Microsoft

Clasificación y moderación de texto con Azure Content Moderator

BIENVENIDA

En este módulo, se presentará Azure Content Moderator y se mostrará cómo usarlo para la moderación de texto.

En este módulo, aprenderá a:

- Conocer en qué consiste la moderación de contenido de texto.
- Obtener información sobre las características clave de Azure Content Moderator para la moderación de texto.
- Probar la moderación de texto con la consola de prueba de API basada en web.

Requisitos previos

Ninguno

Introducción 1 min

Introducción a la moderación de texto 4 min

Creación de un recurso de Content Moderator y suscripción a este 7 min

Ejercicio: Moderación de texto de prueba mediante la consola de prueba de API 10 min

Comprobación de conocimientos 7 min

Resumen 2 min

INTRODUCCIÓN

La moderación de contenido para aspectos problemáticos puede llevar mucho tiempo. Microsoft Azure Content Moderator proporciona moderación asistida automáticamente de imágenes, texto y vídeo. Este módulo abarca los conceptos clave relacionados con el uso de Content Moderator para llevar a cabo la moderación de texto.

Nota

Este módulo requiere una suscripción de Azure. Los servicios que cree y use son gratuitos, pero necesitará una suscripción activa o una versión de prueba para completar los ejercicios. Si no tiene una suscripción a Azure, cree una **cuenta gratuita** antes de empezar.

Objetivos de aprendizaje

En este módulo aprenderá a hacer lo siguiente:

- Conocer en qué consiste la moderación de contenido de texto.
- Obtener información sobre las características clave de Azure Content Moderator para la moderación de texto.
- Probar la moderación de texto con la consola de prueba de API.

INTRODUCCIÓN A LA MODERACIÓN DE TEXTO

Al usar la moderación de contenido asistida automáticamente, el contenido se bloquea, aprueba o revisa en función de las directivas y los umbrales. Esta asistencia automática se usa para aumentar la moderación humana de los entornos donde socios, empleados y consumidores generan contenido de texto. Entre estos entornos se incluyen los siguientes:

- Salas de chat
- Paneles de discusión
- Bots de chat
- Catálogos de comercio electrónico
- Documentos

La respuesta de Text Moderation API incluye la información siguiente:

- Una lista de palabras potencialmente no deseadas que se encuentran en el texto.
- El tipo de palabras potencialmente no deseadas.
- La posible información de identificación personal que se ha encontrado.

Palabras soeces

Al pasar texto a la API, se identifican los posibles términos soeces en el texto y se devuelven en una respuesta JSON. El elemento soez se devuelve como un Term en la respuesta JSON, junto con un valor de

índice en el que se muestra la ubicación del término en el texto proporcionado.

También se pueden usar listas de términos personalizadas con esta API. En ese caso, si se identifica un término soez en el texto, también se devuelve un valor ListId para identificar la palabra personalizada específica que se ha encontrado. Aquí se muestra un ejemplo de una respuesta JSON:

JSON

```
"Terms": [
{
    "Index": 118,
    "OriginalIndex": 118,
    "ListId": 0,
    "Term": "crap"
}
```


Clasificación

Esta característica de la API puede colocar texto en categorías específicas según las especificaciones siguientes:

- **Categoría 1**: posible presencia de lenguaje que se puede considerar sexualmente explícito o para adultos en ciertas situaciones.

- **Categoría 2:** posible presencia de lenguaje que se puede considerar sexualmente insinuante o para adultos en ciertas situaciones.
- **Categoría 3:** posible presencia de lenguaje que se puede considerar ofensivo en ciertas situaciones.

Cuando se devuelve la respuesta JSON, proporciona un valor booleano para una revisión recomendada del texto. Si es true, debe revisar el contenido de forma manual para determinar posibles problemas.

