LiveData 文本审核控制台使用手册

1、数据统计

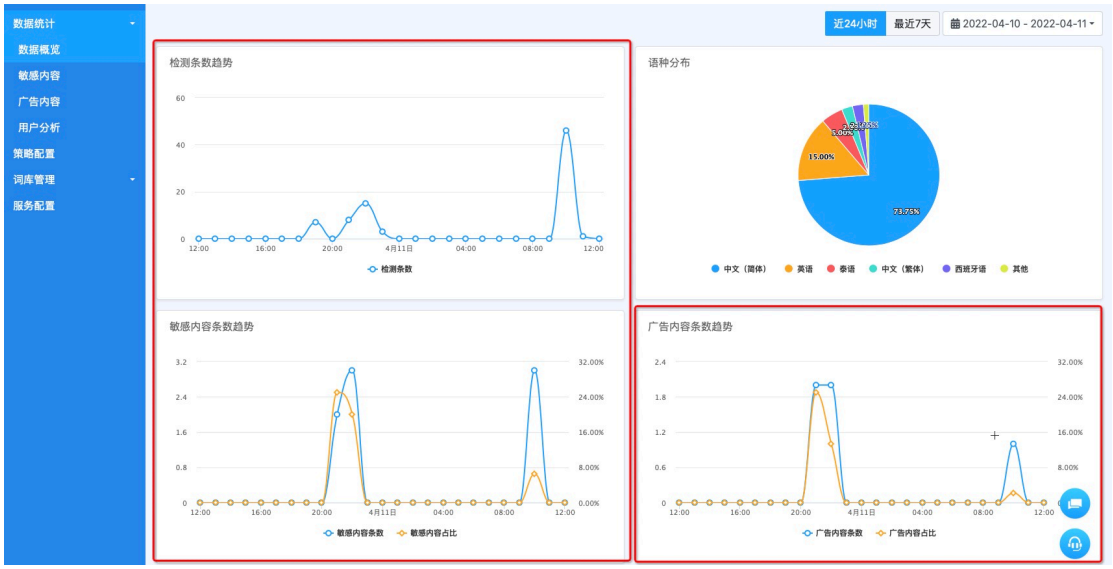
1.1 数据概览

您可以在此总览该项目一段时间内检测趋势变化、检测语种分布、检出类型分布情况。

1.1.1 今日概览

此处展示当日 0 时至当前时刻的检测条数、敏感内容条数、广告内容条数、敏感内容占比、广告内容占比以及各数据项相较于上一自然日的环比变化情况。

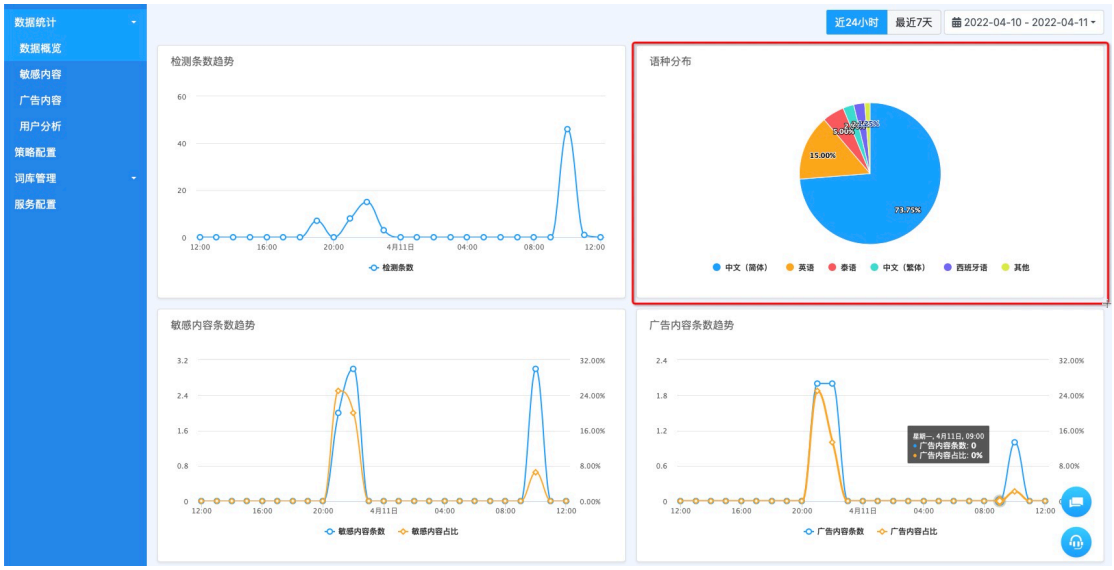
1.1.2 检测条数&敏感内容条数&广告内容条数趋势



以上三个模块以折线统计图形式展示您筛选时间范围内检测条数、敏感内容条数、广告内容条数随时间的数据变化趋势。

