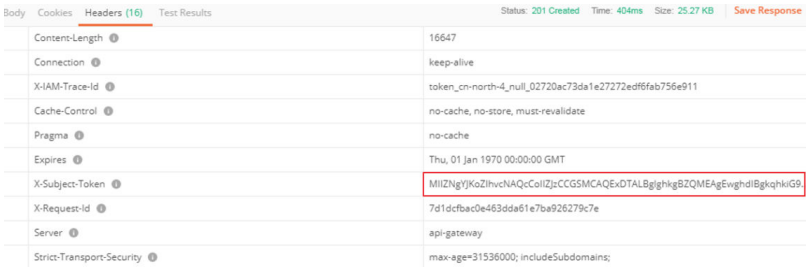
图中代码红框内加粗的蓝色字段需要根据实际值填写，其中username为用户名，domainname为用户所属的账号名称，********为用户登录密码，project name为服务的部署区域，获取方法请登录**我的凭证**获取。



内容审核服务部署的区域必须与调用的服务所在区域一致，本示例中为ap-southeast-1。

```
{
    "auth": {
        "identity": {
            "methods": [
                "password"
            ],
            "password": {
                "user": {
                    "name": "username",
                    "password": "**********",
                    "domain": {
                        "name": "domainname"
                    }
                }
            }
        },
        "scope": {
            "project": {
                "name": "ap-southeast-1"
            }
        }
    }
}
```

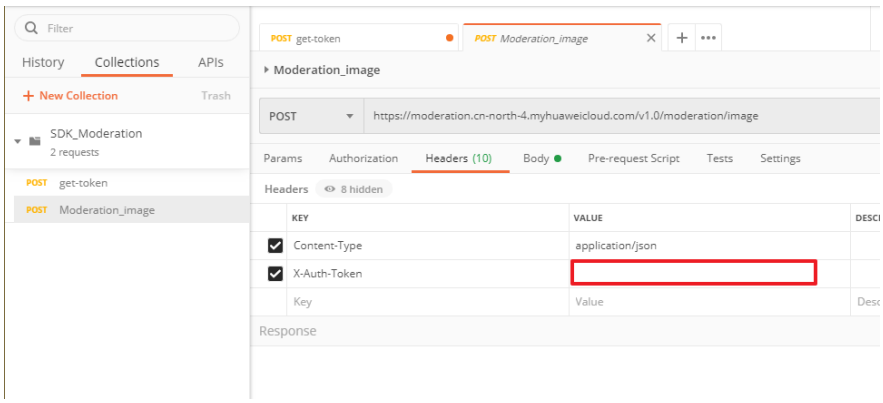
6. 单击右上角“**Send**”按钮发送请求。获取返回结果中的Token值（Token有效期为24小时）。



调用服务

1. 在Postman左侧导航栏的“Collections”目录下单击“Moderation_image”配置文件。
2. 单击“Headers”配置项，将获取的Token复制到“X-Auth-Token”值中。

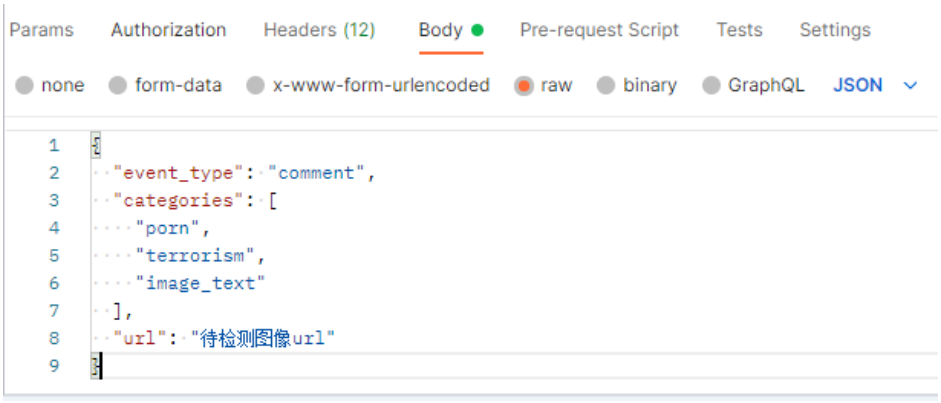
图 3-5 填入 Token



3. 单击“Body”配置项，将获取的图片url填写到“url”参数中。

- 关于body体中其他参数说明，请参考**图像内容审核API**。

图 3-6 修改参数



4. 单击“Send”，发送请求，获取调用结果。

```
{
   "request_id": "bcacaef5367b525620ec92531246af71",
   "result": {
      "details": [],
      "suggestion": "pass"
   }
}
```