Cada categoría también se devuelve con una puntuación entre 0 y 1 para indicar la categoría predicha para el texto evaluado. Cuanto mayor sea la puntuación, más probable es que se aplique la categoría en cuestión. Aquí se muestra una respuesta JSON de ejemplo:

JSON

```
"Classification": {
   "ReviewRecommended": true,
   "Category1": {
      "Score": 1.5113095059859916E-06
      },
   "Category2": {
      "Score": 0.12747249007225037
      },
   "Category3": {
```

```
        "Score": 0.98799997568130493
    }
}
```

Información de identificación personal

La información de identificación personal es de vital importancia en muchas aplicaciones. Esta característica de la API puede ayudarle a detectar si algún valor del texto se considera información de identificación personal antes de publicarlo. Entre los aspectos clave que se detectan se incluyen los siguientes:

- Direcciones de correo electrónico
- Direcciones de correo postal de EE. UU.
- Direcciones IP
- Números de teléfono de EE. UU.
- Números de teléfono de Reino Unido
- Números del seguro social

Si se encuentran valores que podrían ser información de identificación personal, la respuesta JSON incluirá la información pertinente sobre el texto y la ubicación de índice dentro del texto. Aquí se muestra un ejemplo de una respuesta JSON:

JSON

```
"PII": {
    "Email": [{
        "Detected": "abcdef@abcd.com",
```

```
    "SubType": "Regular",
    "Text": "abcdef@abcd.com",
    "Index": 32
    }],
"IPA": [{
    "SubType": "IPV4",
    "Text": "255.255.255.255",
    "Index": 72
    }],
"Phone": [{
    "CountryCode": "US",
    "Text": "5557789887",
    "Index": 56
    }, {
    "CountryCode": "UK",
    "Text": "+44 123 456 7890",
    "Index": 208
    }],
```

```
"Address": [{
    "Text": "1 Microsoft Way, Redmond, WA 98052",
    "Index": 89
    }],
"SSN": [{
    "Text": "999-99-9999",
    "Index": 267
    }]
}
```

CREACIÓN DE UN RECURSO DE CONTENT MODERATOR Y SUSCRIPCIÓN A ESTE

Antes de empezar a probar la moderación de contenido o integrarla en aplicaciones personalizadas, tendrá que crear un recurso de Content Moderator, suscribirse a este y obtener la clave de suscripción para tener acceso al servicio.

En este ejercicio, creará un recurso de Content Moderator en Azure Portal.

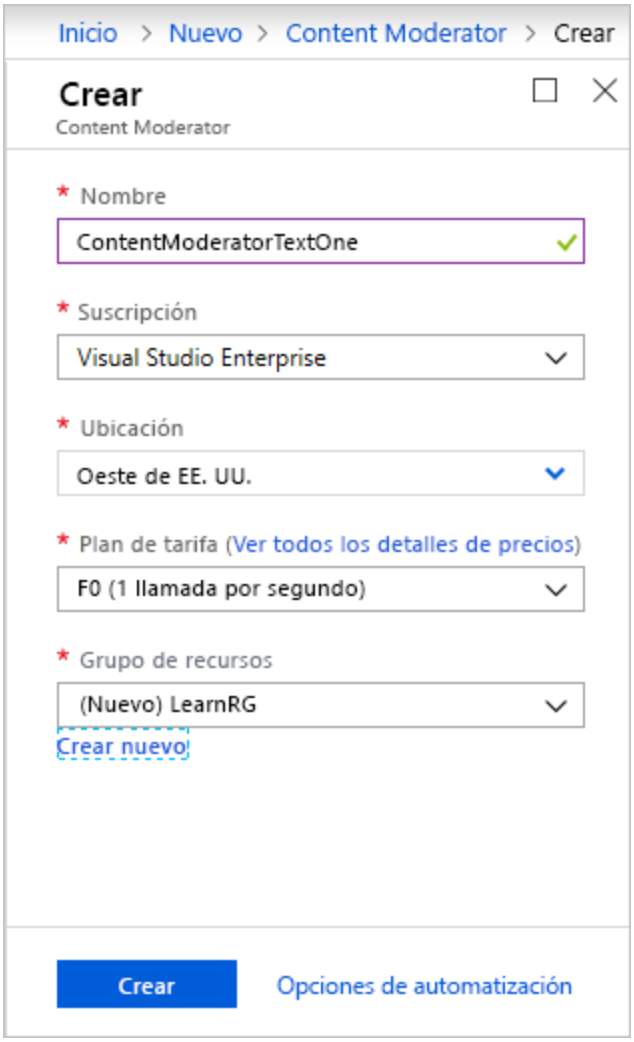
Creación de un recurso de Content Moderator y suscripción a este

