Abstract

Quickly extract document data using a sample Content Analyzer API application.

Learn how to use the sample application written with Content Analyzer APIs to classify documents and extract unstructured document data. You can learn to extend the value of your current applications quickly with simple API calls. Content Analyzer APIs can help you get a more complete picture of the data in your documentation by applying AI algorithms to your data.

This sample application can be used as a companion tool to your new IBM Business Automation Content Analyzer cloud solution.

Introduction

The IBM Business Automation Content Analyzer is a new cloud-based API web service that can help you rapidly accelerate extraction and classification of data in your documents. Content Analyzer can digitize, classify and extract unstructured document content using OCR and PDF text extraction, and enable Watson and other AI technologies to reveal business insight from your documentation.

Once you have used the web interface to “train” your Content Analyzer instance to recognize your specific ontology of document classes, you will need to incorporate the Content Analyzer API calls into your workflow to integrate the data extraction and document classification capabilities.

Instead of waiting for your custom application to be written to call the Content Analyzer APIs, you can use this sample tool to get started right away. This sample tool can be installed on MacOS or Windows, and is written with the Content Analyzer RESTful APIs as a quick verification tool for your ontology and to jump start your integration. The intuitive user interface of the API Sample tool allows you to upload multiple documents, check status, download output, and clean up resources when you are done.

Audience: This tutorial is intended for the business analyst who have just customized an ontology of document classes for the enterprise, and is ready to hand off the system to the application developer who will be updating the business logic to integrate the Content Analyzer capabilities into the work flow. While the Content Analyzer web interface allows some API testing with the industry standard SwaggerUI interface, this API Sample tool will more closely simulate a production level API usage and help to verify that the system is ready for API use.

While this Sample tool does not show the source code, an application developer may still find the API Sample tool useful to see a usage flow of how the APIs process the documents.

Prerequisites

1. You must have access to a Content Analyzer cloud deployment. You will need a user ID and password. You can get this information from your Content Analyzer administrator. The administrator can use the IBM Digital Business Automation on Cloud user portal to create a functional user ID for you.
2. You might want to access the Content Analyzer Knowledge Center web page as a reference. The link is in the Related Links section below.
3. You will need a Content Analyzer API key and the API request URL. You can get this information from your Content Analyzer administrator. The administrator can use the Content Analyzer web interface to generate an API key specifically for you. This API key will identify you as the caller of the APIs.
4. You should also decide what output you want to be produced for each file: JSON, UTF-8 Text, and/or PDF. The JSON output will contain the extracted key-value pair information. The UTF-8 Text will be the raw OCR text results. The PDF will be an enhanced searchable PDF.
5. If you selected JSON output, you should know what subset of JSON options you want to be included. Select all or see documentation for details.

   ## Acronyms

   The acronymns below can be used to add more options to JSON options.

Abbreviation	Full Meaning	Conditions
OCR	Optical Character Recognition	
KVP	Key Value Pair	
MT	Mandatory	MT can only be used when SN is selected
DC	Document Classification	
HR	Headers	HR can only be used when OCR is selected
SN	Semantic Normalization	SN can only be used when DC and KVP selected
TH	Table Headers	TH can only be used when OCR is selected
WI	Watson Integration	WI can only be used when DC is selected
SHW	Sentence with headers (Watson Discovery)	SHW can only be used when HR is selected

6. You can put all the prerequisite configuration settings into a JSON format file that can be loaded into the sample tool. See the “Sample Configuration JSON”

