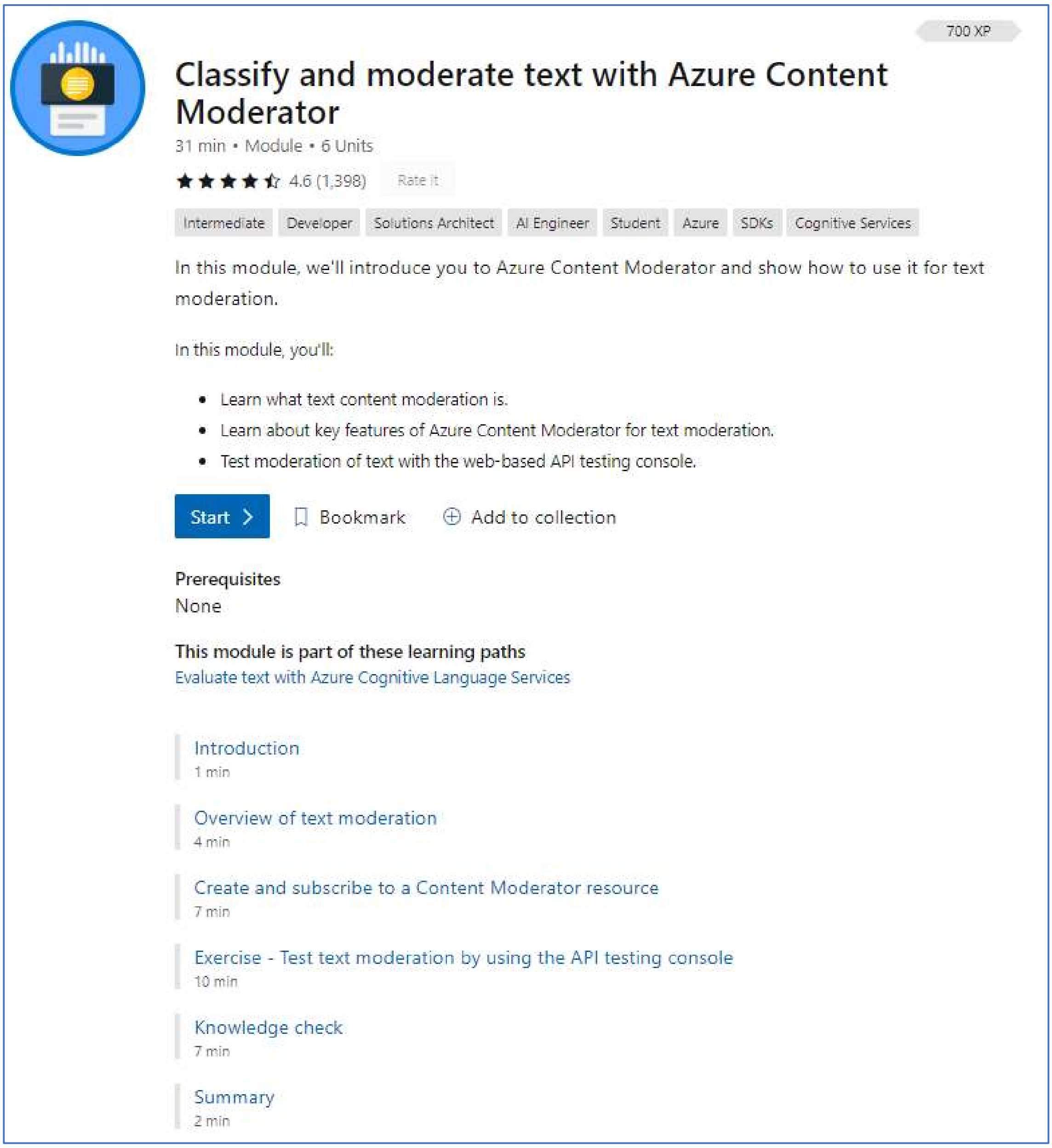


Classify and moderate text with Azure Content Moderator

31 min • Module • 6 Units

★★★★⯪ 4.6 (1,398)

In this module, we'll introduce you to Azure Content Moderator and show how to use it for text moderation.

In this module, you'll:

- Learn what text content moderation is.
- Learn about key features of Azure Content Moderator for text moderation.
- Test moderation of text with the web-based API testing console.

Prerequisites
None

This module is part of these learning paths
Evaluate text with Azure Cognitive Language Services

https://docs.microsoft.com/en-us/learn/modules/classify-and-moderate-text-with-azure-content-moderator/

Summary

Introduction

Moderating content for problematic aspects can be a time-consuming task. Microsoft Azure Content Moderator provides machine-assisted content moderation for images, text, and video. This module covers the key concepts involved in using Content Moderator to do text moderation.

Note

This module requires an Azure subscription. The services you create and use are free, but you'll need an active subscription or trial to complete the exercises. If you don't have an Azure subscription, **create a free account** before you begin.

Learning objectives

In this module, you'll:

- Learn what text content moderation is.
- Learn about key features of Azure Content Moderator for text moderation.
- Test moderation of text with the API test console.

Overview of text moderation

When you're using machine-assisted content moderation, you either block, approve, or review the content based on your policies and thresholds. You can use machine assistance to augment human moderation of environments where partners, employees, and consumers generate text content. These places include:

- Chat rooms
- Discussion boards
- Chatbots
- E-commerce catalogs
- Documents

The response from the Text Moderation API includes the following information:

- A list of potentially unwanted words found in the text.
- What type of potentially unwanted words were found.
- Possible personally identifiable information (PII) found in the text.

Profanity

When you pass text to the API, any potentially profane terms in the text are identified and returned in a JSON response. The profane item is returned as a `Term` in the JSON response, along with an index value showing where the term is in the supplied text.

You can also use custom term lists with this API. In that case, if a profane term is identified in the text, a `ListId` is also returned to identify the specific custom word that was identified. A sample JSON response is shown here:

JSON

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

Classification

This feature of the API can place text into specific categories based on the following specifications:

- **Category 1:** Potential presence of language that might be considered sexually explicit or adult in certain situations.
- **Category 2:** Potential presence of language that might be considered sexually suggestive or mature in certain situations.
- **Category 3:** Potential presence of language that might be considered offensive in certain situations.

When the JSON response is returned, it provides a Boolean value for a recommended review of the text. If `true`, you should review the content manually to determine the potential for any issues.