您可通过移动鼠标到折线上某一个点的方式，查看对应时间点的具体数据。

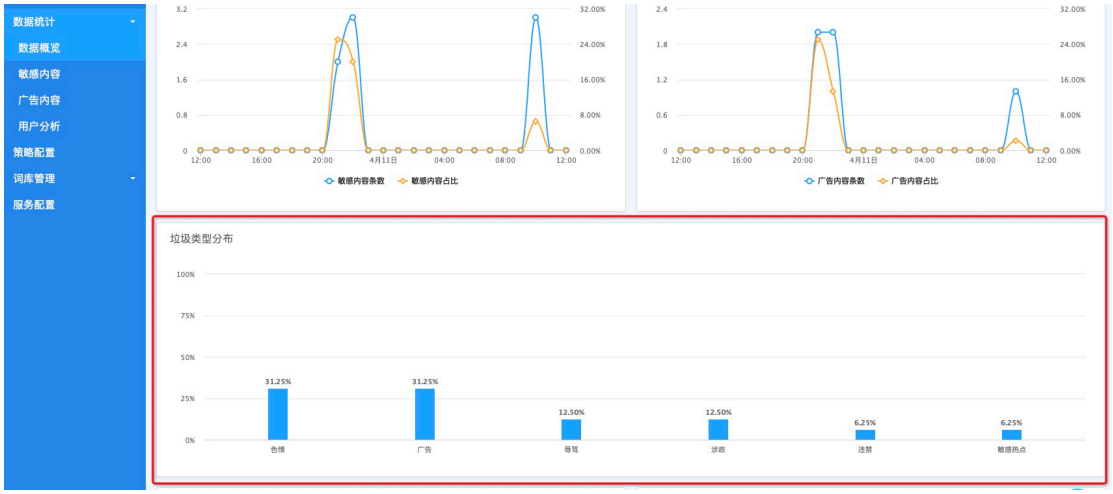
1.1.3 语种分布



此处以饼图的形式展示您筛选时间范围内，检测文本的语种分布情况。

您可通过移动鼠标到饼图上某一扇型区域的方式，查看对应语种的具体检测条数。

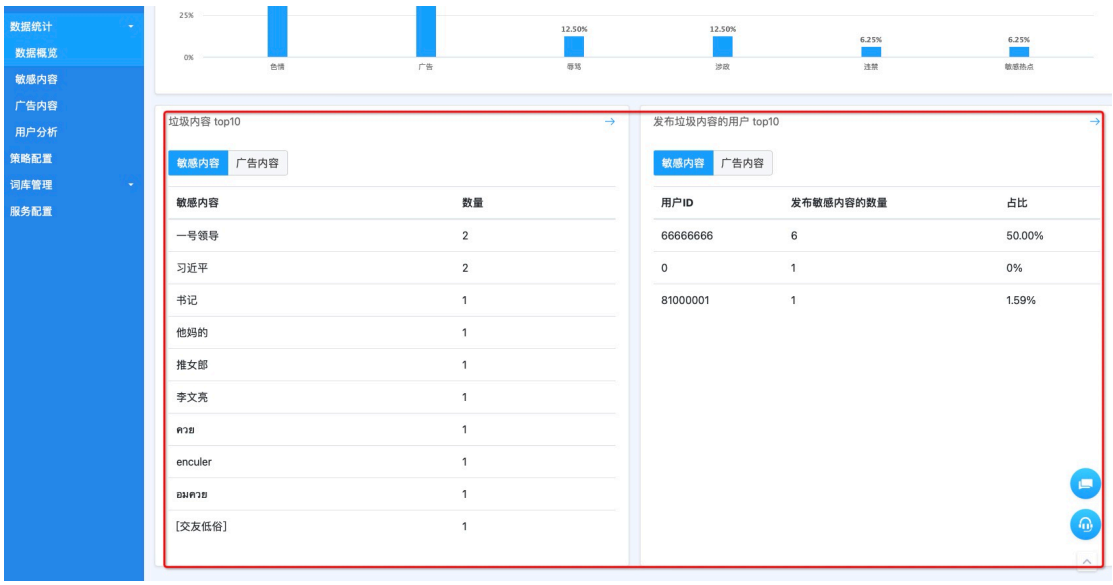
1.1.4 垃圾类型分布



此处以条形统计图的形式展示您筛选时间范围内，检出敏感内容的类型分布情况。

您可通过移动鼠标到某一矩形的方式，查看对应类型的具体检出条数。

1.1.5 垃圾内容&发布垃圾内容用户 top10



（1）垃圾内容 top10

