Content Moderation

Getting Started

Issue   01

Date   2025-07-03

Huawei Technologies Co., Ltd.

Address:	Huawei Industrial Base Bantian, Longgang Shenzhen 518129 People's Republic of China
Website:	https://www.huawei.com
Email:	support@huawei.com

Security Declaration

Vulnerability

Huawei's regulations on product vulnerability management are subject to the *Vul. Response Process.* For details about this process, visit the following web page:
https://www.huawei.com/en/psirt/vul-response-process
For vulnerability information, enterprise customers can visit the following web page:
https://securitybulletin.huawei.com/enterprise/en/security-advisory

Contents

1 Content Moderation Overview

Content Moderation adopts image, text, and audio and video detection technologies that detect pornography and images and text violating related laws or regulations. This reduces potential business risks.

Content Moderation provides services through open application programming interfaces (APIs). You can obtain the inference result by calling APIs. It helps you build an intelligent service system and improves service efficiency. Currently, Content Moderation includes Moderation (Image) and Moderation (Text).

The following methods are available to call Content Moderation APIs:

- A software tool such as curl and Postman

  These are good options if you are more comfortable writing code, HTTP requests, and API calls. For details, see **Using Postman to Call Content Moderation**.

- Software development kit (SDK)

  SDKs for Java, Python, .NET, and GO are available for quick integration.

2 Subscribing to the Service

> **NOTE**
>
> This service istemporarily available only to enterprise users.

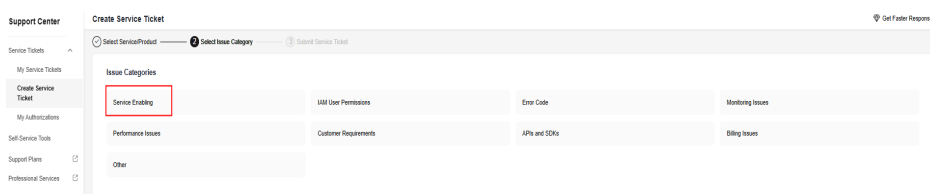To subscribe to Content Moderation, perform the following steps:

1. Sign up for a HUAWEI ID and complete real-name authentication.

2. Log in to the Content Moderation console and select a region. For details about the region where the Content Moderation service is available, see **Endpoints**.

3. In the navigation pane on the left, choose **Overview** and perform the following operations:

    a. Click **Subscribe to Service**.

    Figure 2-1 Overview page
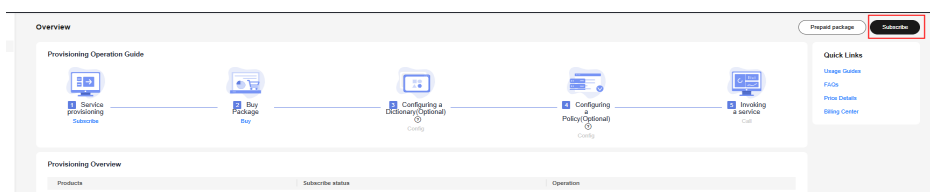
    b. On the displayed **Create Service Ticket** page, click **Service Enabling**.

    Figure 2-2 Service Enabling

    c. Enter **Subscribe to Content Moderation** in the text box and click **Next**.

Figure 2-3 Human agent

d. Click **Submit Service Ticket**. The chatbot will create a service ticket for
    you.

e. Wait for the customer service to review your request and enable the
    service for you.

    ☐ NOTE

    ● You only need to subscribe to the service once.

4. After the subscription is successful, the Content Moderation service is
    displayed on the **Overview** page. You can call APIs to use the service.

5. In the upper right corner, click **Prepay to Get Discounts**. On the displayed
    page, select your desired type and specifications, set other parameters, and
    click **Next**. Then confirm the information, click **Submit**, and pay the order to
    enable the service.

    ☐ NOTE

    Content Moderation supports both the pay-per-use and yearly/monthly billing modes.
    For details, see **Billing**. If you want to use the pay-per-use billing mode, see **Content
    Moderation Pricing Details**.

3 Using Postman to Call Content Moderation

3.1 Moderation (Text)

This section provides an example of how to use Postman to call Moderation (Text), helping you quickly get familiar with the service. The procedure is as follows:

Step 1 **Subscribe to Moderation (Text)** on the Content Moderation console.

Step 2 **Configure allowlist and blocklist dictionaries.**

Step 3 **Configure an environment**. Import the prepared configuration file into the development environment.

Step 4 **Use a token for authentication**. When you call an API, a token is required for authentication.

Step 5 **Call the service API** and view the status code and error code whenever necessary.

Step 6 **(Optional) View call statistics** and call details.

----End

Subscribing to the Service

Subscribe to Moderation (Text) by following the instructions in **Subscribing to the Service**.

