

Content Moderation User Guide

Issue 01
Date 2023-01-03



Copyright © Huawei Technologies Co., Ltd. 2023. All rights reserved.

No part of this document may be reproduced or transmitted in any form or by any means without prior written consent of Huawei Technologies Co., Ltd.

Trademarks and Permissions



HUAWEI and other Huawei trademarks are trademarks of Huawei Technologies Co., Ltd.

All other trademarks and trade names mentioned in this document are the property of their respective holders.

Notice

The purchased products, services and features are stipulated by the contract made between Huawei and the customer. All or part of the products, services and features described in this document may not be within the purchase scope or the usage scope. Unless otherwise specified in the contract, all statements, information, and recommendations in this document are provided "AS IS" without warranties, guarantees or representations of any kind, either express or implied.

The information in this document is subject to change without notice. Every effort has been made in the preparation of this document to ensure accuracy of the contents, but all statements, information, and recommendations in this document do not constitute a warranty of any kind, express or implied.

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

Contents

1 Process of Using Content Moderation.....	1
2 Subscribing to the Service.....	3
3 Preparing Data.....	5
4 (Optional) Configuring Custom Dictionaries.....	6
5 Calling APIs or SDKs.....	8
5.1 Calling APIs or SDKs Locally.....	8
6 Viewing the Number of Calls.....	11
7 Viewing Metrics.....	13

1 Process of Using Content Moderation

Content Moderation adopts image and text detection technologies that detect pornography and images and text violating related laws or regulations. By calling the APIs provided by Content Moderation, you can have your uploaded images, text, and audio reviewed and obtain the inference results. This helps you build an intelligent system that delivers improved efficiency.

The following figure shows the procedure for using Content Moderation.

Figure 1-1 Procedure for using Content Moderation

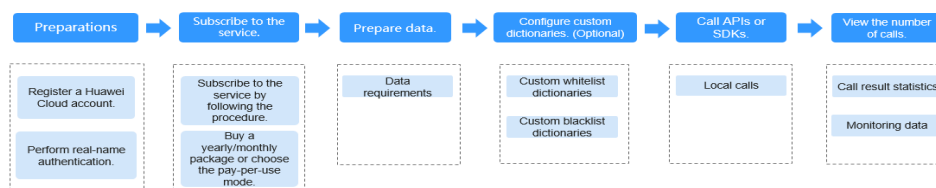


Table 1-1 Procedure for using Content Moderation

Procedure	Sub Task	Description	Instruction
Preparations	Registering an account with Huawei Cloud	Before using Content Moderation, register a Huawei Cloud account.	2-Registering a Huawei Cloud Account
Subscribing to the service	Subscribing to the Content Moderation service by following the procedure	You need to subscribe to the service according to the procedure.	Subscribing to the Content Moderation Service

Procedure	Sub Task	Description	Instruction
	Buying a yearly/ monthly package or choosing the pay-per-use billing mode	After subscribing to the service, you need to purchase the service. Two billing modes are available.	Purchasing the Service
Preparing data	Data requirements	There are restrictions on the data format and the number of concurrent calls. Before using the service, you need to prepare the data to be reviewed by referring to the restrictions.	Preparing Data
(Optional) Configuring custom dictionaries	Custom whitelist and blacklist dictionaries	When using the Text Moderation service, you can configure a custom whitelist dictionary or blacklist dictionary to filter and detect specified text.	(Optional) Configuring Custom Dictionaries
Calling APIs or SDKs	Local calls	Use the moderation SDK for local development. You can directly call functions to use SDKs.	Calling APIs or SDKs Locally
Viewing the number of calls	Viewing call result statistics	You can view the review details and the number of API calls on the management console.	Viewing the Number of Calls
	Viewing monitoring data	You can view the Content Moderation monitoring metrics on the Cloud Eye console or by using the APIs provided by Cloud Eye.	Viewing Metrics

2 Subscribing to the Service

This section introduces the procedure for subscribing to the service.

NOTE

This service is available only to enterprise users for now.

Registering a Huawei Cloud Account

Skip this step if you already have registered one.

1. Log in to the [Huawei Cloud](#) official website.
2. Click **Register** in the upper right corner to access the registration page.
3. Complete the registration as instructed. For details, see [Account Registration Process](#).