- 敏感内容：此处展示您筛选时间范围内被检出的敏感内容命中次数最多的前 10 个敏感词及对应数量。
-广告内容：此处展示您筛选时间范围内检出的广告内容中出现次数最多的前 10 个广告内容及对应数量。

（2）垃圾内容用户 top10

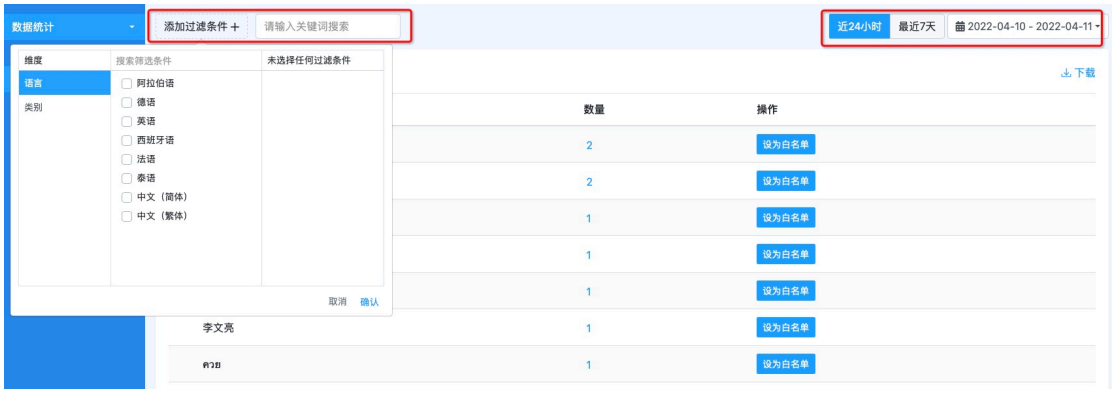
若该项目在调用文本审核接口时传入了用户 id，那么此处将分别展示您筛选时间范围内发送敏感内容、广告内容数量最多的前 10 个用户的 id、发送数量以及占其总共发送内容的比例。

（3）“—>”按钮
- 点击“垃圾内容 top10”右侧的“—>”，将进入「敏感内容」页面；
- 点击“发布垃圾内容用户 top10”右侧的“—>”，将进入「用户分析」页面。

