内容审核

用户指南

文档版本	01
发布日期	2024-04-23

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 服务使用流程

内容审核（Content Moderation），是基于图像、文本的检测技术，可自动检测涉黄、图文违规等内容，用户通过调用API对上传的图片、文字、音视频进行内容审核，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。

使用本服务的操作流程如下所示：

图 1-1 使用流程

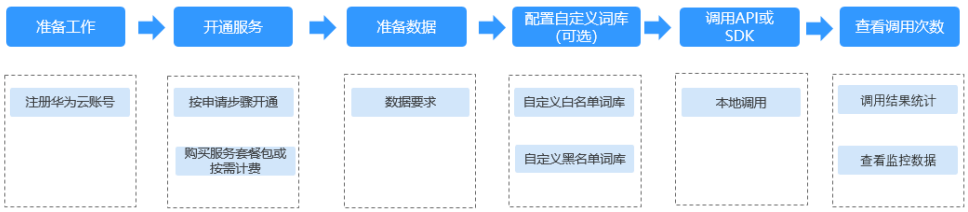


表 1-1 使用流程说明

流程	子任务	说明	详细指导
准备工作	注册华为账号	使用内容审核服务之前，您需要注册华为账号。	**注册华为账号**
开通服务	按申请步骤开通	需要您按照步骤操作说明来申请开通服务。	**开通服务**
	购买服务套餐包或按需计费	成功开通服务后需要购买服务，有两种计费方式可供选择。	**购买服务**
准备数据	数据要求	数据格式和调用并发数有相应的约束限制，需要您在使用服务前参考约束准备好待审核的数据。	**准备数据**
配置自定义词库（可选）	自定义白名单词库/自定义黑名单词库	使用**文本内容审核**服务，您可以配置自定义白名单词库或自定义黑名单词库，来帮助您过滤和检测指定文本内容。	**配置自定义词库（可选）**

流程	子任务	说明	详细指导
调用API或SDK	本地调用	介绍使用Moderation SDK在本地进行开发，用户直接调用接口函数即可使用SDK功能。	**本地调用**
查看调用次数	调用结果统计	开始使用服务后，可以在管理控制台上查看服务审核详情和调用次数统计。	**查看调用次数**
	查看监控数据	云监控服务提供的管理控制台或API接口来检索内容审核服务产生的监控指标。	**查看监控指标**

2 开通服务

您可以按照如下步骤操作申请开通本服务。

说明

本服务暂时仅面向企业用户开放，个人用户暂不支持开通。

注册华为账号

如果您已完成华为账号注册，可跳过该步骤。

1. 登录**华为云**官方网站。
2. 单击华为云官网右上角“注册”进入注册页面。
3. 在注册页面，根据提示信息完成注册。具体操作可参见**账号注册**。

开通服务

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. 单击页面右下角的“联系客服”按钮，联系客服帮助您开通本服务。
4. 商用服务申请成功后，在“服务管理”页面，“我的服务”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。

计费方式

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**。

- 按需计费

  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**。
- 预付套餐包计费

  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。

3 准备数据

服务不同功能部署的区域，数据格式和调用并发数有相应的约束限制，需要您在使用服务前参考约束准备好待审核的数据。

服务功能的使用约束请参见**约束与限制**。

例如文本内容审核，输入数据存在以下约束：

- 文本内容审核V2版本：支持“中国-香港、亚太-新加坡、拉美-圣地亚哥”区域。
- 文本内容审核V3版本：支持“亚太-新加坡”区域。
- 只支持中文文本内容审核。
- 默认API调用最大并发为50，如需调整更高并发限制请通过**工单**联系专业工程师为您服务。

4 配置自定义词库（可选）

使用**文本内容审核**服务前，您可以配置自定义白名单词库或自定义黑名单词库，来帮助您过滤和检测指定文本内容。

配置自定义词库 V2请看**具体操作**。

5 调用 API 或 SDK

5.1 本地调用

内容审核软件开发工具包（Moderation SDK）是对内容审核提供的REST API进行的封装，以简化用户的开发工作。用户通过添加依赖或下载的方式调用API即可实现使用内容审核业务能力的目的。