Configuring Custom Dictionaries

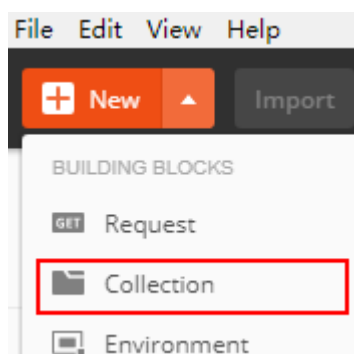You can configure allowlist and blocklist dictionaries for text moderation.

For details, see .
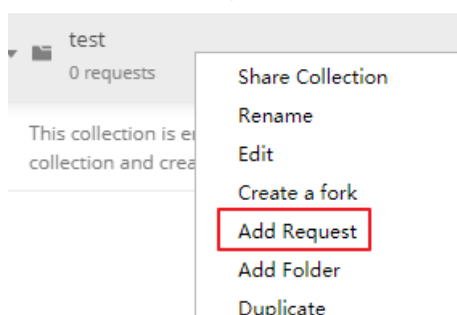
Configuring an Environment

Download and install Postman. Postman 7.24.0 is recommended.
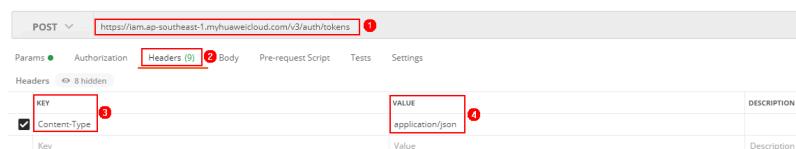
Using a Token for Authentication

1. On the Postman page, choose **New** > **Collection**, set the name, and click **Create** to create a collection.
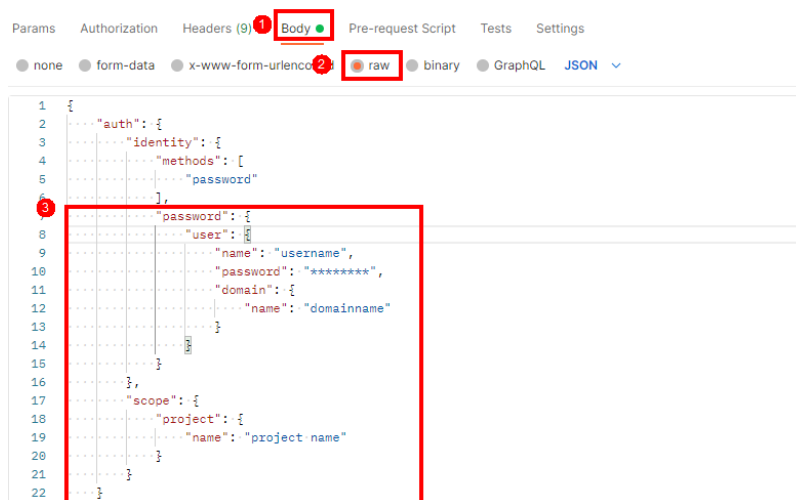
2. Right-click the created collection and select **Add Request** from the shortcut menu. Set the request name and click **Save**.

3. Change the request mode to **POST** and enter the URL.

   For example, if your region is **ap-southeast-1**, set the URL to **https://iam.ap-southeast-1.myhuaweicloud.com/v3/auth/tokens**.

4. On the **Headers** page, set **KEY** to **Content-Type** and **VALUE** to **application/json**.

5. Click **Body**, select **raw**, and enter the following code in the blank area.

Figure 3-1 Token-based authentication

Replace *username*, ******** (login password), *domainname*, and *project name* (region where the service is deployed) with the actual values. To obtain the values of these parameters, log in to the management console and click **My Credentials**.

The region where Content Moderation is deployed must be the same as the region where you call the service. In this example, the region is **ap-southeast-1**.

```
{
    "auth": {
        "identity": {
            "methods": [
                "password"
            ],
            "password": {
                "user": {
                    "name": "username",
                    "password": "**********",
                    "domain": {
                        "name": "domainname"
                    }
                }
            }
        },
        "scope": {
            "project": {
                "name": "ap-southeast-1"
            }
        }
    }
}
```
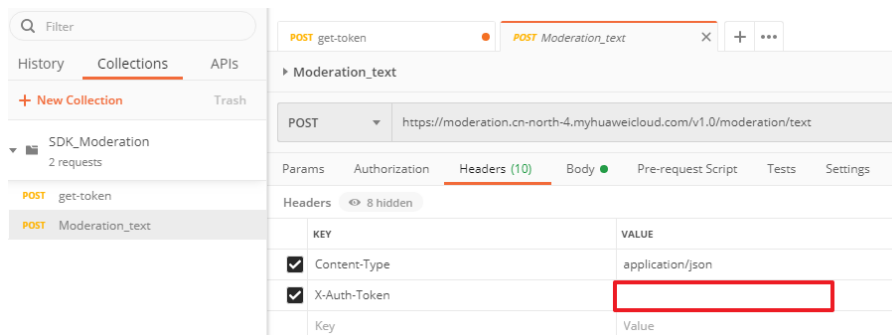
6. Click **Send** in the upper right corner to send the request. Obtain the token from the returned result. The token is valid for 24 hours.