Each category is also returned with a score between 0 and 1 to indicate the predicted category for the evaluated text. The higher the score, the more likely it is that the category might apply. Here's a sample JSON response:

JSON

```json
"Classification": {
    "ReviewRecommended": true,
    "Category1": {
        "Score": 1.5113095059859916E-06
        },
    "Category2": {
        "Score": 0.12747249007225037
        },
    "Category3": {
        "Score": 0.98799997568130493
    }
}
```

Personally identifiable information

Personally identifiable information (PII) is of critical importance in many applications. This feature of the API can help you detect if any values in the text might be considered PII before you release it publicly. Key aspects that are detected include:

- Email addresses
- US mailing addresses
- IP addresses
- US phone numbers
- UK phone numbers

- Social Security numbers

If possible PII values are found, the JSON response includes relevant information about the text and the index location within the text. A sample JSON response is shown here:

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
        }]
    }
```

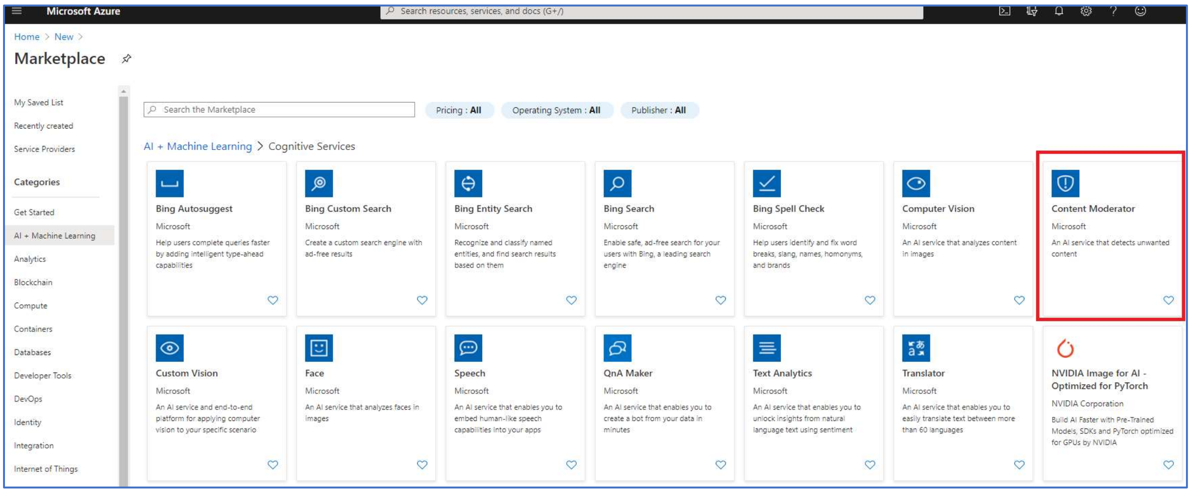
Create and subscribe to a Content Moderator resource

Before you can begin to test content moderation or integrate it into your custom applications, you need to create and subscribe to a Content Moderator resource and get the subscription key for accessing the resource.

In this exercise, you'll create a new Content Moderator resource in the Azure portal.

Create and subscribe to a Content Moderator resource

1. Sign in to the [Azure portal](https://portal.azure.com).
2. In the left pane, select **Create a resource**.
3. In the search box, enter **Content Moderator**, and then press Enter.



Bing Autosuggest
Microsoft
Help users complete queries faster by adding intelligent type-ahead capabilities

Bing Custom Search
Microsoft