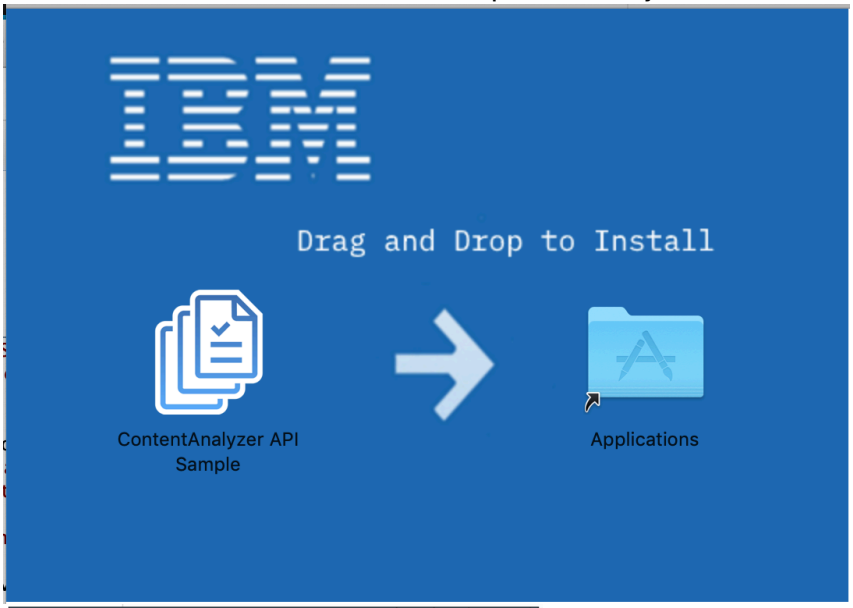
below. You can also input the same information directly into the sample tool user interface.

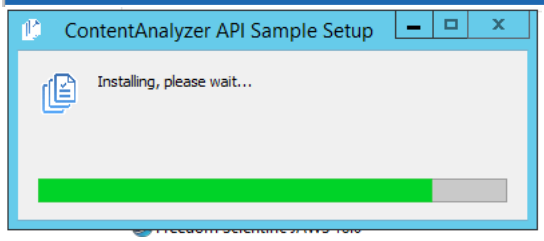
Sample Configuration JSON


```
{
"url":"https://9.30.123.456/backendsp/ca/rest/content/v1",
"apiKey":"273a3575-0961-4956-8ced-3789688e477b",
"jsonOptions":["OCR","DC","KVP","TH","HR","SN","MT","WI","SHW"],
"responseTypeOptions":["JSON","PDF","UTF8"],
"logpath":"/Users/janedoe/Library/Application Support/ContentAnalyzer/logs",
"SSO":"'SSO'", "functionId":"janedoe","password":"password4janedoe"
}
```