Calling a Service

1. In the **Collections** directory in the navigation pane of Postman, click the **Moderation_text** configuration file.

2. Click **Headers** and copy the token value to **X-Auth-Token**.

   Figure 3-2 Entering the token

3. Click **Body** and enter the text to be checked for the **text** parameter.
   – For details about other parameters in the body, see the **Text Moderation API**.

   Figure 3-3 Modifying the parameters
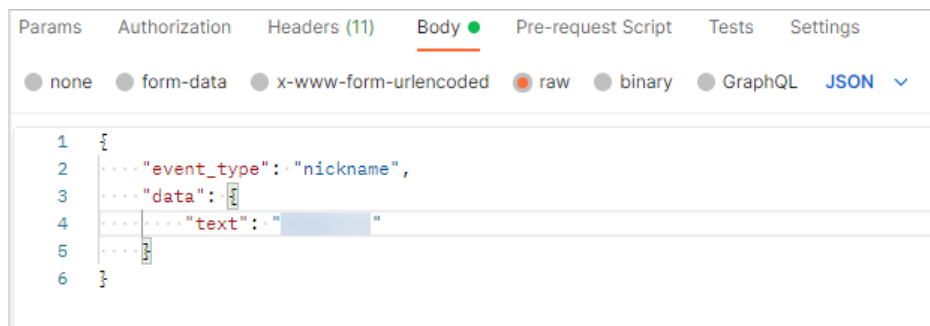
4. Click **Send** to send the request and obtain the API call results.

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

(Optional) Viewing the Number of Calls

Step 1 Log in to the Content Moderation management console.

Step 2 In the left navigation pane, choose **Content Moderation V3** > **Identification Statistics**. On the **Text Moderation** page, you can view the identification statistics. You can set a time range and select a policy (event type) to view the change of the number of API calls within the time range.

- **Identification Statistics**: total number of calls to Content Moderation, number of rejected calls, number of suspected calls, and number of passed calls within a specified period of time, helping you better learn the calls and moderation status of Content Moderation.

  - **Total**: total number of calls made to Content Moderation

  - **Rejected**: number of calls made to Content Moderation that are rejected because the text contains sensitive information

  - **Suspected**: number of calls made to Content Moderation that require manual review

  - **Passed**: number of calls made to Content Moderation that are approved

- **Data Trend**: trend of the total number of calls, number of rejected calls, number of suspected calls, and number of passed calls within the specified period of time

- **Causes of Rejected Data**: proportion of detection scenarios that fail to pass the review within the specified period of time

- **Causes of Suspected Data**: proportion of detection scenarios that require manual review within the specified period of time

----End

3.2 Moderation (Image)

This section provides an example of how to use Postman to call Moderation (Image), helping you quickly get familiar with the service. The procedure is as follows:

Step 1 **Subscribe to Moderation (Image)** on the Content Moderation console.

Step 2 **Configure allowlist and blocklist dictionaries.**

Step 3 **Configure an environment**. Import the prepared configuration file into the development environment.

Step 4 **Use a token for authentication**. When you call an API, a token is required for authentication.

Step 5 **Call the service API** and view the status code and error code whenever necessary.

Step 6 **(Optional) View call statistics** and call details.

----End

Subscribing to the Service

Subscribe to Moderation (Image) by following the instructions in **Subscribing to the Service**.

Configuring Custom Dictionaries

You can configure allowlist and blocklist dictionaries for image moderation.
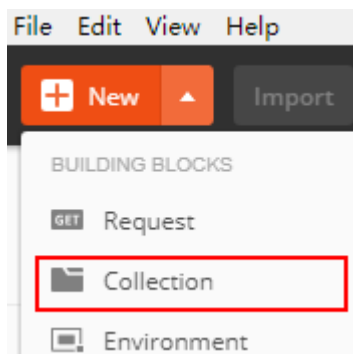
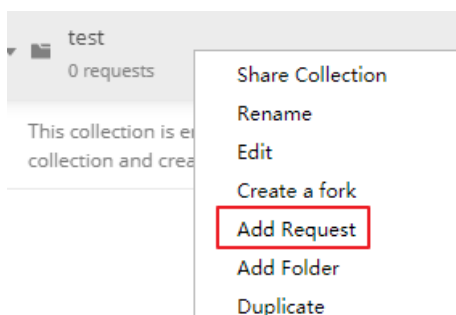For details, see .