Create a custom search engine with ad-free results

Bing Entity Search
Microsoft
Recognize and classify named entities, and find search results based on them

Bing Search
Microsoft
Enable safe, ad-free search for your users with Bing, a leading search engine

Bing Spell Check
Microsoft
Help users identify and fix word breaks, slang, names, homonyms, and brands

Computer Vision
Microsoft
An AI service that analyzes content in images

Content Moderator
Microsoft
An AI service that detects unwanted content

Custom Vision
Microsoft
An AI service and end-to-end platform for applying computer vision to your specific scenario

Face
Microsoft
An AI service that analyzes faces in images

Speech
Microsoft
An AI service that enables you to embed human-like speech capabilities into your apps

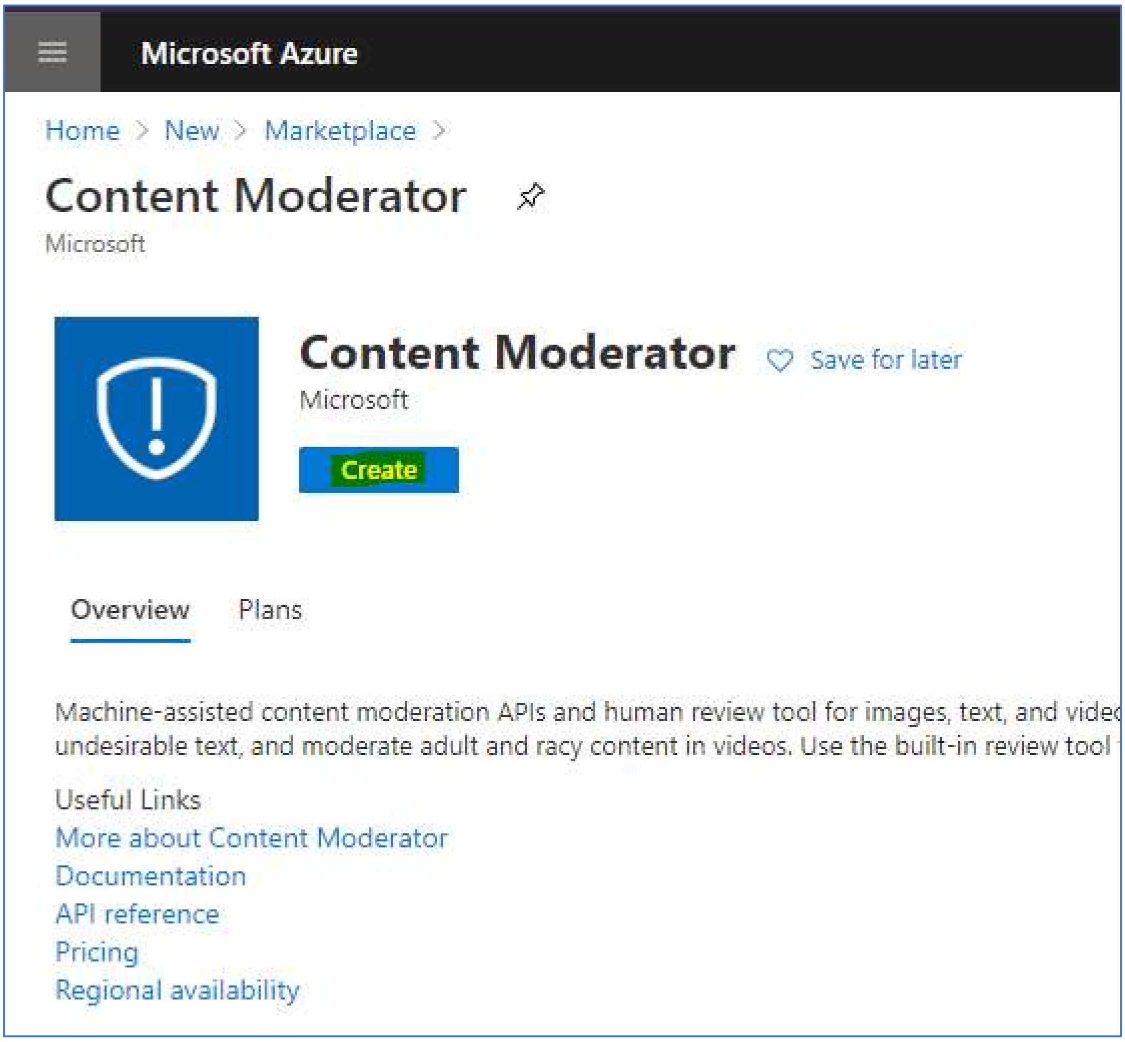
QnA Maker
Microsoft
An AI service that enables you to create a bot from your data in minutes

Text Analytics
Microsoft
An AI service that enables you to unlock insights from natural language text using sentiment

Translator
Microsoft
An AI service that enables you to easily translate text between more than 60 languages

NVIDIA Image for AI - Optimized for PyTorch
NVIDIA Corporation
Build AI Faster with Pre-Trained Models, SDKs and PyTorch optimized for GPUs by NVIDIA

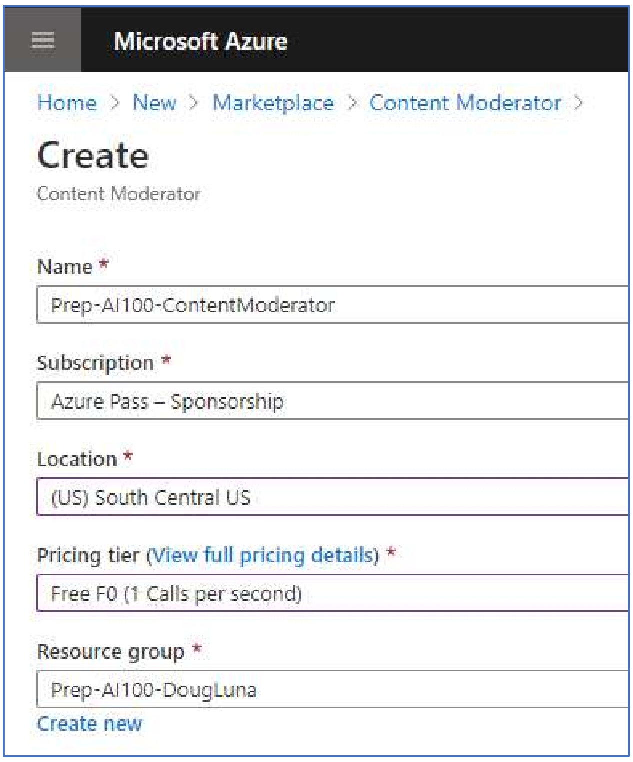
4. From the search results, select **Content Moderator**.
5. Select **Create**.



Machine-assisted content moderation APIs and human review tool for images, text, and vide
undesirable text, and moderate adult and racy content in videos. Use the built-in review tool

Useful Links
More about Content Moderator
Documentation
API reference
Pricing
Regional availability

6. Enter a unique name for your resource, choose a subscription, and select a location close to you.
7. Select the pricing tier for this resource, then choose **S0**.



Create

Content Moderator

Name *

Prep-AI100-ContentModerator

Subscription *