Subscribing to the Content Moderation Service

To subscribe to Content Moderation, perform the following steps:

1. Register a Huawei Cloud account and perform 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 lower right corner of the page, click **Customer Service** to subscribe to the Content Moderation (Text) service.
4. After a commercial service is successfully subscribed, the service is displayed in **My Services** on the **Service Management** page. In this case, you can call the API to use Content Moderation.

Billing Modes

Content Moderation supports both the pay-per-use and yearly/monthly billing modes. For details, see [Billing](#).

- Pay-per-use

If you want to use the pay-per-use billing mode, see [Content Moderation Pricing Details](#).

- Yearly/Monthly
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.

3 Preparing Data

The functions of Content Moderation vary depending on the region. There are restrictions on the data format and the number of concurrent calls. Before using the service, you need to prepare the data to be reviewed by referring to the restrictions.

For details about the restrictions on using service functions, see [Constraints](#).

For example, Text Moderation has the following restrictions on the input data:

- It is available in **CN-Hong Kong**, **AP-Singapore**, and **LA-Santiago**.
- It only supports Chinese text.
- By default, the maximum number of concurrent API calls is 50. To increase the concurrency, [submit a service ticket](#).

4 (Optional) Configuring Custom Dictionaries

Function

Before using the **Text Moderation** service, you can configure a custom whitelist dictionary or blacklist dictionary to filter and detect specified text.

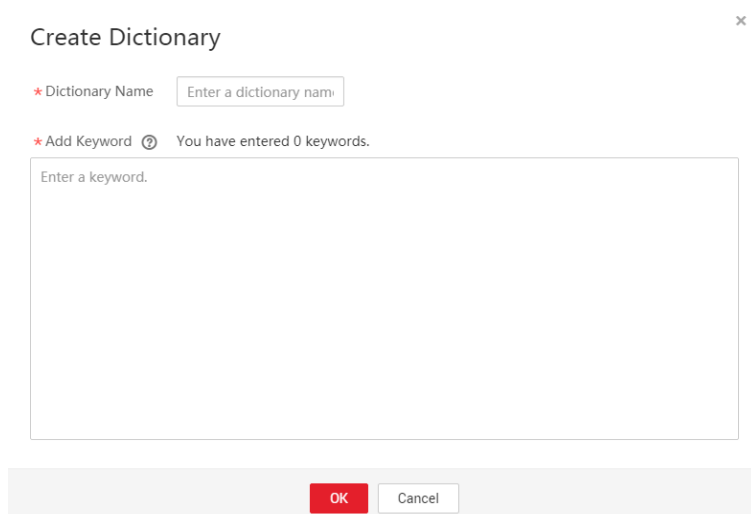
NOTE

1. You can use custom dictionaries for free.
2. A maximum of 10 custom blacklist dictionaries and 10 custom whitelist dictionaries can be created for an account.
3. Custom whitelist dictionaries cannot be used in the flood scenario.

Procedure


1. Log in to the Content Moderation console. In the left navigation pane, choose **Settings > Custom Dictionary**.
2. On the displayed page, click the **Blocklist Libraries** or **Allowlist Libraries** tab and then **Create Dictionary**.

Figure 4-1 Creating a dictionary



Create Dictionary ×

* Dictionary Name

* Add Keyword  You have entered 0 keywords.

3. In the displayed dialog box, enter a dictionary name and keywords.
 - **Dictionary Name:** A dictionary name contains a maximum of 40 characters. Only letters, digits, hyphens (-), and underscores (_) are allowed.
 - **Add Keyword:** A dictionary contains a maximum of 10,000 keywords. Each keyword consists of a maximum of 50 characters and ends with a carriage return.
4. Click **OK**.

 **NOTE**

- The dictionary name cannot be changed.
- After a whitelist dictionary is configured, the whitelist takes effect by default in any detection scenario.
- The created dictionary can be used in the request parameter **categories**. Setting **categories** to the dictionary name is setting a custom scenario.

5 Calling APIs or SDKs

5.1 Calling APIs or SDKs Locally

Content Moderation Software Development Kit (Moderation SDK for short) encapsulates the RESTful APIs provided by Content Moderation to simplify application development. You can add dependencies or download SDKs to call APIs to use Content Moderation.