Configuring an Environment

Download and install Postman. Postman 7.24.0 is recommended.
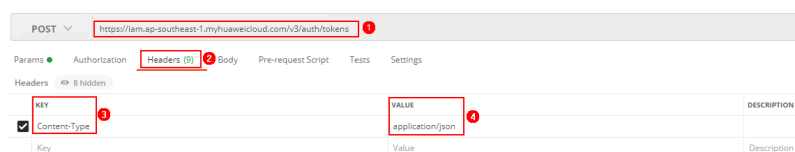
Using a Token for Authentication

1. On the Postman page, choose **New** > **Collection**, set the name, and click **Create** to create a collection.

2. Right-click the created collection and select **Add Request** from the shortcut menu. Set the request name and click **Save**.
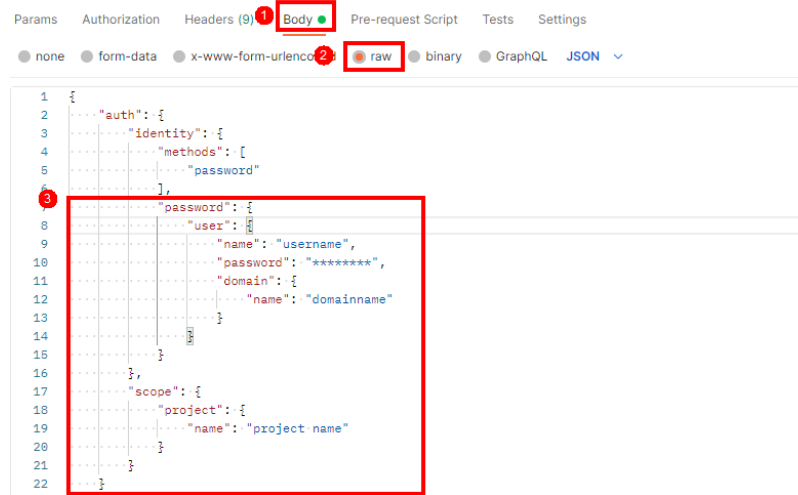

3. Change the request mode to **POST** and enter the URL.

   For example, if your region is **ap-southeast-1**, set the URL to **https://iam.ap-southeast-1.myhuaweicloud.com/v3/auth/tokens**.

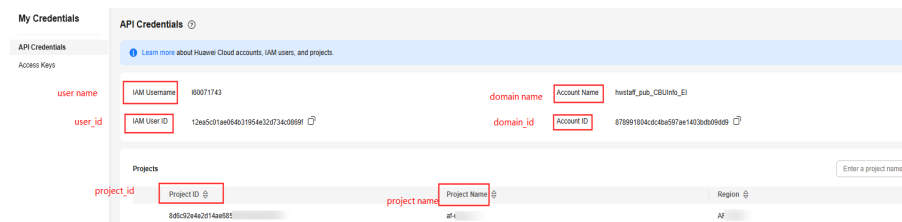4. On the **Headers** page, set **KEY** to **Content-Type** and **VALUE** to **application/json**.

5. Click **Body**, select **raw**, and enter the following code in the blank area.

Figure 3-4 Token-based authentication

Replace *username*, ******** (login password), *domainname*, and *project name* (region where the service is deployed) with the actual values. To obtain the values of these parameters, log in to the management console and click **My Credentials**.

The region where Content Moderation is deployed must be the same as the region where you call the service. In this example, the region is **ap-southeast-1**.

```
{
    "auth": {
        "identity": {
            "methods": [
                "password"
            ],
            "password": {
                "user": {
                    "name": "username",
                    "password": "**********",
                    "domain": {
                        "name": "domainname"
                    }
                }
            }
        },
        "scope": {
            "project": {
                "name": "ap-southeast-1"
            }
        }
    }
}
```
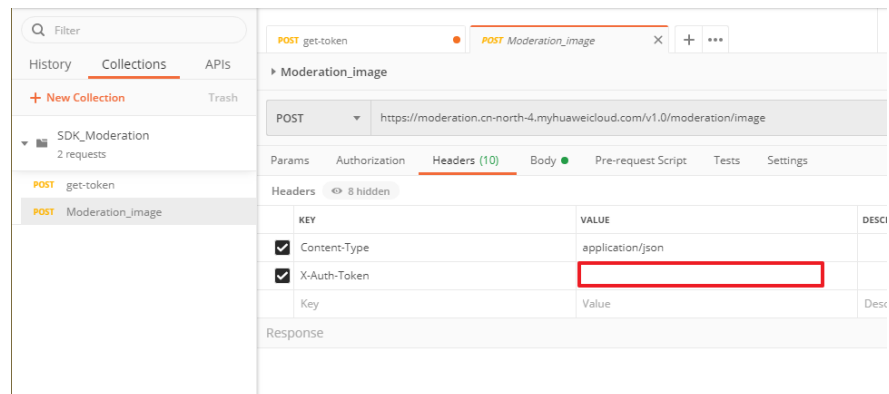
6. Click **Send** in the upper right corner to send the request. Obtain the token from the returned result. The token is valid for 24 hours.

Calling a Service

1. In the **Collections** directory in the navigation pane of Postman, click the **Moderation_image** configuration file.

2. Click **Headers** and copy the token value to **X-Auth-Token**.

Figure 3-5 Entering the token

3. Click **Body** and enter the URL of the image for **url**.
   – For details about other parameters in the body, see the **Image Moderation API**.