1.2 敏感内容

该页面按照检出敏感内容命中的敏感词分类统计数量。您可以在此快速浏览筛选时间段内命中不同敏感词的数据。

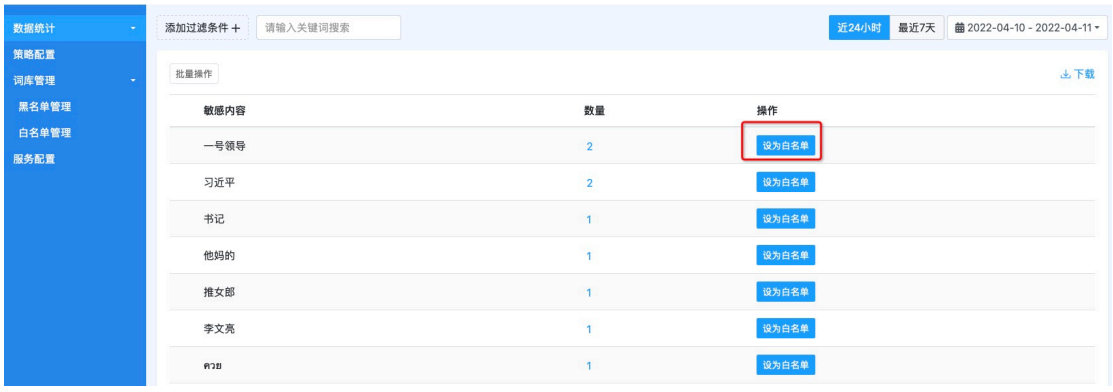
1.2.1 过滤条件



您可通过勾选语言、命中敏感类别，输入关键词，选择时间四个维度共同筛选要查看的敏感内容。

在此处修改所选时间段，下方将展示您所选时间段对应的敏感数据。

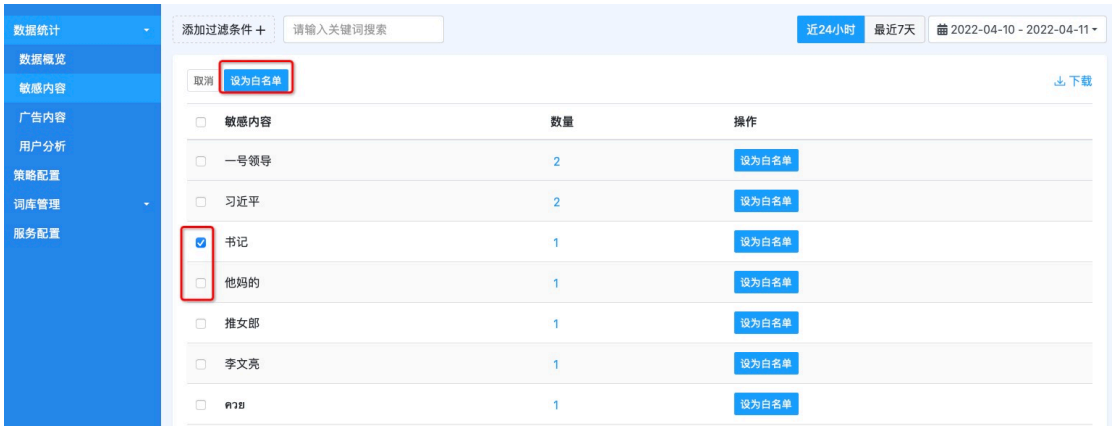
1.2.2 设为白名单



您可通过此功能将敏感词设置为“白名单”词汇。

点击每个词汇后的“设为白名单”按钮，系统将提示是否确认执行操作。在您确认后，对应敏感词将进入白名单，您可在「白名单管理」页面查看。同时，系统对仅命中该敏感词的文本将不再检出。

1.2.3 批量操作

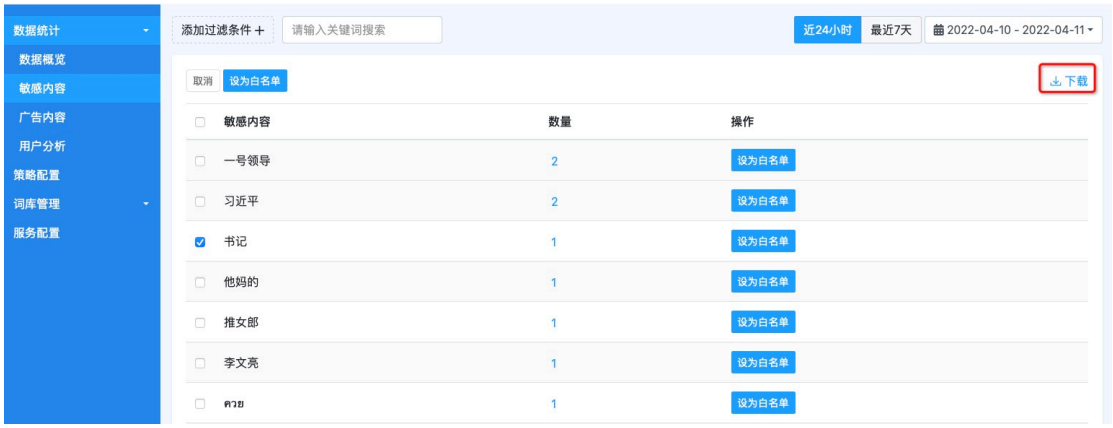


此功能可将敏感词批量设置为“白名单”词汇。

点击“批量操作”按钮，页面将进入如上图所示状态，您可勾选多个想要设置为白名单的词汇。勾选完毕，点击“设为白名单”按钮并确认操作后的处理逻辑与 1.2.3 一致，同时该页面自动退出上图所示状态。