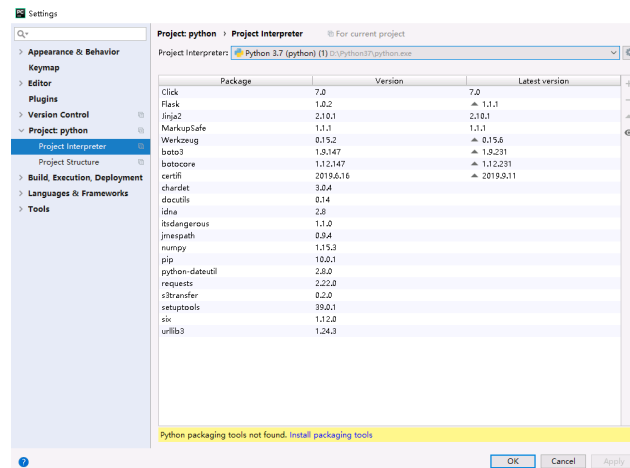
This section uses **Text Moderation** as an example to describe how to use the Moderation Python SDK for local development. You can use the SDK by calling API functions.

Prerequisites

- You have registered an account with Huawei Cloud and completed real-name authentication. Your account cannot be in arrears, frozen, or deregistered.
- You have learned the [constraints on Text Moderation](#).
- You have [subscribed to the Text Moderation service](#).

Procedure

1. Install the Python environment and obtain the SDK.
 - a. Download Python of a proper version from [Python's official website](#) and install it. Python 3.3 or later is recommended. This section uses Python 3.7 as an example.
 - b. Download the latest version of PyCharm from [PyCharm's official website](#).
 - c. Start the PyCharm development tool and choose **File > Settings > Project Interpreter** to configure the Python environment.
 - d. Select the Python installation path. See [Figure 5-1](#). After selecting the target Python, click **Apply** at the bottom of the page to complete the configuration.

Figure 5-1 Configuring the python environment using PyCharm

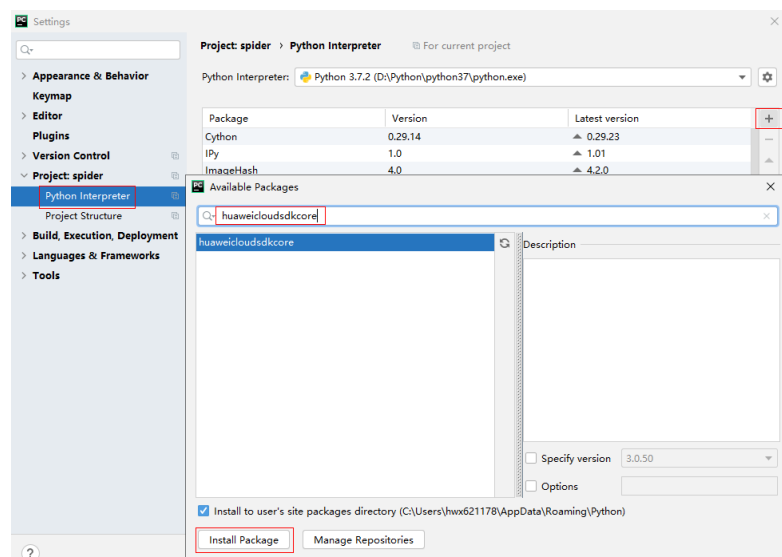
2. Create a project in PyCharm and click **Terminal** in the lower left corner. Run the following commands to install the SDK (the SDK supports Python 3 or later):

Install the SDK using pip commands:

```
# Install the core library.
pip install huaweicloudsdkcore
```

```
# Install the moderation service library.
pip install huaweicloudsdkmoderation
```

On PyCharm, choose **File > Settings > Project > Python Interpreter**. Click **+** in the upper right corner, search for **huaweicloudsdkcore** and **huaweicloudsdkmoderation** respectively, and click **Install Package** in the lower left corner to install them.