Figure 3-6 Modifying the parameters
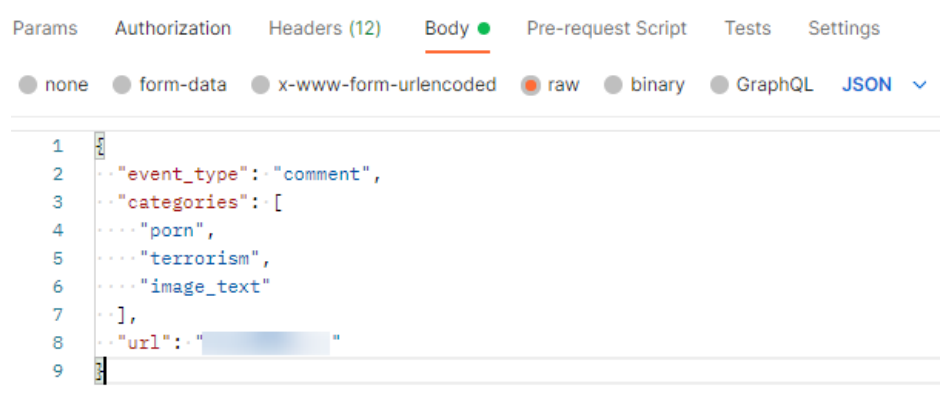
4. Click **Send** to send the request and obtain the API call results.

```
{
    "request_id": "bcacaef5367b525620ec92531246af71",
    "result": {
```

```
        "details": [],
        "suggestion": "pass"
    }
}
```

(Optional) Viewing the Number of Calls

Image Moderation analyzes and identifies sensitive content (such as pornography) in images uploaded by users and returns the result.

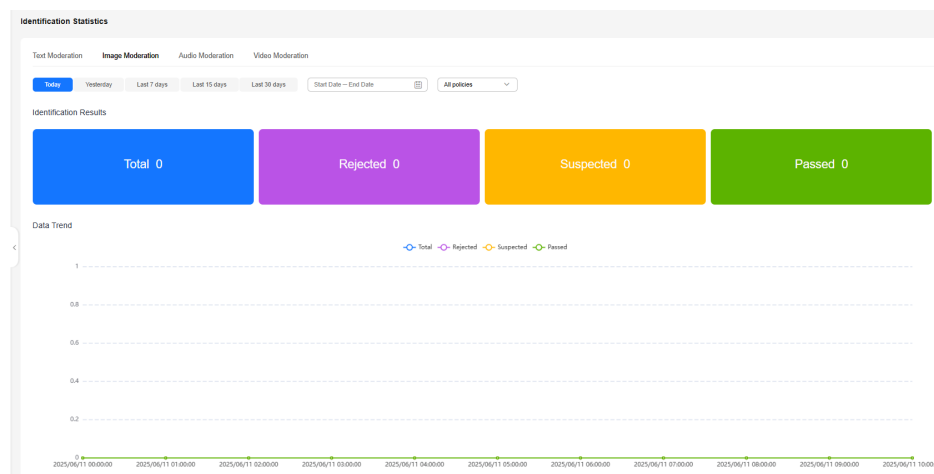To call Image Moderation, perform the following steps:

Step 1 Log in to the management console and choose **Service List** > **Content Moderation**.

Step 2 In the left navigation pane, choose **Content Moderation V3** > **Identification Statistics**. On the **Image Moderation** page, you can view the identification statistics. You can set a time range and select a policy (event type) to view the change of the number of API calls within the time range.

Figure 3-7 Identification statistics

- **Identification Statistics**: total number of calls to Content Moderation, number of rejected calls, number of suspected calls, and number of passed calls within a specified period of time, helping you better learn the calls and moderation status of Content Moderation.
  - **Total**: total number of calls made to Content Moderation
  - **Rejected**: number of calls made to Content Moderation that are rejected because the text contains sensitive information
  - **Suspected**: number of calls made to Content Moderation that require manual review
  - **Passed**: number of calls made to Content Moderation that are approved
- **Data Trend**: trend of the total number of calls, number of rejected calls, number of suspected calls, and number of passed calls within the specified period of time

Figure 3-8 Causes

Causes of Rejected Data   Causes of Suspected Data

porn(14.31%)
politics(37.1%)
ban(28.45%)
abuse(2.83%)
ad(8.66%)
terrorism(7.59%)
customized(1.06%)

No data available.

Subscription Time: May 27, 2021 17:23:39 GMT+08:00

- **Causes of Rejected Data**: proportion of detection scenarios that fail to pass the review within the specified period of time
- **Causes of Suspected Data**: proportion of detection scenarios that require manual review within the specified period of time

----End

4 Using an SDK to Call Content Moderation

4.1 Moderation (Text)

This section provides an example of how to use a Java SDK to call Moderation (Text). You can directly call APIs to use SDK functions. The procedure is as follows:

Step 1 **Subscribe to the service**. In the left navigation pane, choose **Service List** or **Service Management**, and subscribe to Moderation (Text).

Step 2 **(Optional) Configure allowlist and blocklist dictionaries.**

Step 3 **Configure the Java development environment** and development tools.