本章节以**文本内容审核**为例，介绍如何使用Moderation Python SDK在本地进行开发，用户直接调用接口函数即可使用SDK功能。

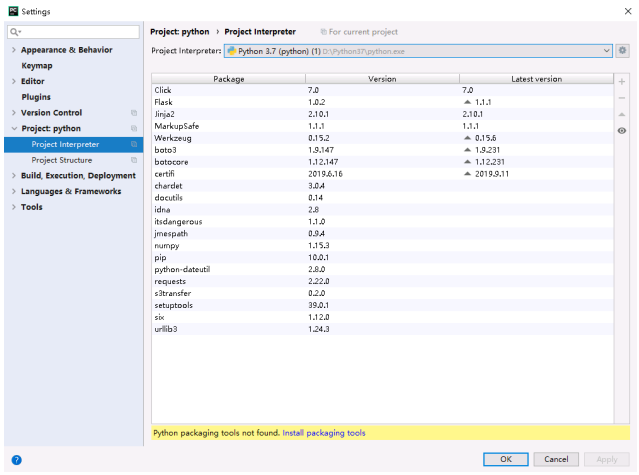
前提条件

- 已注册华为账号，并完成实名认证，账号不能处于欠费、冻结、被注销等异常状态。
- 了解**文本内容审核约束限制**。
- 已**开通文本内容审核服务**。

操作步骤

1. 安装Python环境并获取SDK软件包。
   a. 从**Python官网**下载并安装合适的Python版本。请使用Python3.3以上版本，如下以Python3.7 版本为例进行说明。
   b. 从**PyCharm官网**下载并安装最新版本。
   c. 在PyCharm开发工具中配置Python环境，在菜单依次选择“File > Settings > Project Interpreter”。
   d. 在页面上方选择您的Python安装路径，如图 **PyCharm配置python环境所示**。选择好目标Python之后单击页面下方“Apply”完成配置。

图 5-1 PyCharm 配置 python 环境



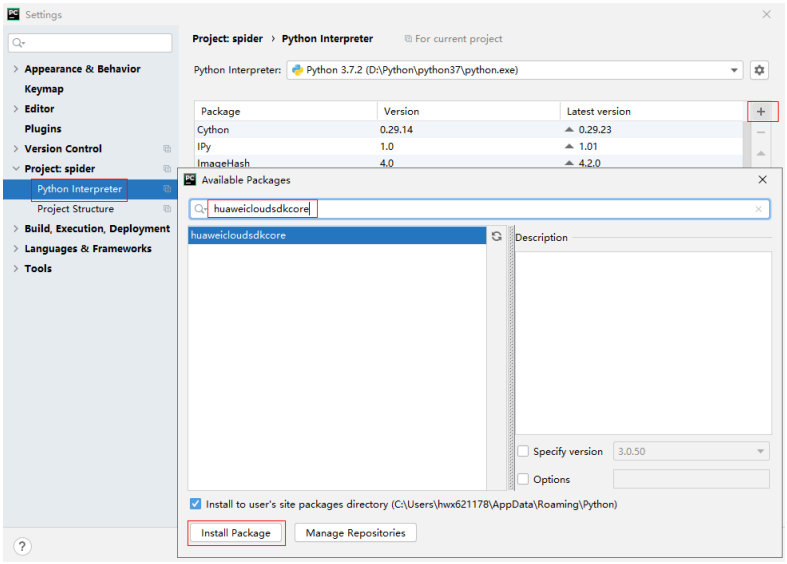
2. 在PyCharm中新建一个项目，并单击左下方“Terminal”按钮。分别执行以下命令安装SDK（该SDK支持Python3及以上版本）。参考方法如下：

```
pip 安装：
# 安装核心库
pip install huaweicloudsdkcore

# 安装Moderation服务库
pip install huaweicloudsdkmoderation
```

在pycharm中，选择“File > Settings > Project > Python Interpreter”单击右上角+，分别搜索huaweicloudsdkcore及huaweicloudsdkmoderation，搜索到包内容单击左下角Install Package完成安装。