查看调用次数（可选）

内容审核-图像API接口，分析并识别用户上传的图像内容是否有敏感内容（如色情等），并将识别结果返回给用户。

调用内容审核-图像API接口，您可以按照如下步骤操作：

步骤1 进入内容审核管理控制台。

步骤2 在左侧导航栏中选择“内容审核 V3>识别统计>图像审核”，可以查看识别统计详情，如**图3-7**所示。您可以设置时间范围，策略（事件类型）来观察这段时间内的调用次数变化情况。

图 3-7 识别统计

- 识别结果统计：显示一段时间范围，内容审核的调用总数，拒绝数，疑似数和通过数，帮助您更好了解服务的调用情况和审核情况。
  - 总数：指的是审核调用总次数。
  - 拒绝数：指的是block总数，即文本中包含敏感信息，审核不通过的次数。
  - 疑似数：指的是review总数，即人工复查审核的次数。
  - 通过数：指的是pass总数，即通过审核的次数。
- 数据趋势：显示您设置的这段时间范围内，总数，拒绝数，疑似数和通过数的变化趋势。

图 3-8 原因分布

- 拒绝数据原因分布：显示您设置的这段时间范围内，审核不通过的检测场景占比数。
- 疑似数据原因分布：显示您设置的这段时间范围内，需要人工复查的检测场景占比数。

----结束

4 调用 SDK 实现内容审核功能

4.1 内容审核-文本

本章节提供了通过Java SDK调用“内容审核-文本”服务的样例，用户直接调用接口函数即可使用SDK功能。具体流程如下：