在上图所示状态下，点击“取消”按钮，将直接退出该状态。

1.2.4 下载

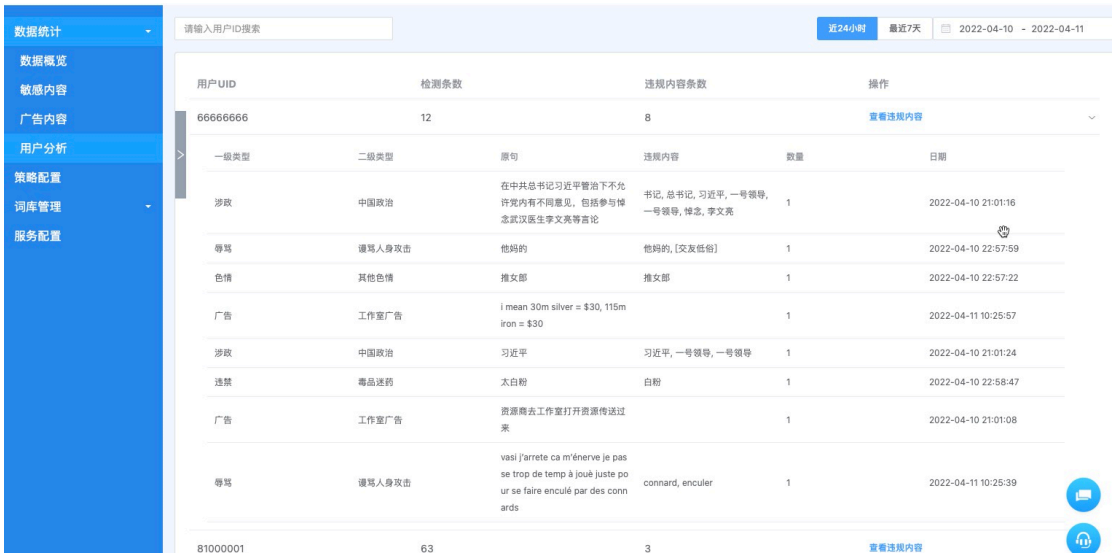


此功能可根据您的筛选结果将对应具体文本内容以.csv 格式的文件下载至本地。

1.3 广告内容

该页面展示筛选时间段内检出的广告内容数据。与「敏感内容」页面基本相同，支持通过勾选语言、命中敏感类别，选择时间三个维度共同筛选要查看的广告内容。

1.4 用户分析



默认情况下，该页面以用户 id 为主维度展示对应用户在筛选时间范围内检测内容数量、违规内容数量以及具体文本内容。

1.4.1 筛选条件

您可从输入用户 uid 和选择时间两个维度筛选要展示的内容。

1.4.2 展示字段选择

您可通过上图所示的“展示字段配置器”选择要展示的字段。当选择展示“违规内容占比”时，系统将在“违规内容条数”列后增加展示“违规内容占比”列；当选择展示“违规类型”时，系统将根据“用户 uid+类型”对筛选数据结果分行展示。

