内容安全

产品概述

产品文档

腾讯云

文档目录

产品概述
产品介绍

最近更新时间：2019-08-09 10:08:11

什么是内容安全

内容安全是腾讯云提供的过滤网站和 App 不良信息的安全服务，包括文本识别和图片识别两个子服务，帮助客户告别大规模的人工识别，在节约成本的同时，提高识别效率，净化网络环境，增强产品体验，保护业务健康发展。

文本识别

产品概述

在互联网蓬勃发展的过程当中，用户生成了海量的文本内容，其中不可避免地会夹杂一些垃圾文本，轻则影响产品体验，重则违反法律法规，甚至会导致关站。这些垃圾文本，大都是由专业的互联网黑产从业者制造出来的。为了帮助客户免受垃圾文本内容的困扰和牵制，从与黑产繁杂的对抗中解放出来，天御推出了文本识别服务。
文本识别服务应用于网站和 App 的内容场景，包括发帖、评论、弹幕、聊天等，可准确识别这些场景的恶意文字和图片内容，包括违法违规、垃圾广告、恶意营销等，有效打击各类违法违规广告和色情低俗内容。

产品功能

反干扰

由于产品安全策略的打击，黑产团伙现在多通过符号、图标、拼音和外文等方式进行文本干扰。传统的文本挖掘对抗方式，已经很难有效识别业务中的“牛皮癣”广告。天御研发出反干扰引擎这一创新方式来扫除黑产的恶意广告，该引擎通过文本预处理和 OCR 识别的方式，将内容中的符号、图标等干扰内容进行转换，还原文本内容，轻松识破黑产团伙的“障眼法”。

脏词库

在清晰还原干扰文本的基础上，天御提供了最快最全面的脏词扫描技术，基于腾讯生态积累的业界最全的脏词库和 Tree 树多模型匹配技术，天御文本识别服务可以最快最全面的识别出恶意文本和广告内容。天御脏词库包括 30 多个类别和几十万的关键词，从主流的色情、赌博到生僻的开锁、茶叶等广告均能全面覆盖。

黑产特征

- 文本特征：天御的文本特征主要包括获利特征、引流文本特征、噪音文本特征等内容。利用这些文本特征，加上天御的文本挖掘能力，能够发现可疑的用户内容。除需提取引流、噪音等较为直观的特征外，天御还采用了 LDA 来自动发掘更多的特征。

- 图片引流识别：图片引流作为一种高门槛的对抗行为，经常被黑产团伙利用，目前天御文本识别接口中已经集成了 OCR 图片识别能力，该能力已经在 QQ/Qzone 等腾讯产品运用多年，可以对广告、色情类图片发挥良好的识别效果。

风险等级

天御文本识别服务对用户产生的内容识别分为恶意、可疑和可信三种风险等级。建议客户根据不同的风险等级，采用不同的策略进行处理。

- 可信：正常使用业务。
- 可疑：建议结合该用户行为上的特征（包括 IP 聚集、内容频繁等）进行打击。
- 恶意：建议直接对该用户进行业务限制。

风险通知

如果用户产生的内容被识别为恶意风险等级，天御文本识别服务会明确的告知客户存在的恶意内容，包括命中关键词和恶意类别。

产品优势

优势	优势说明
技术强	十万量级的关键词，超强抗干扰的滤噪技术和干扰手段还原技术，LDA 文档主题特征提取技术等文本层面的识别技术配以 URL 识别，黄图识别和图片 OCR 技术，使得在不依赖于 IP、手机号等辅助信息的情况下,也具有较高的恶意文本识别率。
数据广	通过用户举报等途径，腾讯积累了多年的海量黑产数据库，如黑 IP 库，黑手机库，黑 QQ 库，黑微信号库等。不论是重新换个 IP，还是重新换个手机号、QQ 号，都逃不过天御文本识别服务布下的天罗地网。
多维度	通过手机号、邮箱、QQ、微信号等多个维度，精准识别黑产帐号等级，深度打击恶意广告行为。
低成本	每 100 次查询最低 0.3 元，只需低廉的成本即可抵御恶意的风险。
快速	依托腾讯云的先进架构，服务毫秒级响应，每秒超过万级并发，更支持动态扩容，使您无需担心性能损耗。
便捷	您无需安装任何脚本文件，通过 API 方式即可直接使用，只需三步轻松接入，支持非腾讯云客户使用。
准确	文本识别服务恶意识别率高于 95%，国内最大的同城网站和最早的社区网站均使用此项服务，成功阻挡海量恶意广告。

图片识别

产品概述

天御图片识别服务基于腾讯优图的深度图片鉴黄技术，可以高效准确地鉴别色情图片和性感图片，解决网站和 App 开发者的鉴黄难题，告别大规模的人工鉴黄，在节约成本的同时，大幅提高鉴黄效率，净化网络环境。

产品功能

色情识别

天御图片识别服务通过开发者提供的图片地址，自动下载并进行鉴黄识别，通过置信分数的方式返回给客户，83 分以上判定为色情可疑图片，开发者可以对这部分图片的用户行为，进行跟踪分析。

性感识别

天御图片识别服务针对开发者提供的图片，免费进行性感度识别，当性感度超过 83 分时，开发者可以对这部分图片的用户行为，进行跟踪分析。

自动鉴黄

天御图片识别服务使用腾讯优图的 DeepEye 主动色情识别技术引擎，对图片进行色情置信度分析，依托腾讯社交的海量图片样本优势进行深度识别训练，算法识别准确率达到 99.9% 以上，远超人工识别水平，实际工作中可以取代 90% 人力，而且针对图片自动识别领域最难的色情与性感的界定问题，鉴黄引擎采用了分离图谱技术，精准识别性感和色情的差别。

产品优势

优势	优势说明
准确	自动色情图片识别准确率超过 99%，精确区分正常、色情及性感图片，方便业务发展。
节约人力	图片识别已经为多家直播平台提供自动鉴黄服务，最多减少 90% 的人力审核。
置信度	图片识别服务会提供置信度数据，使您精准区分正常、色情及性感图片，助力业务健康发展。
低成本	每 100 次查询最低 0.1 元，只需低廉的成本即可抵御色情风险。
快速	依托腾讯云的先进架构，服务毫秒级响应，每秒超过万级并发，更支持动态扩容，使您无需担心性能损耗。
便捷	您无需安装任何脚本文件，只需通过 API 方式，三步轻松接入，即可使用，支持非腾讯云客户使用。

应用场景

社区论坛

内容安全可以广泛应用于 BBS、博客等有用户 UGC 内容的各类网站，包括发帖、回帖、站内信等场景。应用智能文本反垃圾技术，实时检测文本中的涉黄、广告、灌水、谩骂等垃圾文本。

直播鉴黄

天御和腾讯视频云为直播客户提供视频截图、存储、图片自动识别的整体解决方案，您只需要将直播服务部署在腾讯云上，就可以一键开启自动鉴黄服务。
以直播行业为例，在部署了天御图片识别服务后，主播们视频直播的内容先经过腾讯云视频鉴黄服务部署的环境，再呈现到观看者的屏幕。也就是说，直播同时，天御能对视频截图进行图片识别，做到第一时间截图、发现和通知回调业务处理，更可以根据不同时间段对不同房间进行不同的截图频率鉴别，对重点房间进行监控。

IM

内容识别可以对昵称、头像、签名、C2C 消息、群发消息等藏匿的垃圾内容进行针对性地检测识别，防止恶意用户的骚扰，预防风险诈骗。