步骤1 **开通服务**，用户在“服务列表”或“服务管理”页面选择内容审核-文本服务申请开通。

步骤2 **配置自定义词库（可选）**，用户可配置自定义白名单词库和自定义黑名单词库。

步骤3 **配置环境**，配置JAVA开发环境和开发工具。

步骤4 **获取示例代码**，获取SDK和样例工程，导入到开发环境中。

步骤5 **认证鉴权**，使用AK/SK方式进行认证。

步骤6 **调用服务**，调用API接口使用服务，使用过程中可以随时查看状态码与错误码。

----结束

开通服务

内容审核-文本服务开通步骤请参考**开通服务步骤说明**。

配置自定义词库（可选）

文本内容审核服务可支持用户配置自定义白名单词库和自定义黑名单词库进行文本审核。

具体步骤请参考**创建自定义词库**。

配置环境

您可以基于内容审核SDK通过编写代码的方式调用内容审核-文本API。在使用SDK和调用API时您需要进行环境配置。具体操作步骤如下：

1. 最新版本内容审核SDK软件包和文档，请在**SDK中心**获取。
2. 环境配置请参见**Java开发环境配置**。

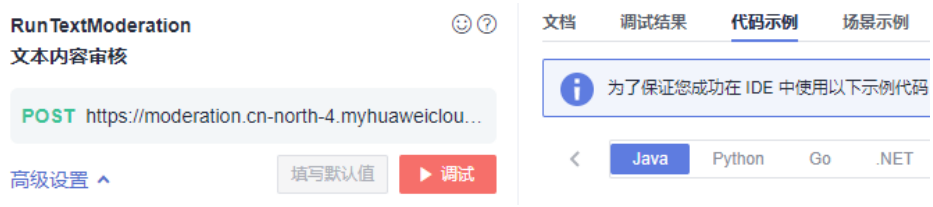
3. 获取SDK参见**SDK获取和安装**。

获取示例代码

获取文本审核V3 Java SDK示例。

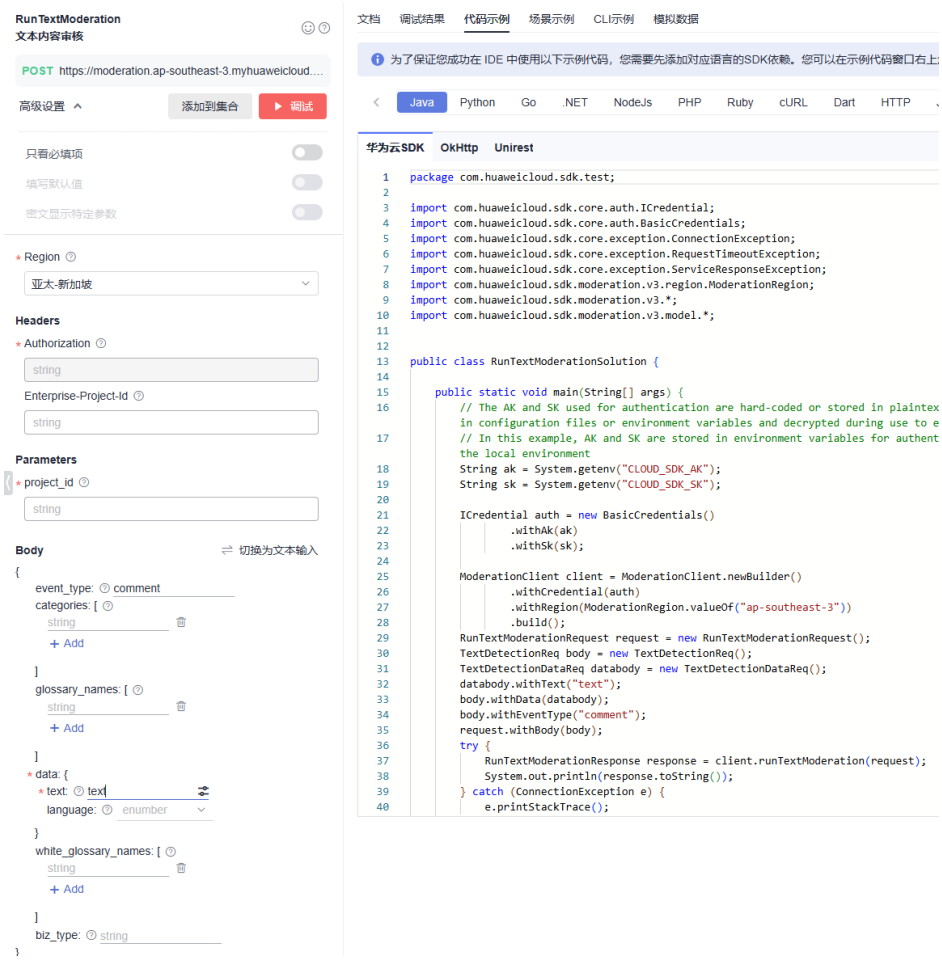
1. 登录**API Explorer**，在“代码示例”中选择“Java”。见下图。