Azure Pass – Sponsorship

Location *

(US) South Central US

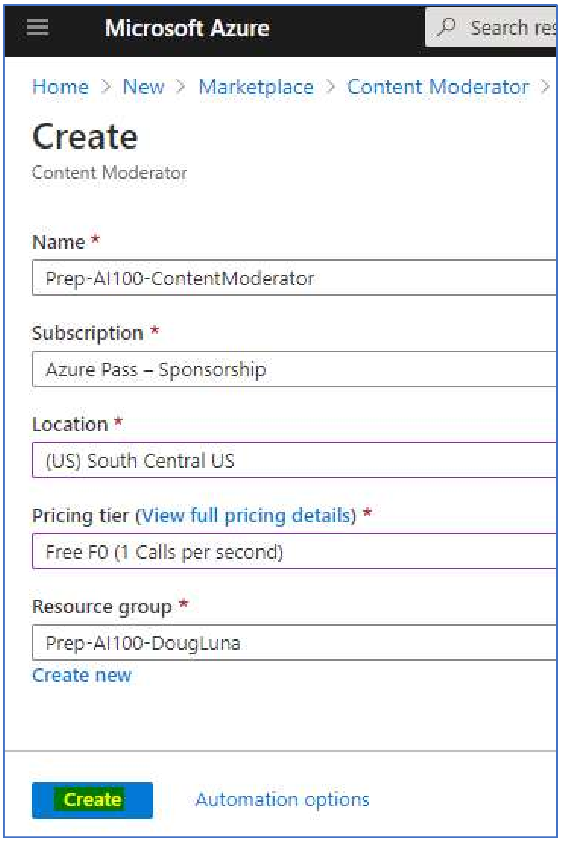
Pricing tier (View full pricing details) *

Free F0 (1 Calls per second)

Resource group *

Prep-AI100-DougLuna

8. Create a new resource group named **LearnRG**.
9. Select **Create**.



Create

Content Moderator

Name *

Prep-AI100-ContentModerator

Subscription *

Azure Pass – Sponsorship

Location *

(US) South Central US

Pricing tier (View full pricing details) *

Free F0 (1 Calls per second)

Resource group *

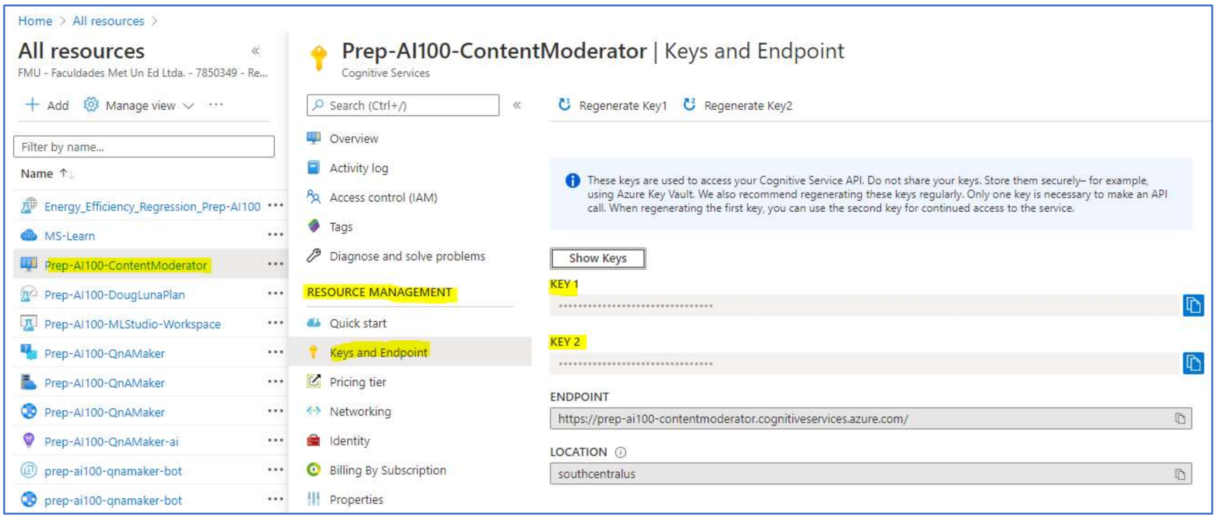
Prep-AI100-DougLuna

The resource will take a few minutes to deploy. After it does, go to the new resource.

Copy the subscription key

To access your Content Moderator resource, you'll need a subscription key:

1. In the left pane, under **RESOURCE MANAGEMENT**, select **Keys and Endpoints**.
2. Copy one of the subscription key values for later use.



These keys are used to access your Cognitive Service API. Do not share your keys. Store them securely– for example, using Azure Key Vault. We also recommend regenerating these keys regularly. Only one key is necessary to make an API call. When regenerating the first key, you can use the second key for continued access to the service.

KEY 1: ••••••••••••••••••••••••••••••••

KEY 2: ••••••••••••••••••••••••••••••••

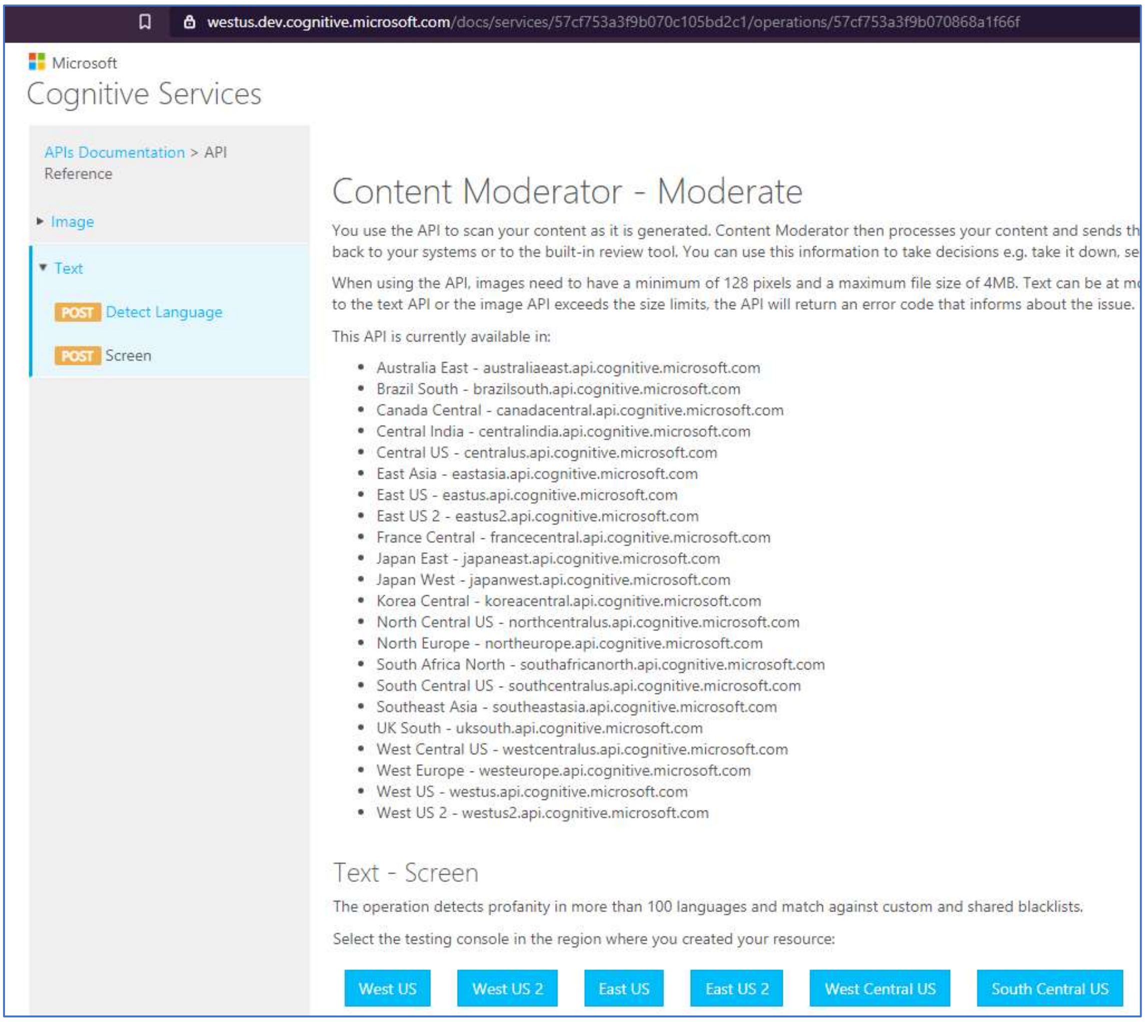
ENDPOINT: https://prep-ai100-contentmoderator.cognitiveservices.azure.com/

LOCATION: southcentralus

Exercise - Test text moderation by using the API testing console

Now that you have a resource available in Azure for content moderation, and you have a subscription key for that resource, let's run some tests by using the API web-based testing console.

1. Go to the Content Moderator API Reference page. This page is available in a number of regions for testing in the API console.



westus.dev.cognitive.microsoft.com/docs/services/57cf753a3f9b070c105bd2c1/operations/57cf753a3f9b070868a1f66f