1. Inicie sesión en Azure Portal .
2. En el panel izquierdo, seleccione **Crear un recurso**.
3. En el cuadro de búsqueda, escriba **Content Moderator** y presione ENTRAR.
4. Seleccione **Content Moderator** en los resultados de la búsqueda.
5. Haga clic en **Crear**.
6. Escriba un nombre único para el recurso, elija una suscripción y seleccione una ubicación cercana a usted.

7. Seleccione el plan de tarifa para este recurso y, después, seleccione **S0**.

8. Cree un grupo de recursos denominado **LearnRG**.

9. Haga clic en **Crear**.



El recurso tardará unos minutos en implementarse. Una vez que lo haya hecho, vaya al nuevo recurso.

Copia de la clave de suscripción

Para tener acceso al recurso de Content Moderator, necesitará una clave de suscripción:

1. En el panel de la izquierda, en **ADMINISTRACIÓN DE RECURSOS**, seleccione **Claves**.

2. Copie uno de los valores de las claves de suscripción para usarlo posteriormente.

EJERCICIO: MODERACIÓN DE TEXTO DE PRUEBA MEDIANTE LA CONSOLA DE PRUEBA DE API

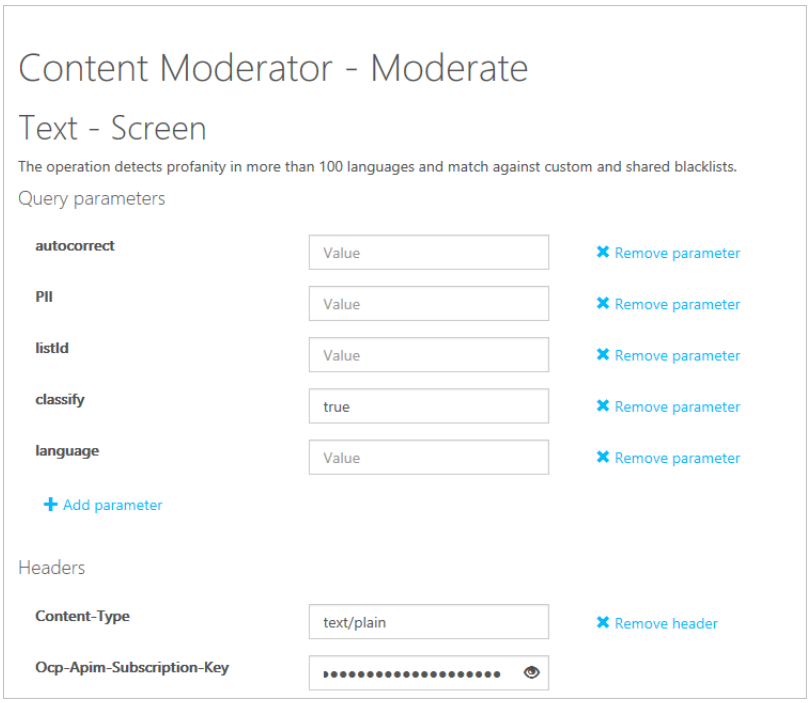
Ahora que tiene un recurso disponible en Azure para la moderación de contenido y que dispone de una clave de suscripción para ese recurso, ejecutaremos algunas pruebas mediante la consola de prueba de API basada en web.