Step 4 **Obtain the sample code**, SDK, and sample project, and import them to the development environment.

Step 5 **Perform authentication** using AK/SK.

Step 6 **Call the service API** and view the status code and error code whenever necessary.

----End

Subscribing to the Service

Subscribe to Moderation (Text) by following the instructions in **Subscribing to the Service**.

(Optional) Configuring Custom Dictionaries

You can configure allowlist and blocklist dictionaries for text moderation.

For details, see

Configuring an Environment

You can compile code to call the API of Moderation (Text) based on the SDK. You need to configure the environment when using the SDK and calling APIs. The procedure is as follows:

1. Obtain the latest Content Moderation SDK software package and document from the **SDK Center**.

2. For details about how to configure the environment, see **Preparing a Java Development Environment**.

3. Obtain the SDK. For details, see **Obtaining and Installing the Java SDK**.

Obtaining Sample Code

Obtain the Java SDK sample of Text Moderation V3.

1. Log in to the **API Explorer** and click **Java** on the **Sample Code** tab page, as shown in the following figure.

   Figure 4-1 Obtaining the Java SDK
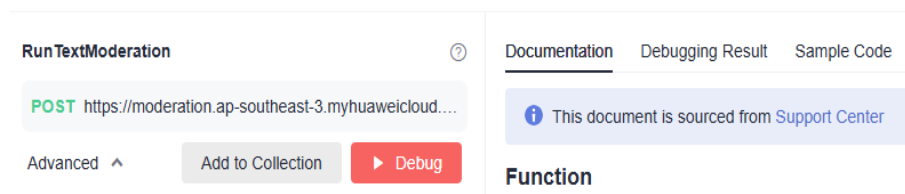
2. Set request body parameters, including **event_type** and **data.text**, as shown in the following figure.

Figure 4-2 Request body parameters

3. Copy the sample code to the Maven project.

Performing Authentication

Content Moderation supports token-based and AK/SK-based authentication. This section uses AK/SK-based authentication as an example.

1. Obtain an AK/SK.

   The AK/SK is the access key. To obtain the AK/SK, log in to the **My Credentials** page, choose **Access Keys** in the left navigation pane, and click **Create Access Key** in the right pane.
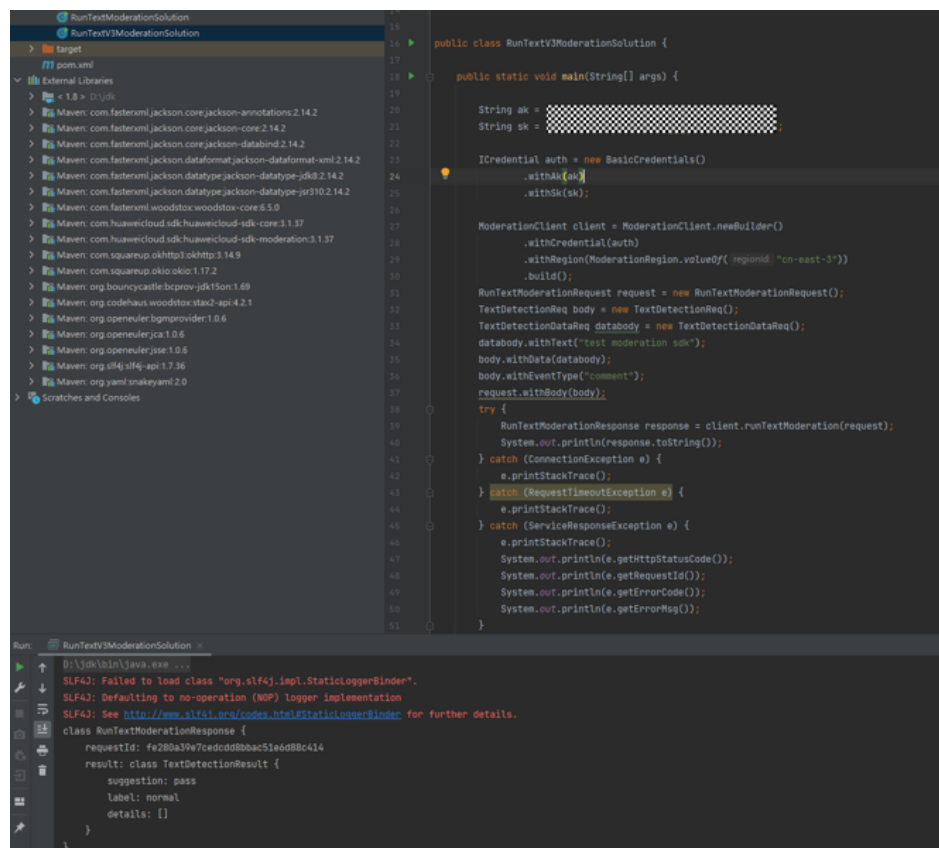
Figure 4-3 Creating an access key

2. Configure the AK/SK in the Java SDK for AK/SK authentication.

Replace the AK/SK parameters in the following figure.

```java
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);
```

Calling a Service

1. Run the sample code to obtain the recognition result, as shown in the following figure.

2. View the number of API calls. In the left navigation pane, choose **Content Moderation V3** > **Identification Statistics**. On the **Text Moderation** page, you can view the call details and statistics.