图 4-1 获取 java SDK



2. 填写请求Body参数。输入event_type、data.text参数。见下图。

图 4-2 请求 Body 参数



3. 复制示例代码至Maven项目。

认证鉴权

内容审核服务认证方式有Token和AK/SK两种方式，本示例中使用AK/SK方式进行认证。

1. 获取AK/SK。

   AK/SK即访问密钥，请登录**我的凭证**页面，选择“访问密钥 > 新增访问密钥”获取。

   图 4-3 新增访问密钥

2. 配置Java SDK中的AK/SK，进行AK/SK认证鉴权。

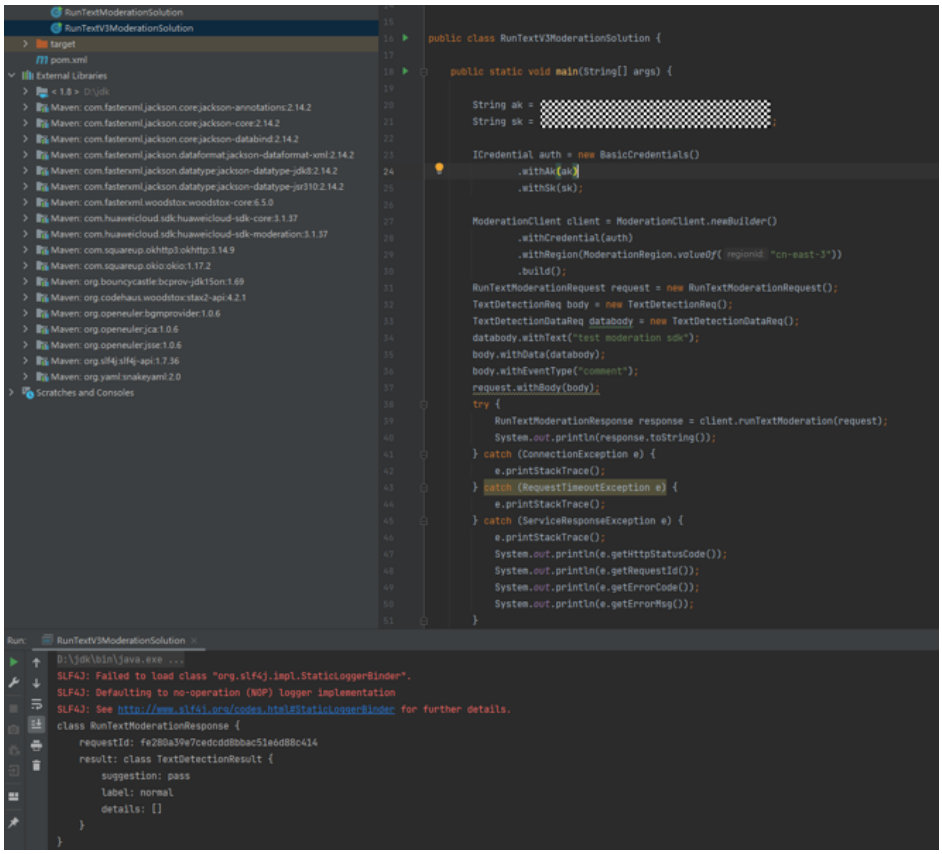
   替换下图中的AK/SK参数。

```
public static void main(String[] args) {
    // The AK and SK used for authentication are hard-coded or stored in plaintext, which has gre
    in configuration files or environment variables and decrypted during use to ensure security.
    // In this example, AK and SK are stored in environment variables for authentication. Before
    the local environment
    String ak = System.getenv("CLOUD_SDK_AK");
    String sk = System.getenv("CLOUD_SDK_SK");

    ICredential auth = new BasicCredentials()
            .withAk(ak)
            .withSk(sk);
```

调用服务

1. 运行代码示例，获取识别结果，如下图所示。



2. 查看调用次数。您可以在“内容审核 V3>识别统计>文本审核”页查看调用详情和调用次数统计。
   - 识别结果统计：显示一段时间范围，内容审核的调用总数，拒绝数，疑似数和通过数，帮助您更好了解服务的调用情况和审核情况。
     - 总数：指的是审核调用总次数。
     - 拒绝数：指的是block总数，即文本中包含敏感信息，审核不通过的次数。
     - 疑似数：指的是review总数，即人工复查审核的次数。
     - 通过数：指的是pass总数，即通过审核的次数。
   - 数据趋势：显示您设置的这段时间范围内，总数，拒绝数，疑似数和通过数的变化趋势。
   - 拒绝数据原因分布：显示您设置的这段时间范围内，审核不通过的检测场景占比数。
   - 疑似数据原因分布：显示您设置的这段时间范围内，需要人工复查的检测场景占比数。