1. Vaya a la página de Referencia de Content Moderator API. Esta página está disponible en diferentes regiones para realizar pruebas en la consola de API.

2. Seleccione el botón con la ubicación adecuada para la región geográfica más cercana para abrir la consola.

3. Observe los parámetros de consulta que puede seleccionar para la prueba. Mantenga las opciones predeterminadas para la primera serie de pruebas.

4. Pegue la clave de suscripción en el cuadro **Ocp-Apim-Subscription-Key**.



5. Mantenga el texto de ejemplo y haga clic en **Enviar**.

Evaluación de la respuesta

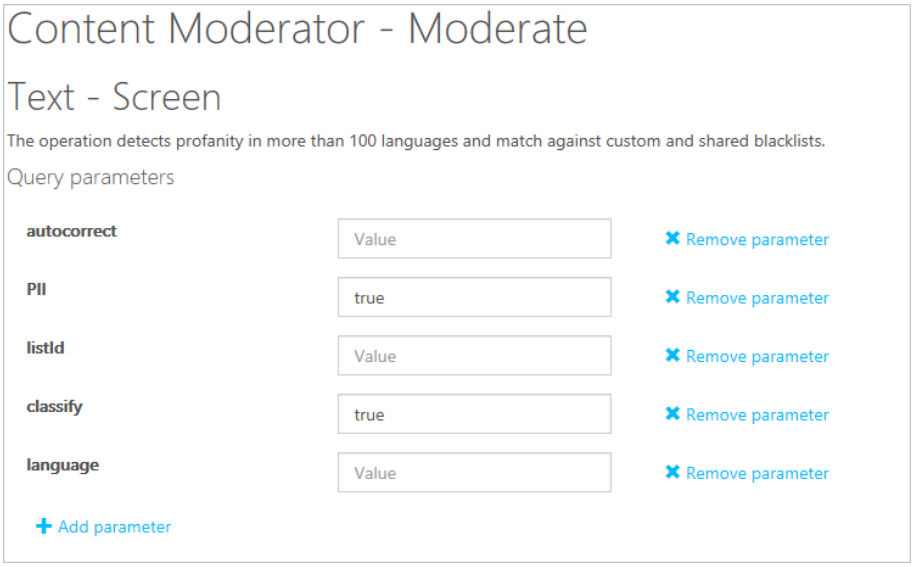
- Desplácese hacia abajo en la página y evalúe la respuesta de la consola de pruebas.

En esta primera prueba se usó *classification* porque el parámetro classify estaba establecido en true en la parte superior de la página. La respuesta contiene la información siguiente:

- Se recomienda una revisión.
- El texto se ha clasificado como Categoría 3 (posible presencia de lenguaje que se podría considerar ofensivo en ciertas situaciones).
- El término que podría resultar ofensivo es "crap" (porquería).

Ejecución de pruebas adicionales

1. Para ejecutar la segunda prueba, desplácese hasta la parte superior de la página y establezca el parámetro PII en true.



2. Haga clic en **Enviar**.

Observe que se muestra el contenido de información de identificación personal. Ahora se considera que la dirección de correo electrónico, la dirección IP, el número de teléfono, la dirección de correo postal y el número del seguro social podrían ser información de identificación personal.

3. Si quiere ejecutar otras pruebas, escriba algunos valores de texto propios de un documento existente y vuelva a ejecutar las pruebas para ver los resultados devueltos.

4. Evalúe la respuesta JSON y la sintaxis de la dirección URL de la solicitud para analizar la manera en que las aplicaciones personalizadas pueden llamar a esta API.

Sugerencia

Para probar esta API mediante una aplicación de C#, vea **Guía de inicio rápido: Análisis de contenido de textos para detectar material inapropiado en C#**.