   – **Identification Statistics**: total number of calls to Content Moderation, number of rejected calls, number of suspected calls, and number of passed calls within a specified period of time, helping you better learn the calls and moderation status of Content Moderation.

     ▪ **Total**: total number of calls made to Content Moderation

     ▪ **Rejected**: number of calls made to Content Moderation that are rejected because the text contains sensitive information

- **Suspected**: number of calls made to Content Moderation that require manual review

- **Passed**: number of calls made to Content Moderation that are approved

– **Data Trend**: trend of the total number of calls, number of rejected calls, number of suspected calls, and number of passed calls within the specified period of time

– **Causes of Rejected Data**: proportion of detection scenarios that fail to pass the review within the specified period of time

– **Causes of Suspected Data**: proportion of detection scenarios that require manual review within the specified period of time

4.2 Moderation (Image)

This section provides an example of how to use a Java SDK to call Moderation (Image). You can directly call APIs to use SDK functions. The procedure is as follows:

Step 1 **Subscribe to the service**. In the left navigation pane, choose **Service List** or **Service Management**, and subscribe to Moderation (Image).

Step 2 **(Optional) Configure allowlist and blocklist dictionaries.**

Step 3 **Configure the Java development environment** and development tools.

Step 4 **Obtain the sample code**, SDK, and sample project, and import them to the development environment.

Step 5 **Perform authentication** using AK/SK.

Step 6 **Call the service API** and view the status code and error code whenever necessary.

----End

Subscribing to the Service

Subscribe to Moderation (Image) by following the instructions in **Subscribing to the Service**.

(Optional) Configuring Custom Dictionaries

You can configure allowlist and blocklist dictionaries for image moderation.

For details, see

Configuring an Environment

You can compile code to call the API of Moderation (Text) based on the SDK. You need to configure the environment when using the SDK and calling APIs. The procedure is as follows:

1. Obtain the latest Content Moderation SDK software package and document from the **SDK Center**.

2. For details about how to configure the environment, see **Preparing a Java Development Environment**.

3. Obtain the SDK. For details, see **Obtaining and Installing the Java SDK**.

Obtaining Sample Code
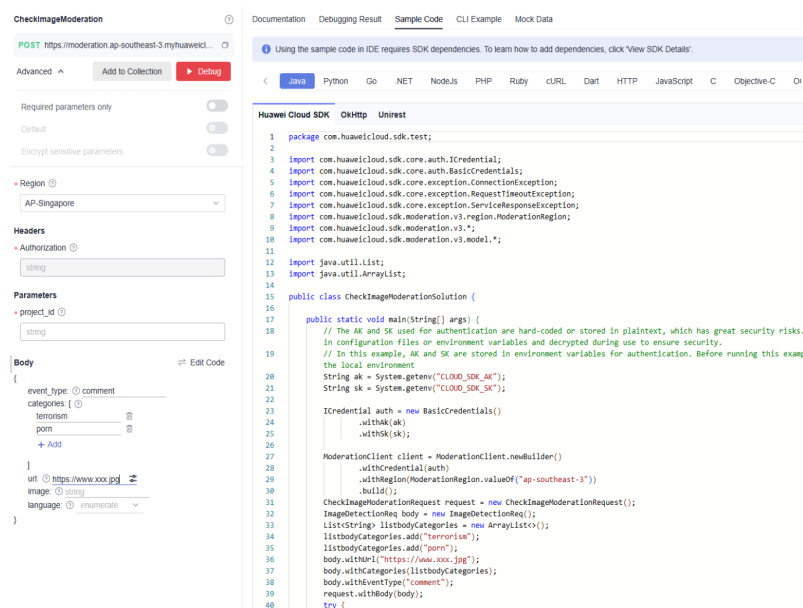
Obtain the Java SDK sample of Text Moderation V3.

1. Log in to the **API Explorer** and click **Java** on the **Sample Code** tab page, as shown in the following figure.

Figure 4-4 Obtaining the Java SDK
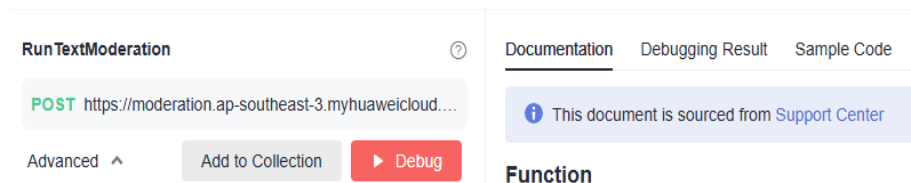
2. Set request body parameters, including **event_type** and **data.text**, as shown in the following figure.