Microsoft
Cognitive Services
APIs Documentation > API Reference
Image
Text
POST Detect Language
POST Screen

Content Moderator - Moderate

You use the API to scan your content as it is generated. Content Moderator then processes your content and sends th back to your systems or to the built-in review tool. You can use this information to take decisions e.g. take it down, se

When using the API, images need to have a minimum of 128 pixels and a maximum file size of 4MB. Text can be at m to the text API or the image API exceeds the size limits, the API will return an error code that informs about the issue.

This API is currently available in:

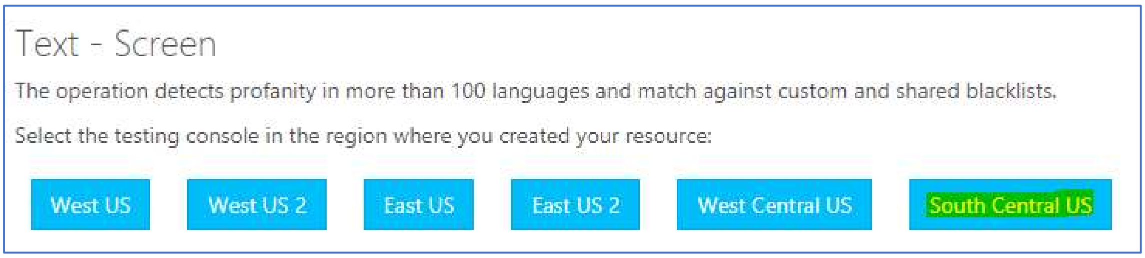
- Australia East - australiaeast.api.cognitive.microsoft.com
- Brazil South - brazilsouth.api.cognitive.microsoft.com
- Canada Central - canadacentral.api.cognitive.microsoft.com
- Central India - centralindia.api.cognitive.microsoft.com
- Central US - centralus.api.cognitive.microsoft.com
- East Asia - eastasia.api.cognitive.microsoft.com
- East US - eastus.api.cognitive.microsoft.com
- East US 2 - eastus2.api.cognitive.microsoft.com
- France Central - francecentral.api.cognitive.microsoft.com
- Japan East - japaneast.api.cognitive.microsoft.com
- Japan West - japanwest.api.cognitive.microsoft.com
- Korea Central - koreacentral.api.cognitive.microsoft.com
- North Central US - northcentralus.api.cognitive.microsoft.com
- North Europe - northeurope.api.cognitive.microsoft.com
- South Africa North - southafricanorth.api.cognitive.microsoft.com
- South Central US - southcentralus.api.cognitive.microsoft.com
- Southeast Asia - southeastasia.api.cognitive.microsoft.com
- UK South - uksouth.api.cognitive.microsoft.com
- West Central US - westcentralus.api.cognitive.microsoft.com
- West Europe - westeurope.api.cognitive.microsoft.com
- West US - westus.api.cognitive.microsoft.com
- West US 2 - westus2.api.cognitive.microsoft.com