2、策略配置

用于针对同一个项目中的不同场景，配置审核松紧程度不同的策略。

2.1 默认策略

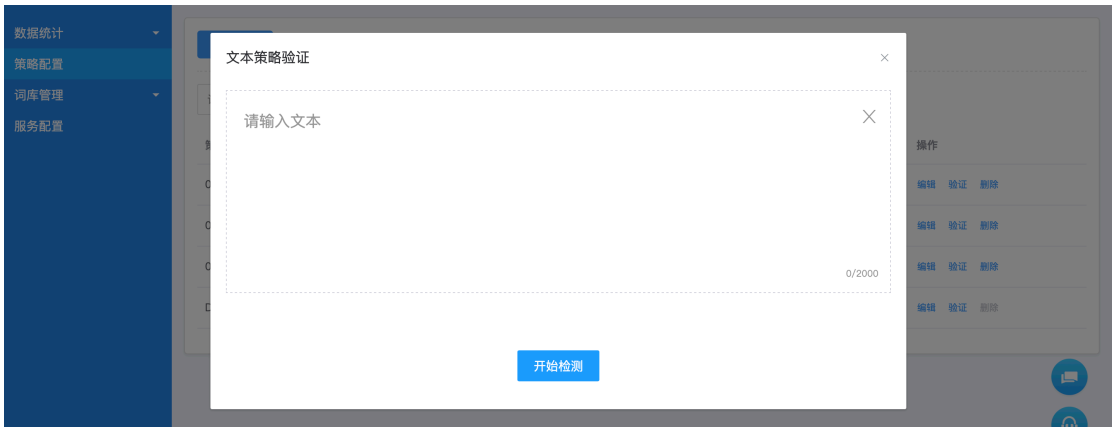
项目创建时，系统自动创建的策略。当调用文本审核接口的入参 strategyId 传入为空值时，将默认执行该策略。

注：默认策略不支持删除。

2.1.1 策略编辑

点击“编辑”，进入「文本策略编辑」页面，可调整对应策略开启检测的类别。

2.1.2 策略验证



点击“验证”，可输入文本测试您配置策略的检测效果。

2.2 创建策略

点击“创建策略”，将进入「文本策略创建」页面。

2.2.1 策略编号

您需要为此策略定义一个唯一编号，该编号将作为区分不同策略的标识，因此其与已有策略的编号不可重复。在调用文本审核接口时，可将编号作为入参 StrategyId 的值传入，系统将调用对应策略检测文本内容。