4.2 内容审核-图像

本章节提供了通过Java SDK调用“内容审核-图像”服务的样例，用户直接调用接口函数即可使用SDK功能。具体流程如下：

步骤1 开通服务，用户在“服务列表”或“服务管理”页面选择内容审核-图像服务申请开通。

步骤2 **配置自定义词库（可选）**，用户可配置自定义白名单词库和自定义黑名单词库。

步骤3 **配置环境**，配置JAVA开发环境和开发工具。

步骤4 **获取示例代码**，获取SDK和样例工程，导入到开发环境中。

步骤5 **认证鉴权**，使用AK/SK方式进行认证。

步骤6 **调用服务**，调用API接口使用服务，使用过程中可以随时查看状态码与错误码。

----结束

开通服务

内容审核-图像服务开通步骤请参考**开通服务步骤说明**。

配置自定义词库（可选）

文本内容审核服务可支持用户配置自定义白名单词库和自定义黑名单词库进行文本审核。

具体步骤请参考**创建自定义词库**。

配置环境

您可以基于内容审核SDK通过编写代码的方式调用内容审核-文本API。在使用SDK和调用API时您需要进行环境配置。具体操作步骤如下：

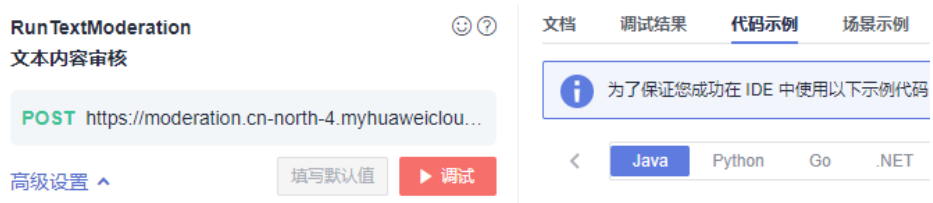
1. 最新版本内容审核SDK软件包和文档，请单击**SDK中心**。
2. 环境配置请参见**Java开发环境配置**。
3. 获取SDK参见**SDK获取和安装**。

获取示例代码

获取文本审核V3 Java SDK示例。

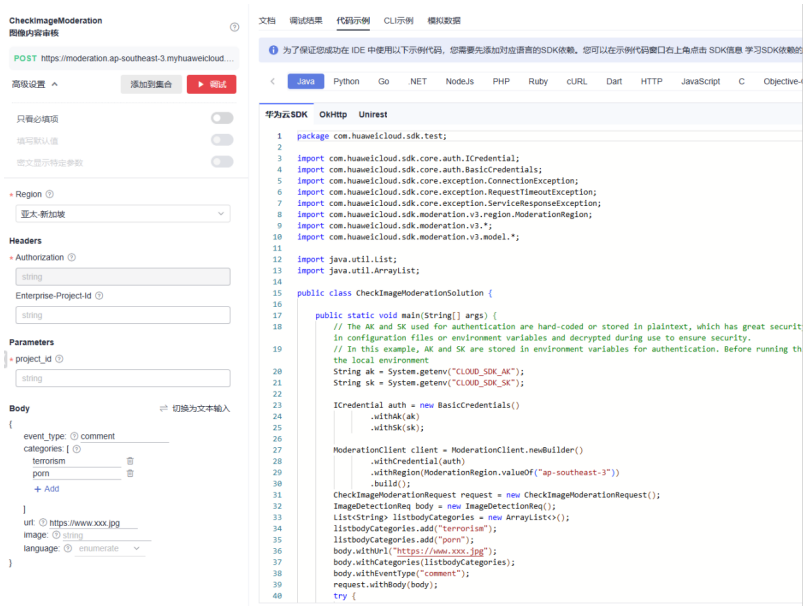
1. 登录**API Explorer**，在“代码示例”中选择“Java”。见下图。