Text - Screen

The operation detects profanity in more than 100 languages and match against custom and shared blacklists.

Select the testing console in the region where you created your resource:

2. For the geographic region closest to you, select the appropriate location button to open the console.

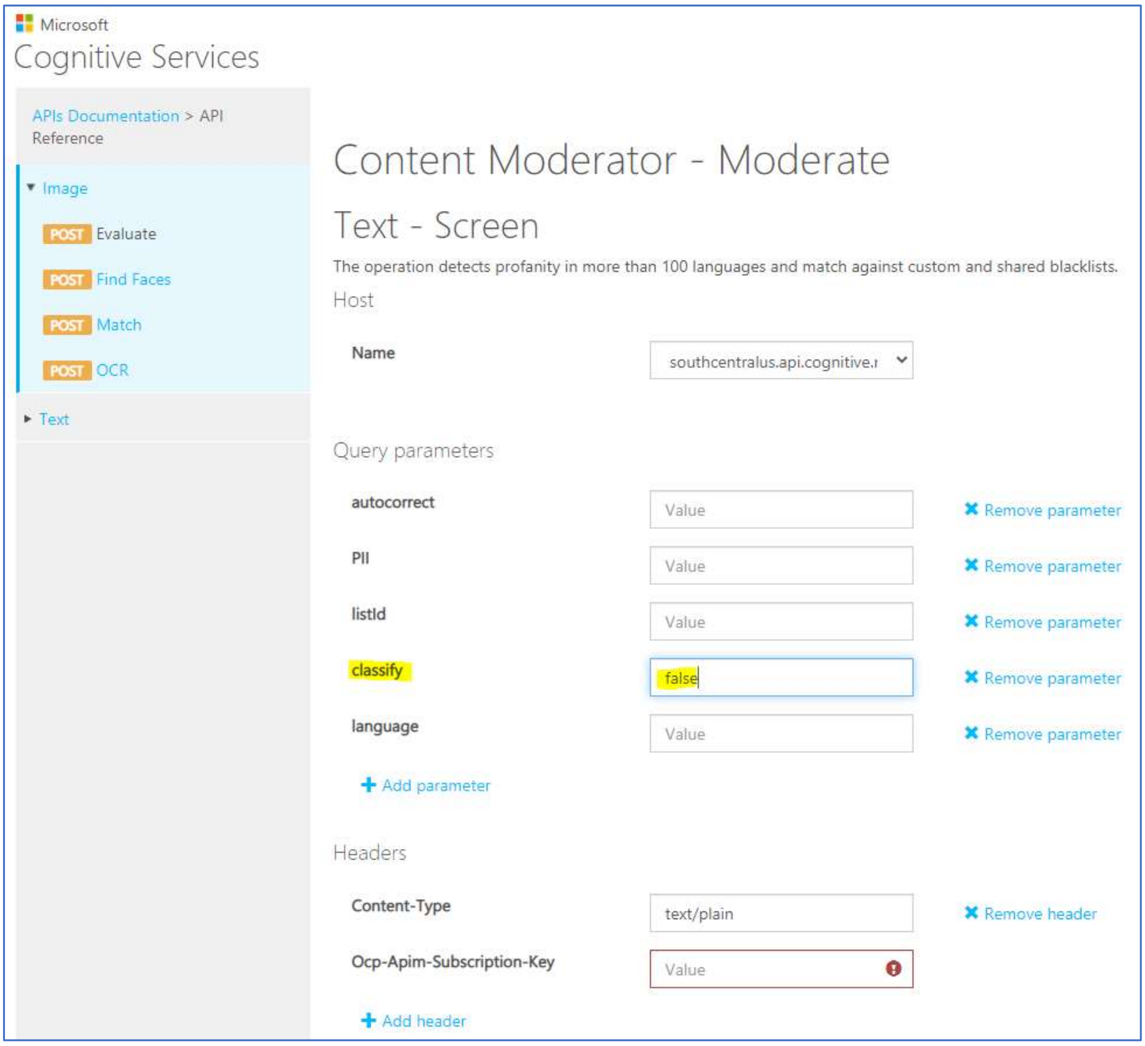


Text - Screen

The operation detects profanity in more than 100 languages and match against custom and shared blacklists.

Select the testing console in the region where you created your resource:

3. Note the query parameters that you can select for your test. For the first test run, ensure that the classify option is set to false. Leave the remaining values at their default.



Content Moderator - Moderate

Text - Screen

The operation detects profanity in more than 100 languages and match against custom and shared blacklists.

Host

Name	southcentralus.api.cognitive.r

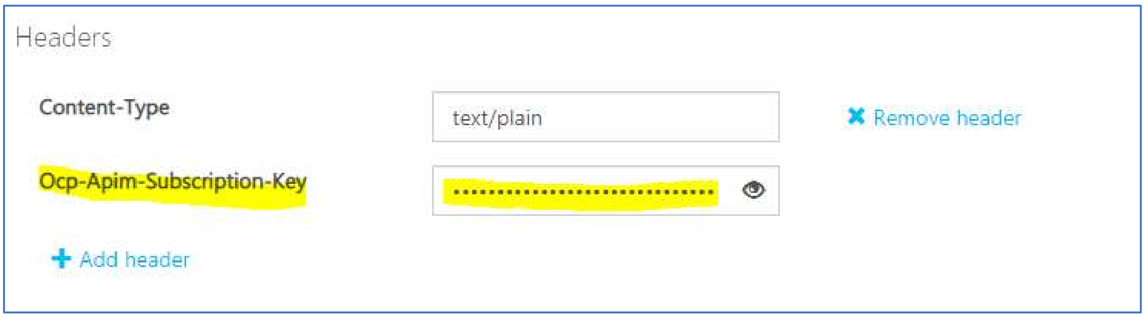
Query parameters

autocorrect	Value
PII	Value
listId	Value
classify	false
language	Value

Headers

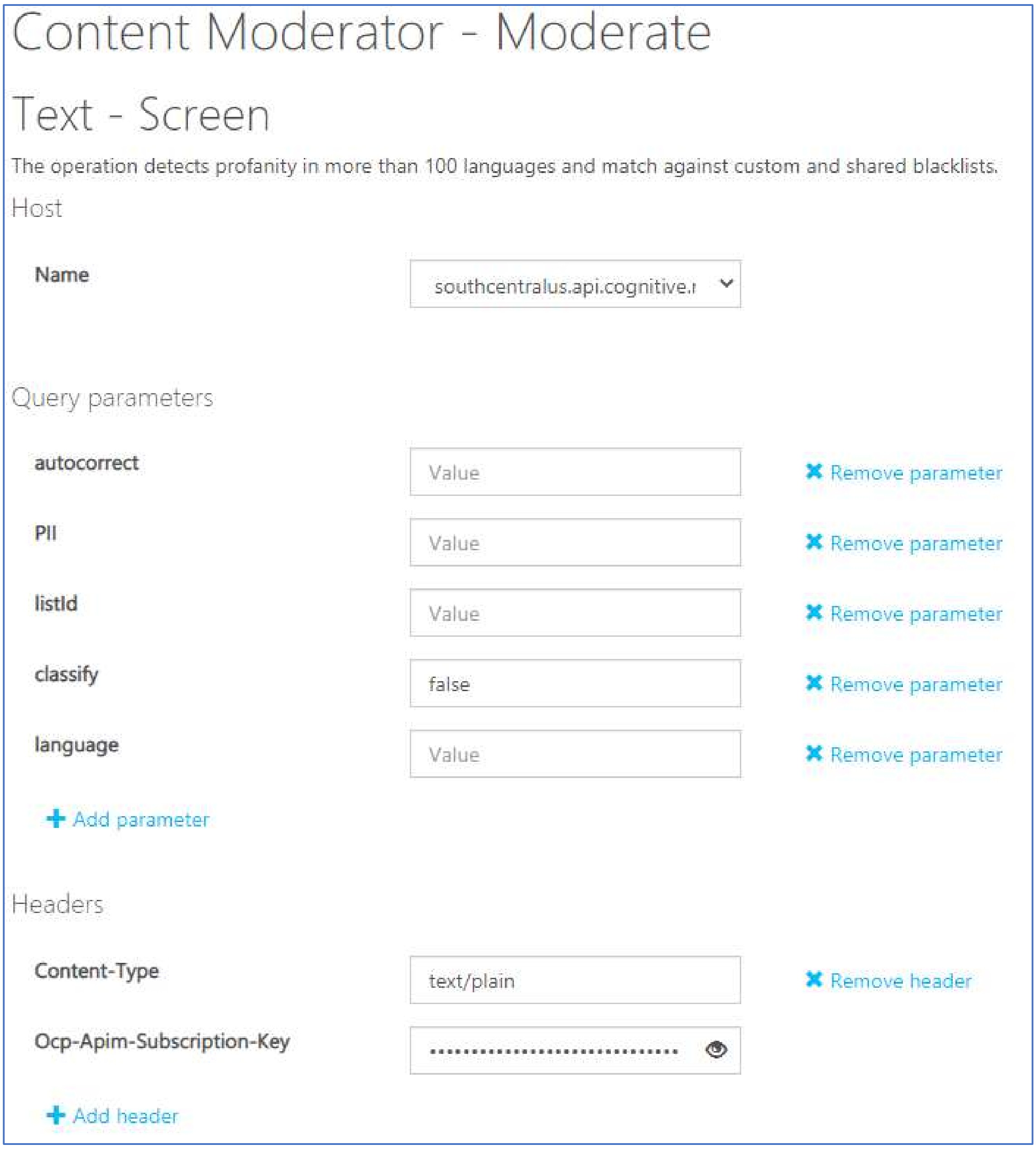
Content-Type	text/plain
Ocp-Apim-Subscription-Key	Value

4. Paste your subscription key into the **Ocp-Apim-Subscription-Key** box.



Headers

Content-Type	text/plain
Ocp-Apim-Subscription-Key	••••••••••••••••••••••••••••••••



Content Moderator - Moderate

Text - Screen

The operation detects profanity in more than 100 languages and match against custom and shared blacklists.

Host

Name	southcentralus.api.cognitive.r

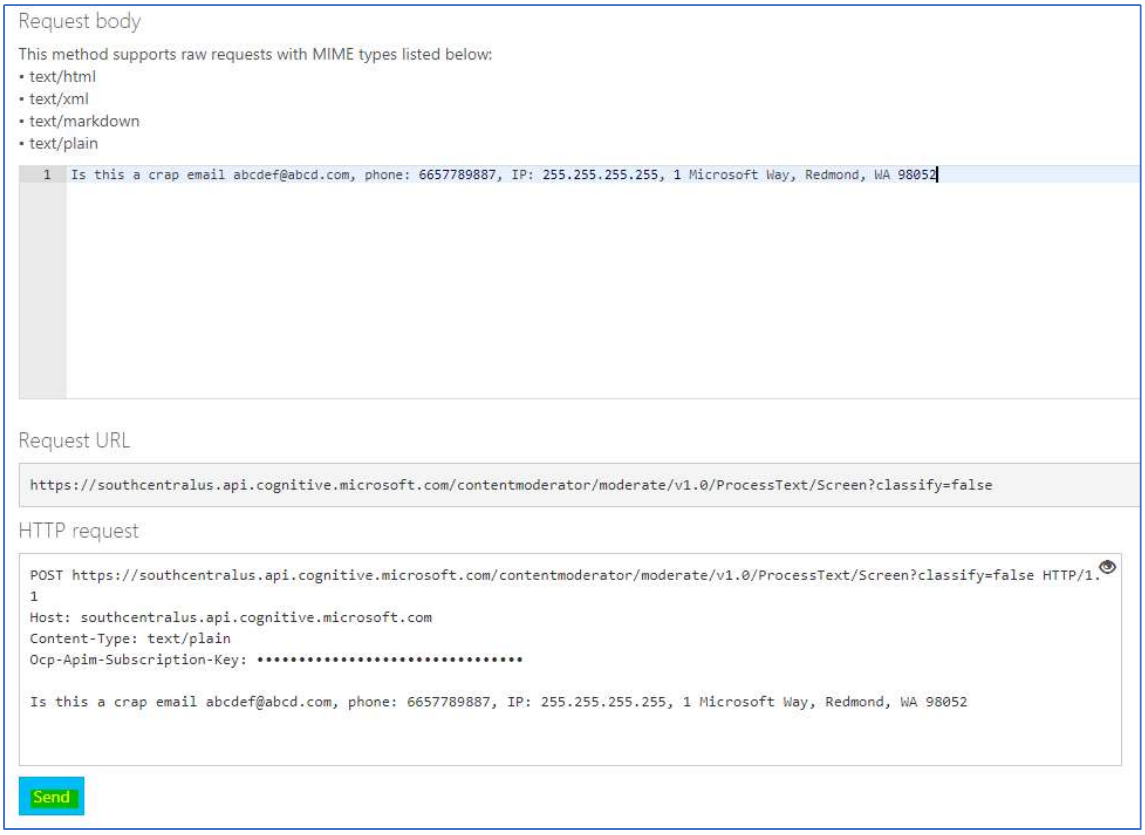
Query parameters

autocorrect	Value
PII	Value
listId	Value
classify	false
language	Value

Headers

Content-Type	text/plain
Ocp-Apim-Subscription-Key	································

5. Leave the sample text in place, and then click **Send**.



Request body

This method supports raw requests with MIME types listed below:

- text/html
- text/xml
- text/markdown
- text/plain