2.2.2 策略名称

您需要在此输入待创建策略的名称，以便于创建后快速查询、区分不同策略。

2.2.3 应用场景

您可直接在系统定义的场景中选择，或者自定义一个场景名称。

以上三项均填写完毕后，可继续根据需要配置该策略要检测的类别；点击“保存”后，策略即创建完成。后续在调用文本审核接口时可通过策略编号使用对应策略。

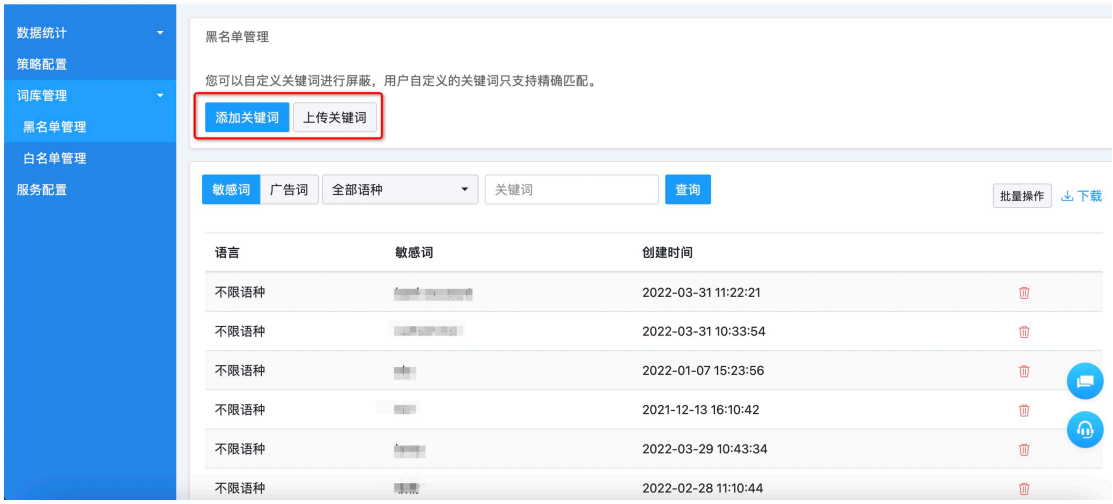
2.3 策略搜索

当列表中存在很多策略时，您可通过在此输入策略名称快速找到需要查询的策略。

3、词库管理

黑名单用于自定义需要屏蔽的特殊敏感词；
白名单用于自定义系统认为是敏感但业务上无需屏蔽的非敏感词以及自定义发送内容无需过文本审核的用户的 id。

3.1 黑名单管理



您可通过直接添加和批量上传两种方式向黑名单中添加敏感词及广告词，添加后的敏感词及广告词，将分类在下方列表中展示。

3.1.1 添加关键词

点击“添加关键词”，进入词汇添加页面。

（1）选择垃圾类型

分为普通敏感词和广告词两种类型（注：黑名单中自定义的广告词需在策略配置中开启“用户自定义广告”类别后，方可正常检测。）

（2）国家

您可在此选择待添加敏感词适用的国家。添加成功后，对应敏感词将只在接口传入的 country 字段值与配置国家 code 相同时生效。

（3）语言

表示敏感词（或广告词）适用的语种；即只有当检测文本中出现该敏感词且其语种识别结果在该敏感词（或广告词）选中的适用语种中时，该敏感词（或广告词）才会被检出。

（4）策略编号

表示敏感词（或广告词）适用的策略；即该敏感词只有在系统使用文本调用配置的对应策略且文本中包含该敏感词时候，才会被检出。

（5）结果

表示当检测文本命中该敏感词，系统返回的结果，包括：不通过和疑似两种情况。

（6）类型

表示当检测文本命中该敏感词时，系统返回的检出类型，一级类型包括：用户自定义、涉政、暴恐、违禁、色情、辱骂、仇恨言论、未成年保护、敏感热点、个人信息保护、私人交易、违规表情；二级类型即为这些一级类下的子类。默认情况下，一级类型系统自动选中“用户自定义”类，此时二级类型不支持选择。

（7）关键词

要自定义的黑名单词汇。

（8）是否报警

如开启，则表示当检测文本中包含该自定义词汇时，文本审核接口将返回 warning 字段，且值为 true。（注：仅垃圾类型选择“广告词”时，可配置该功能）

点击“保存”，系统将自动保存添加词汇，并返回黑名单列表页。

3.1.2 上传关键词

点击“上传关键词”，进入词汇上传页面。

（1）选择垃圾类型

同上。

（2）覆盖已经存在的关键词

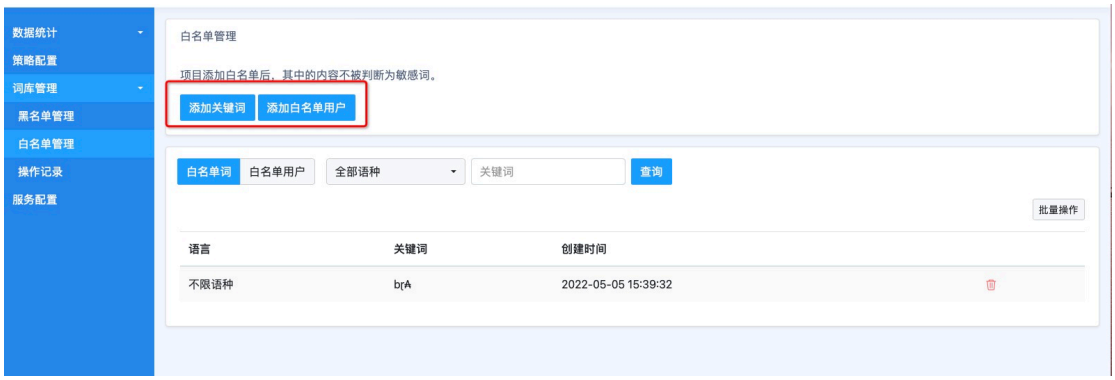
如开启，则在将上传敏感词保存后，系统将用本次上传的敏感词将历史所有添加过的敏感词覆盖，仅保留本次上传结果。