Figure 5-2 Installing the Python SDK for Content Moderation using PyCharm

3. Copy the SDK sample code of Text Moderation to PyCharm as follows:

```
# coding: utf-8
from huaweicloudsdkcore.auth.credentials import BasicCredentials
from huaweicloudsdkmoderation.v2.region.moderation_region import ModerationRegion
from huaweicloudsdkcore.exceptions import exceptions
from huaweicloudsdkmoderation.v2 import *
```

```

if __name__ == "__main__":
    //Enter your AK/SK.
    ak = "<YOUR AK>"
    sk = "<YOUR SK>"
    credentials = BasicCredentials(ak, sk) \
    client = ModerationClient.new_builder() \
        .with_credentials(credentials) \
        .with_region(ModerationRegion.value_of("ap-southeast-1")) \
        .build()
try:
    request = RunTextModerationRequest()
    listItemsbody = [
        TextDetectionItemsReq(
            text="asdfasdf" //Enter the text to be detected, for example, asdfasdf.
        )
    ]
    request.body = TextDetectionReq(
        items=listItemsbody
    )
    response = client.run_text_moderation(request)
    print(response)
except exceptions.ClientRequestException as e:
    print(e.status_code)
    print(e.request_id)
    print(e.error_code)
    print(e.error_msg)

```

4. Obtain the AK and SK and replace <YOUR AK> and <YOUR SK> in the sample code with the AK and SK, respectively.

Log in to the [My Credentials](#) page, choose **Access Keys** in the navigation pane on the left, and click **Create Access Key** in the right pane.

5. Run the sample code to obtain the recognition result. You can interpret the review result based on the response parameter description. For details, see [Text Moderation Result](#).

Figure 5-3 Example

The screenshot shows an IDE with a Python script named 'TextModeration.py'. The code is identical to the sample code provided in the previous block. The script is executed, and the output in the console is: `{'result': {'suggestion': 'pass', 'detail': {}}}`. The IDE interface includes a project explorer on the left, a code editor in the center, and a run console at the bottom.

6 Viewing the Number of Calls

Function

You can view the moderation details and the number of calls on the Content Moderation console to better understand the moderation status and call statistics.

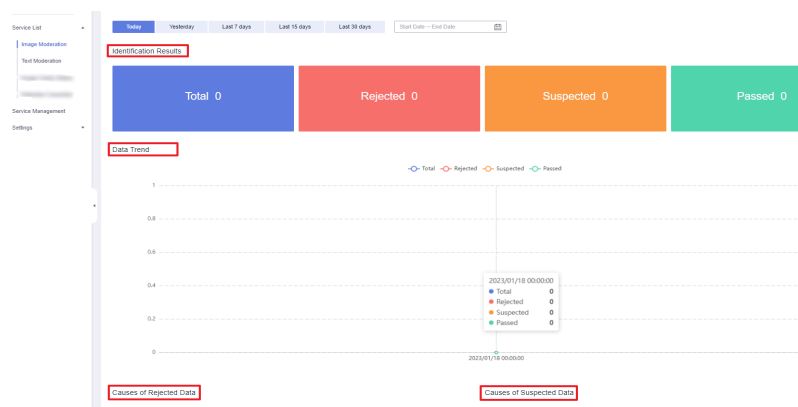
NOTE

This function is applicable to text, image, audio, and video moderation.

Procedure

1. Log in to the Content Moderation management console.
2. In the left navigation pane, choose **Service List** > **Text Moderation** and view identification statistics shown in [Figure 6-1](#). 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 6-1 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
- **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

7 Viewing Metrics

You can view the Content Moderation monitoring metrics on the Cloud Eye console or by using the APIs provided by Cloud Eye.

Namespace

SYS.MODERATION

Content Moderation Metrics

Table 7-1 Content Moderation metrics

Metric ID	Metric Name	Description	Value Range	Monitored Object	Monitoring Period (Raw Metric)
successful_call_times_of_service	Successful Calls of Service	Number of successful calls to the service Unit: calls/min	≥ 0 times/min	Content Moderation APIs	1 min
failed_call_times_of_service	Failed Calls of Service	Number of failed calls to the service Unit: calls/min	≥ 0 times/min	Content Moderation APIs	1 min

Viewing Metrics

The following steps use Text Moderation as an example.

1. Log in to the Content Moderation management console.
2. In the left navigation pane, choose **Service List > Text Moderation**. Move the cursor to the bottom of the page and click **View Metric**.

Figure 7-1 Viewing metrics

[View Metric](#)

Subscription Time: Jul 06, 2021 09:32:52 GMT+08:00

3. On the displayed Cloud Eye console, set the time range and view metrics such as Successful Calls of Service and Failed Calls of Service.

Figure 7-2 Viewing metrics