```
Is this a crap email abcdef@abcd.com, phone: 6657789887, IP: 255.255.255.255, 1 Microsoft Way, Redmond, WA 98052
```

Request URL

```
https://southcentralus.api.cognitive.microsoft.com/contentmoderator/moderate/v1.0/ProcessText/Screen?classify=false
```

HTTP request

```
POST https://southcentralus.api.cognitive.microsoft.com/contentmoderator/moderate/v1.0/ProcessText/Screen?classify=false HTTP/1.
1
Host: southcentralus.api.cognitive.microsoft.com
Content-Type: text/plain
Ocp-Apim-Subscription-Key: ••••••••••••••••••••••••••••••••

Is this a crap email abcdef@abcd.com, phone: 6657789887, IP: 255.255.255.255, 1 Microsoft Way, Redmond, WA 98052
```

Evaluate the response

- Scroll down the page and evaluate the response from the testing console.

You will see the that the email, IP address, phone, and address values are under a JSON array value of PII. You did not have to set the PII value to true for this result.

Response status

200 OK

Response latency

132 ms

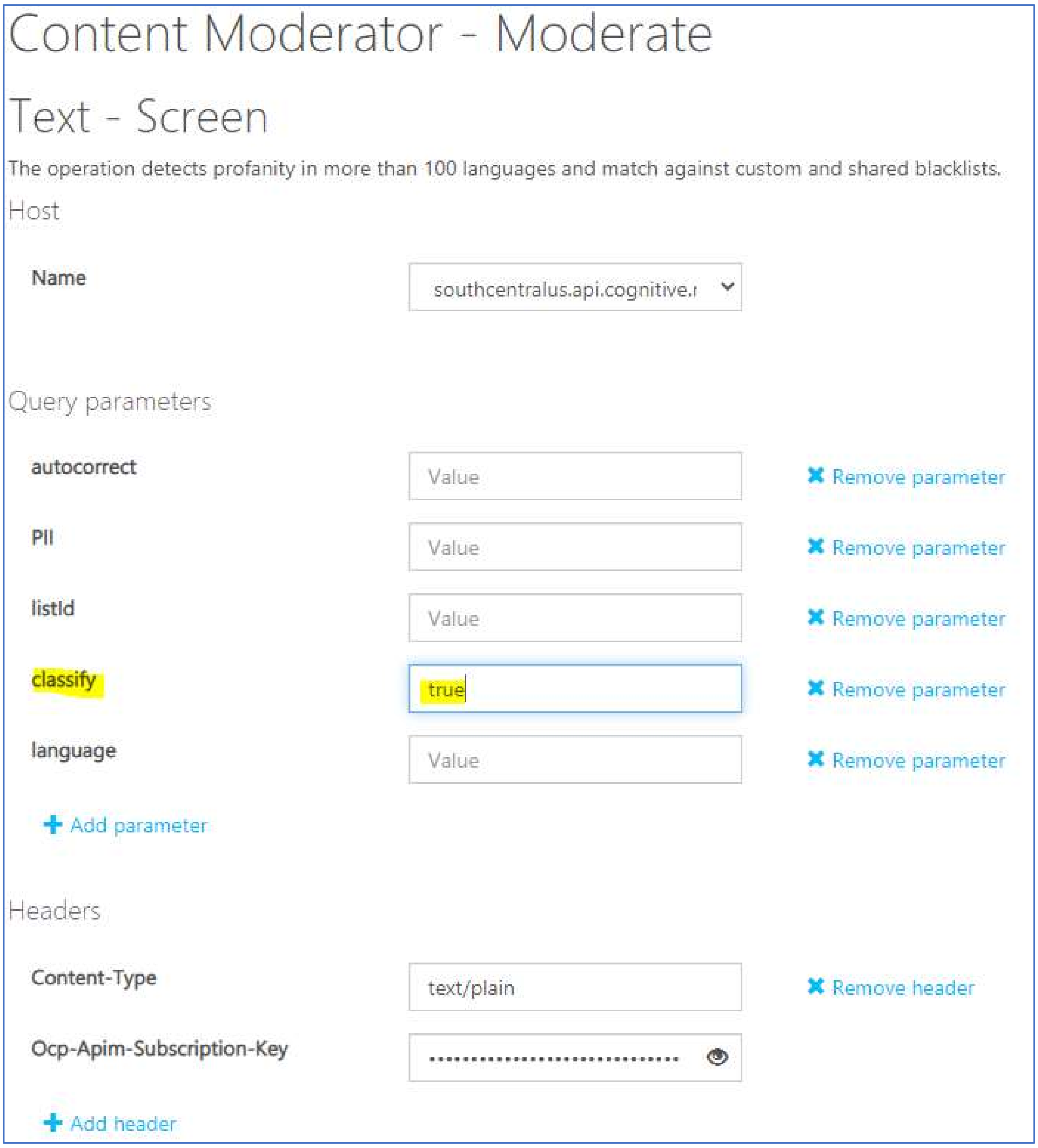
Response content

```
Pragma: no-cache

apim-request-id: 3bc63f52-0b1a-4f6a-84fa-956191796950

Strict-Transport-Security: max-age=31536000; includeSubDomains; preload

x-content-type-options: nosniff

CSP-Billing-Usage: CognitiveServices.ContentModerator.Transaction=1
```

```
Cache-Control: no-cache
Date: Sun, 12 Jul 2020 16:41:42 GMT
X-AspNet-Version: 4.0.30319
X-Powered-By: ASP.NET
Content-Length: 851
Content-Type: application/json; charset=utf-8
Expires: -1

{
  "OriginalText": "Is this a crap email abcdef@abcd.com, phone: 6657789887, IP: 255.255.255.255, 1 Microsoft Way, Redmond, WA 98052",
  "NormalizedText": "   crap email abcdef@abcd.com, phone: 6657789887, IP: 255.255.255.255, 1 Microsoft Way, Redmond, WA 98052",
  "Misrepresentation": null,
  "PII": {
    "Email": [{
      "Detected": "abcdef@abcd.com",
      "SubType": "Regular",
      "Text": "abcdef@abcd.com",
      "Index": 21
    }],
    "IPA": [{
      "SubType": "IPV4",
      "Text": "255.255.255.255",
      "Index": 61
    }],
    "Phone": [{
      "CountryCode": "US",
      "Text": "6657789887",
      "Index": 45
    }],
    "Address": [{
      "Text": "1 Microsoft Way, Redmond, WA 98052",
      "Index": 78
    }],
    "SSN": []
  },
  "Language": "eng",
  "Terms": [{
    "Index": 3,
    "OriginalIndex": 10,
    "ListId": 0,
    "Term": "crap"
  }],
  "Status": {
    "Code": 3000,
    "Description": "OK",
    "Exception": null
  },
  "TrackingId": "USSC_ibiza_d4ce52bb-e3d4-4170-ad4c-f143f1e8494e_ContentModerator.F0_3475a85c-5793-4aba-bbdf-11f3f7f965e1"
}
```

Run additional tests

1. To run the second test, scroll to the top of the page and set the `classify` parameter to `true`.



Content Moderator - Moderate

Text - Screen

The operation detects profanity in more than 100 languages and match against custom and shared blacklists.

Host

Name	southcentralus.api.cognitive.r	

Query parameters

autocorrect	Value	✖ Remove parameter
PII	Value	✖ Remove parameter
listId	Value	✖ Remove parameter
classify	true	✖ Remove parameter
language	Value	✖ Remove parameter

Headers

Content-Type	text/plain	✖ Remove header
Ocp-Apim-Subscription-Key	••••••••••••••••••••••••••••••	

2. Select **Send**.

   Note that there is now a new JSON array section title **Classification**. It indicates that a review is recommended and displays three categories with score values. The categories are pertaining to the text content that may be undesirable.

   - Category 1 - content could be sexually explicit or adult related
   - Category 2 - language may be considered sexually suggestive or mature in certain situations
   - Category 3 - potentially offensive language

Response status

200 OK

Response latency

53 ms

Response content

```
Pragma: no-cache

apim-request-id: 7d8a5433-3194-4d22-bf47-e010c197eff7

Strict-Transport-Security: max-age=31536000; includeSubDomains; preload

x-content-type-options: nosniff

CSP-Billing-Usage: CognitiveServices.ContentModerator.Transaction=1

Cache-Control: no-cache

Date: Sun, 12 Jul 2020 16:43:48 GMT

X-AspNet-Version: 4.0.30319

X-Powered-By: ASP.NET

Content-Length: 1024

Content-Type: application/json; charset=utf-8

Expires: -1


{

  "OriginalText": "Is this a crap email abcdef@abcd.com, phone: 6657789887, IP: 255.255.255.255, 1 Microsoft Way, Redmond, WA 98052",

  "NormalizedText": "   crap email abcdef@abcd.com, phone: 6657789887, IP: 255.255.255.255, 1 Microsoft Way, Redmond, WA 98052",