（3）上传关键词

点击“浏览”按钮，选择要上传的文件（文件格式需与“下载示例”中的文件格式一致）

点击“保存”，系统将自动保存上传词汇，并返回黑名单列表页。

3.2 白名单管理



您可通过直接添加关键词的方式向白名单中添加要忽略的词汇或用户 id，添加后的词汇活用户 id 将在下方的列表中展示。

3.2.1 添加关键词

点击“添加关键词”，进入词汇添加页面。

（1）语言

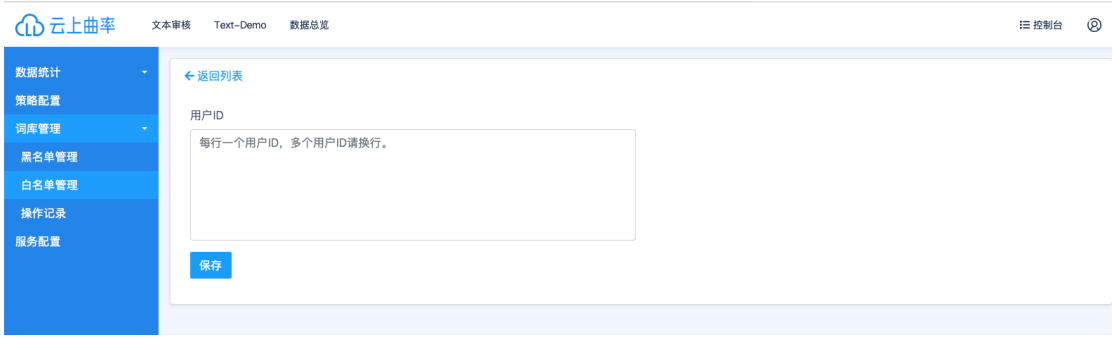
表示该白名单词汇适用的语种；即只有当检测文本中出现该词被系统判定为“敏感”，且其语种识别结果在该词选中的适用语种中时，该词才会再次被修正为“正常”结果。

（2）关键词

要自定义的白名单词汇。

点击“保存”，系统将自动保存添加词汇，并返回白名单列表页。

3.2.2 添加白名单用户

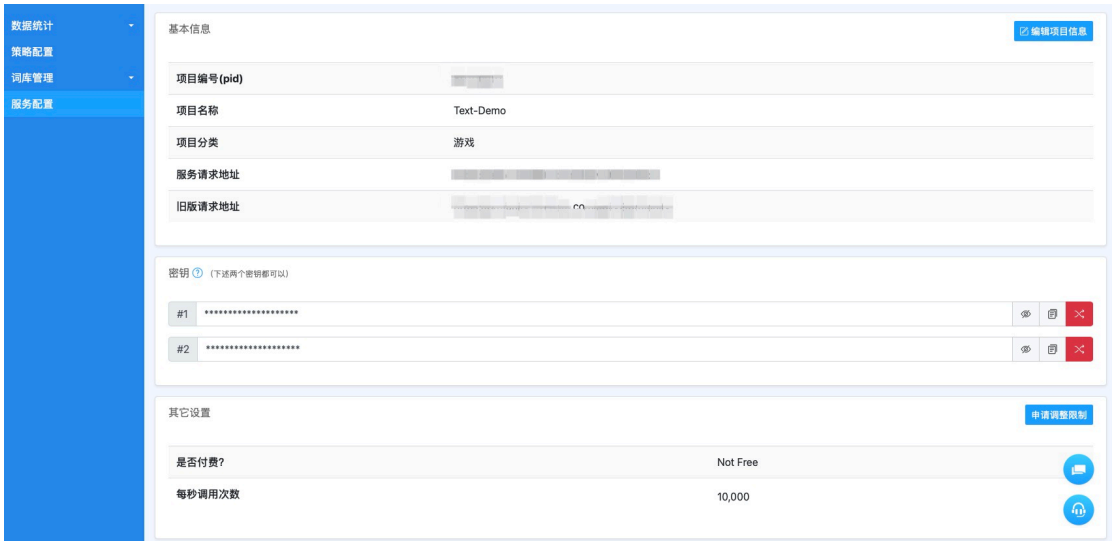


点击“添加白名单用户”按钮，进入用户 id 添加页面。您可按照每行一个用户 id 的形式，同时添加多个白名单用户。保存后，系统对这些用户发送的内容，将均返回通过结果。

3.3 操作记录

操作记录中记录了用户对黑、白名单的操作数据。您可通过操作时间、操作动作、操作人、词汇分类、操作词汇等筛选条件进行记录检索。

4、服务配置



此页面展示项目编号、名称、分类、请求地址、密钥等信息。同时，您可通过“编辑项目信息”按钮，修改项目名称、分类以及描述信息。