图 4-4 获取 java SDK



2. 填写请求Body参数。输入event_type、data.text参数。见下图。

图 4-5 请求 Body 参数



3. 复制示例代码至Maven项目。

认证鉴权

内容审核服务认证方式有Token和AK/SK两种方式，本示例中使用AK/SK方式进行认证。

1. 获取AK/SK。

   AK/SK即访问密钥，请登录**我的凭证**页面，选择“访问密钥 > 新增访问密钥”获取。

图 4-6 新增访问密钥

2. 配置Java SDK中的AK/SK，进行AK/SK认证鉴权。

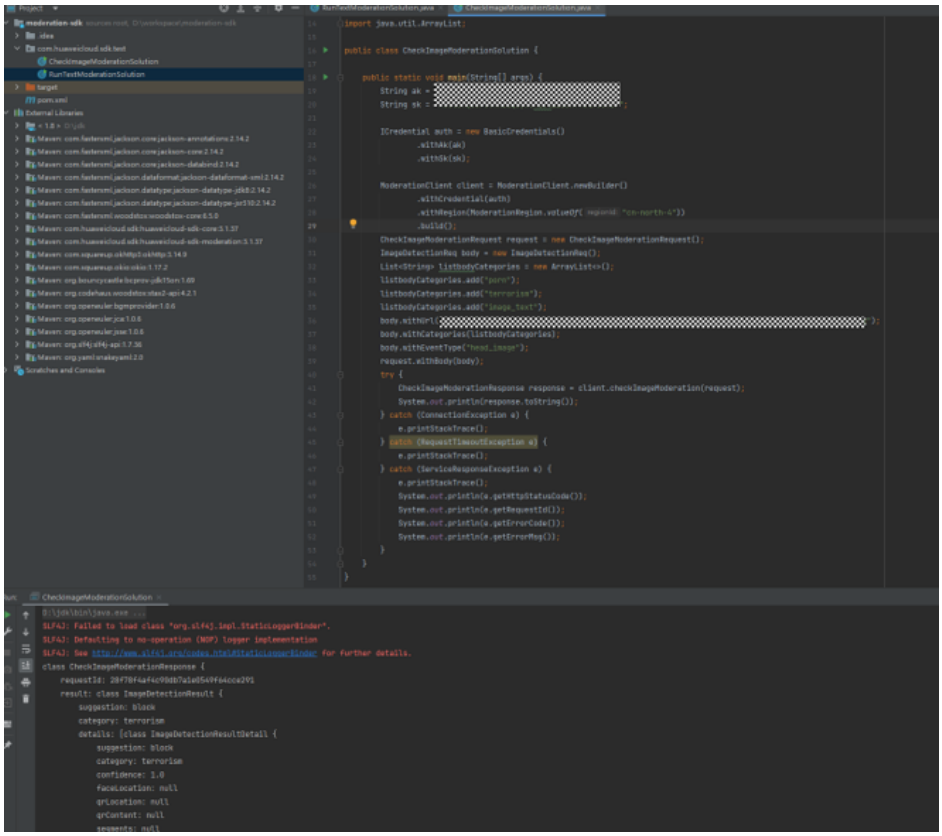
   替换下图中的AK/SK参数。

```
public static void main(String[] args) {
    // The AK and SK used for authentication are hard-coded or stored in plaintext, which has gre
    in configuration files or environment variables and decrypted during use to ensure security.
    // In this example, AK and SK are stored in environment variables for authentication. Before
    the local environment
    String ak = System.getenv("CLOUD_SDK_AK");
    String sk = System.getenv("CLOUD_SDK_SK");

    ICredential auth = new BasicCredentials()
            .withAk(ak)
            .withSk(sk);
```


调用服务

1. 运行代码示例，获取识别结果，如下图所示。。



2. 查看调用次数。您可以在“服务列表”，“图像审核”页查看调用详情和调用次数统计。如图4-7所示。

图 4-7 识别统计

- 识别结果统计：显示一段时间范围，内容审核的调用总数，拒绝数，疑似数和通过数，帮助您更好了解服务的调用情况和审核情况。
  - 总数：指的是审核调用总次数。

- 拒绝数：指的是block总数，即文本中包含敏感信息，审核不通过的次数。
- 疑似数：指的是review总数，即人工复查审核的次数。
- 通过数：指的是pass总数，即通过审核的次数。

– 数据趋势：显示您设置的这段时间范围内，总数，拒绝数，疑似数和通过数的变化趋势。

图 4-8 原因分布

– 拒绝数据原因分布：显示您设置的这段时间范围内，审核不通过的检测场景占比数。
– 疑似数据原因分布：显示您设置的这段时间范围内，需要人工复查的检测场景占比数。