  "Misrepresentation": null,

  "PII": {

    "Email": [{

      "Detected": "abcdef@abcd.com",

      "SubType": "Regular",

      "Text": "abcdef@abcd.com",

      "Index": 21

    }],

    "IPA": [{

      "SubType": "IPV4",

      "Text": "255.255.255.255",

      "Index": 61
    }],
    "Phone": [{
      "CountryCode": "US",
      "Text": "6657789887",
      "Index": 45
    }],
    "Address": [{
      "Text": "1 Microsoft Way, Redmond, WA 98052",
      "Index": 78
    }],
    "SSN": []
  },
  "Classification": {
    "ReviewRecommended": true,
    "Category1": {
      "Score": 0.00040505084325559437
    },
    "Category2": {
      "Score": 0.22345089912414551
    },
    "Category3": {
      "Score": 0.98799997568130493
    }
  },
  "Language": "eng",
  "Terms": [{
    "Index": 3,
    "OriginalIndex": 10,
    "ListId": 0,
    "Term": "crap"
  }],
  "Status": {
    "Code": 3000,
    "Description": "OK",
    "Exception": null
  },
  "TrackingId": "USSC_ibiza_d4ce52bb-e3d4-4170-ad4c-f143f1e8494e_ContentModerator.F0_a4a2da9f-b0d4-47d0-b41b-ea81c75dc7b5"
}
```

3. To run additional tests, enter some of your own text values from an existing document and run the tests again to see the results returned.
4. Study the JSON response and the Request URL syntax to see how your custom applications can call this API.

Response status

200 OK

Response latency

103 ms

Response content

```
Pragma: no-cache
apim-request-id: c2a62b63-7103-4f33-bcdb-a0b25bac7d02
Strict-Transport-Security: max-age=31536000; includeSubDomains; preload
x-content-type-options: nosniff
CSP-Billing-Usage: CognitiveServices.ContentModerator.Transaction=1
Cache-Control: no-cache
Date: Sun, 12 Jul 2020 16:47:33 GMT
X-AspNet-Version: 4.0.30319
X-Powered-By: ASP.NET
Content-Length: 716
Content-Type: application/json; charset=utf-8
Expires: -1

{
  "OriginalText": "I ! My Security ID is 993-98-9992",
  "NormalizedText": "I  !  Security ID  993-98-9992",
  "Misrepresentation": null,
  "Classification": {
    "ReviewRecommended": true,
    "Category1": {
      "Score": 0.507390022277832
    },
    "Category2": {
      "Score": 0.26112818717956543
    },
    "Category3": {
      "Score": 0.98799997568130493
    }
  },
  "Language": "eng",
  "Terms": [{
    "Index": 2,
    "OriginalIndex": 2,
    "ListId": 0,
    "Term": "f***"
  }, {
    "Index": 18,
    "OriginalIndex": 20,
    "ListId": 0,
    "Term": "b****"
  }, {
    "Index": 2,
    "OriginalIndex": 2,
    "ListId": 0,
    "Term": "f******"
  }],
  "Status": {
    "Code": 3000,
    "Description": "OK",
    "Exception": null
  },
  "TrackingId": "USSC_ibiza_d4ce52bb-e3d4-4170-ad4c-f143f1e8494e_ContentModerator.F0_a5b491f7-5c81-4e9c-adbf-6511d5197169"
}
```

Tip

To test this API by using a C# application, see **Quickstart: Analyze text content for objectionable material in C#**.

Knowledge check

Check your knowledge

1. Under the classification of the text, what type of value is returned for the category?

○ A percentage indicating which category is most appropriate

○ A value between 0 and 1

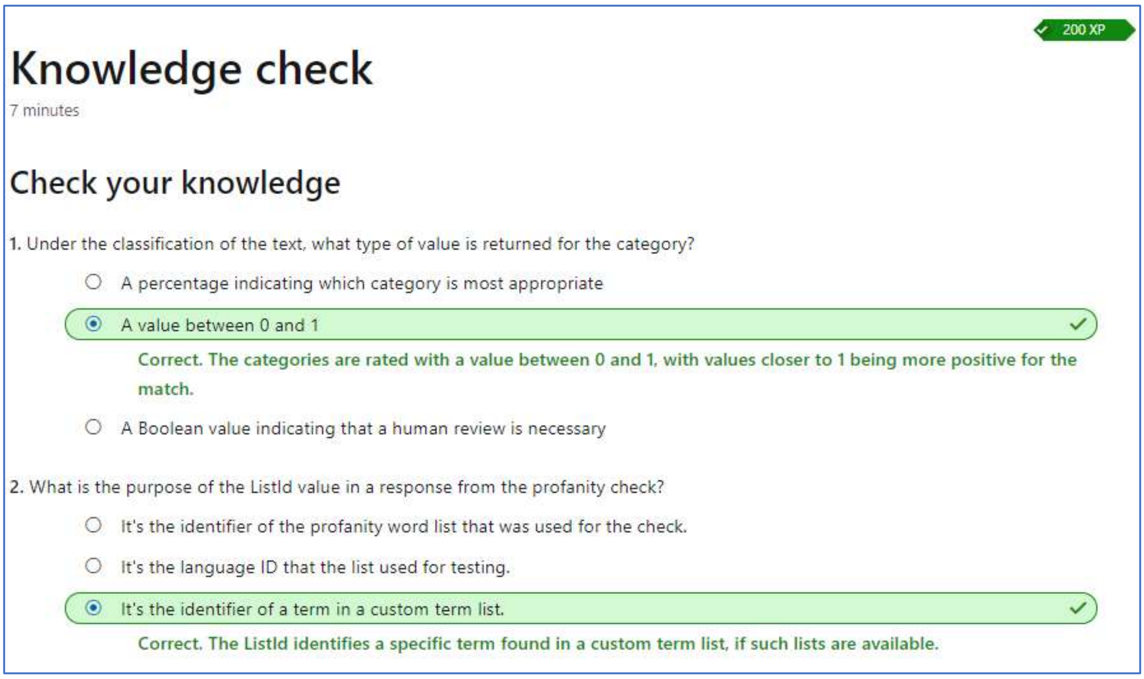
○ A Boolean value indicating that a human review is necessary

2. What is the purpose of the ListId value in a response from the profanity check?

○ It's the identifier of the profanity word list that was used for the check.

○ It's the language ID that the list used for testing.

○ It's the identifier of a term in a custom term list.



200 XP

Knowledge check

7 minutes

Check your knowledge

1. Under the classification of the text, what type of value is returned for the category?

○ A percentage indicating which category is most appropriate

● A value between 0 and 1 ✓

Correct. The categories are rated with a value between 0 and 1, with values closer to 1 being more positive for the match.

○ A Boolean value indicating that a human review is necessary

2. What is the purpose of the ListId value in a response from the profanity check?

○ It's the identifier of the profanity word list that was used for the check.

○ It's the language ID that the list used for testing.

● It's the identifier of a term in a custom term list. ✓

Correct. The ListId identifies a specific term found in a custom term list, if such lists are available.

Summary

As you've seen in this module, it's easy to add text filters to your apps and services with Microsoft Azure Content Moderator. With the Text Moderation API, you can:

- Analyze text to look for unwanted content.
- Classify the potentially offensive content.
- Get insights into the potential PII that's being shared so that you can protect it.

Cleanup

To avoid any unexpected costs in your Azure account, delete the **LearnRG** resource group. Deleting this group will remove all the resources we created in this module. Here are the steps you need to take:

1. Sign in to the Azure portal .
2. In the left pane, select **Resource groups**, and find the **LearnRG** resource group.
3. Select the resource group, and either right-click the row or use the **ellipsis** (...) button to open the context menu.
4. Select **Delete resource group**.
5. Enter the name of the **LearnRG** resource group, and then select **Delete**. Azure will remove all the resources for you.