图 5-2 pycharm 安装内容审核 python 版本 sdk 包



3. 复制文本审核SDK示例代码到PyCharm中，如下所示：

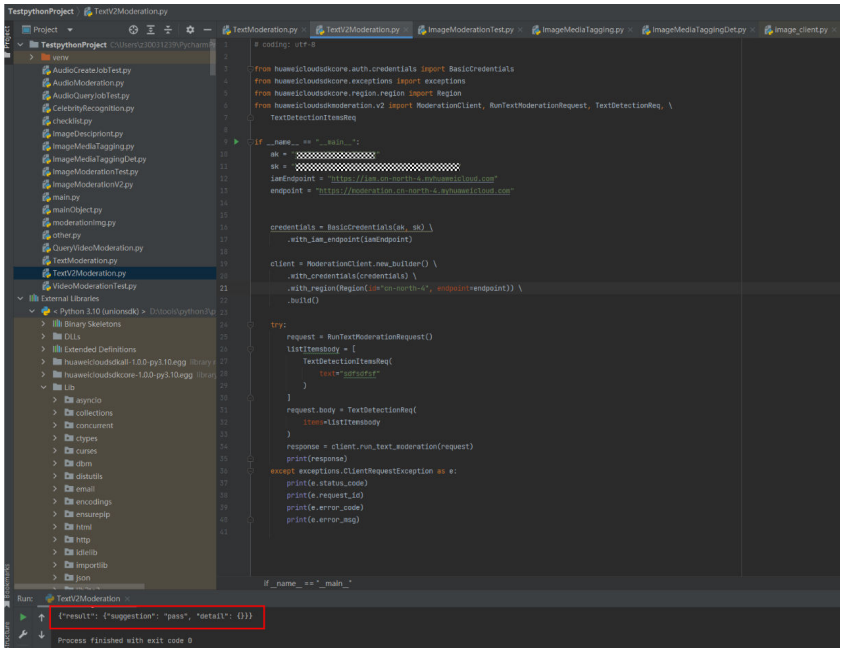
```
# coding: utf-8
from huaweicloudsdkcore.auth.credentials import BasicCredentials
from huaweicloudsdkmoderation.v2.region.moderation_region import ModerationRegion
from huaweicloudsdkcore.exceptions import exceptions
from huaweicloudsdkmoderation.v2 import *
if __name__ == "__main__":
    //此处需要输入您的AK/SK信息
```



```
    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="asdfasdfasdf"   //此处输入待检测文本，以asdfasdfasdf为例
            )
        ]
        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. 获取AK/SK，替换代码示例中的“<YOUR AK>”、“<YOUR SK>”参数。
   登录**我的凭证**界面，选择“管理访问密钥 > 新增访问密钥”获取。
5. 运行代码示例，获取识别结果。您可根据响应参数说明来解读审核结果的含义，具体可参考**文本内容审核结果**。

图 5-3 运行示例



6 查看调用次数

功能介绍

您可以在内容审核服务管理控制台上查看服务审核详情和调用次数统计，帮助您更好了解服务的审核情况和调用情况。

说明

该功能适用于文本/图像/音频/视频审核。

操作步骤

1. 登录内容审核服务管理控制台。
2. 在左侧导航栏中选择“服务列表>文本审核”，可以查看识别统计详情，如图6-1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 6-1 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。
- 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。

7 查看监控指标

您可以通过云监控服务提供的管理控制台或API接口来检索内容审核服务产生的监控指标。

命名空间

SYS.MODERATION

内容审核监控指标

表 7-1 内容审核支持的监控指标

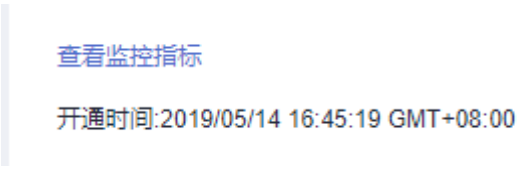
指标ID	指标名称	指标含义	取值范围	测量对象	监控周期（原始指标）
successful_call_times_of_service	调用内容审核成功次数	该指标用于统计调用服务成功次数。 单位：次/分钟	≥ 0 times/min	内容审核接口	1分钟
failed_call_times_of_service	调用内容审核失败次数	该指标用于统计调用服务失败次数。 单位：次/分钟	≥ 0 times/min	内容审核接口	1分钟

查看监控指标

以文本内容审核为例。

1. 登录内容审核服务管理控制台。
2. 在左侧导航栏中选择“服务列表>文本审核”，滑动鼠标至页面低端，单击“查看监控指标”，如图7-1所示。

图 7-1 查看监控指标



3. 进入云监控服务控制台，选择时间范围即可查看服务调用成功和失败的次数等历史数据。

图 7-2 监控数据