Figure 4-5 Request body parameters

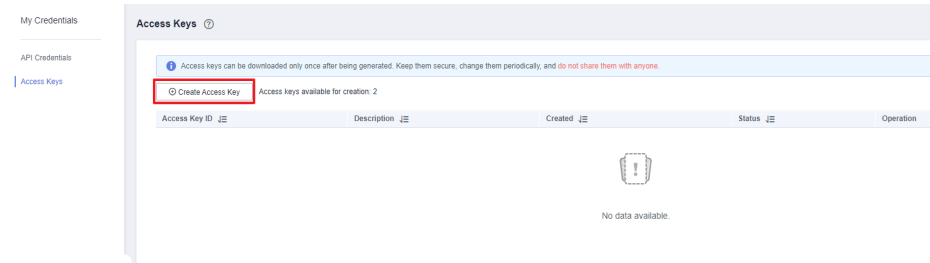3. Copy the sample code to the Maven project.

Performing Authentication

Content Moderation supports token-based and AK/SK-based authentication. This section uses AK/SK-based authentication as an example.

1. Obtain an AK/SK.

The AK/SK is the access key. To obtain the AK/SK, log in to the **My Credentials** page, choose **Access Keys** in the left navigation pane, and click **Create Access Key** in the right pane.

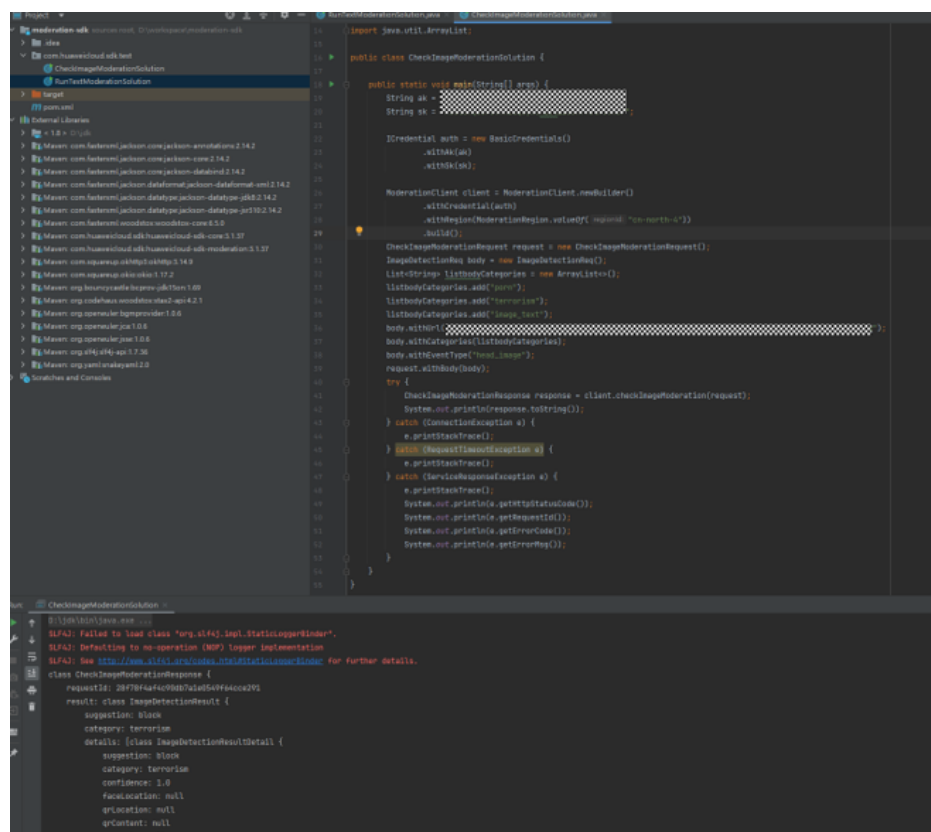**Figure 4-6** Creating an access key

2. Configure the AK/SK in the Java SDK for AK/SK authentication.

   Replace the AK/SK parameters in the following figure.

```java
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);
```

Calling a Service

1. Run the sample code to obtain the recognition result, as shown in the following figure.

2. View the number of API calls. Choose **Service List > Image Moderation**, and view call statistics and call details. See **Figure 4-7**.

Figure 4-7 Identification statistics

– **Identification Statistics**: total number of calls to Content Moderation, number of rejected calls, number of suspected calls, and number of passed calls within a specified period of time, helping you better learn the calls and moderation status of Content Moderation.

■ **Total**: total number of calls made to Content Moderation

■ **Rejected**: number of calls made to Content Moderation that are rejected because the text contains sensitive information

■ **Suspected**: number of calls made to Content Moderation that require manual review

■ **Passed**: number of calls made to Content Moderation that are approved

– **Data Trend**: trend of the total number of calls, number of rejected calls, number of suspected calls, and number of passed calls within the specified period of time

Figure 4-8 Causes

– **Causes of Rejected Data**: proportion of detection scenarios that fail to pass the review within the specified period of time

– **Causes of Suspected Data**: proportion of detection scenarios that require manual review within the specified period of time