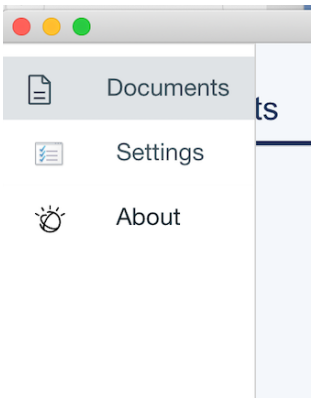

Steps

1. Download and install the API Sample tool to your MacOS or Windows system.

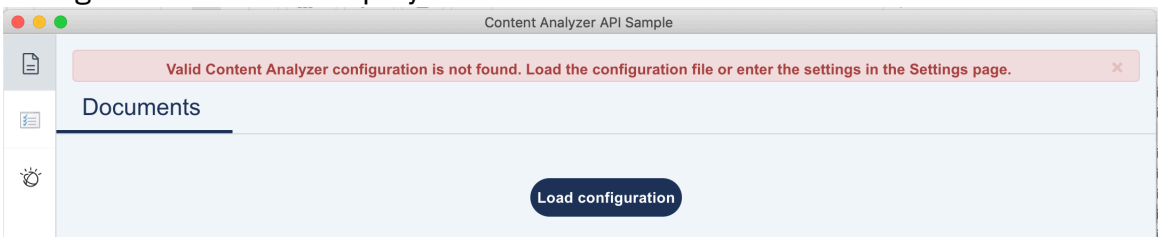




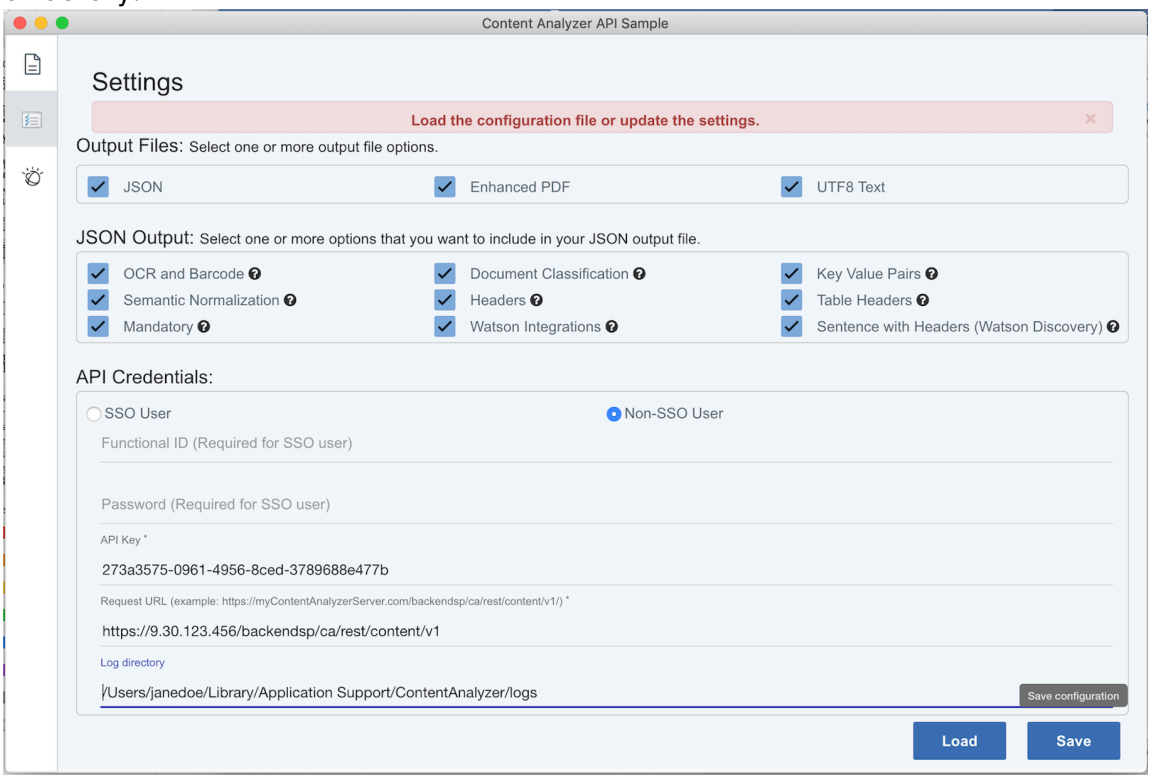
2. Start the API Sample tool.
3. The API Sample tool has a navigation bar on the left with three menu items, Document, Settings, and About.



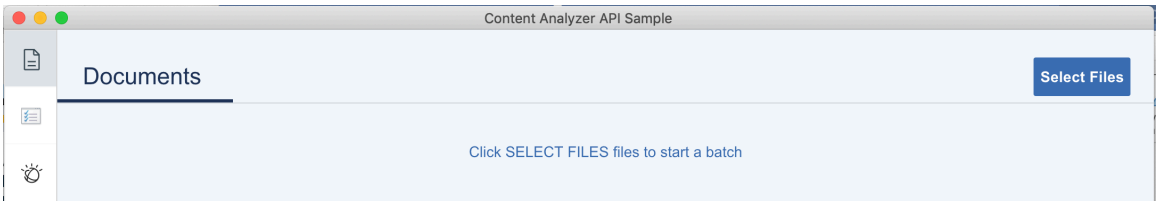
4. When you start the API Sample tool for the first time, the tool will not find any configuration and will display an error.



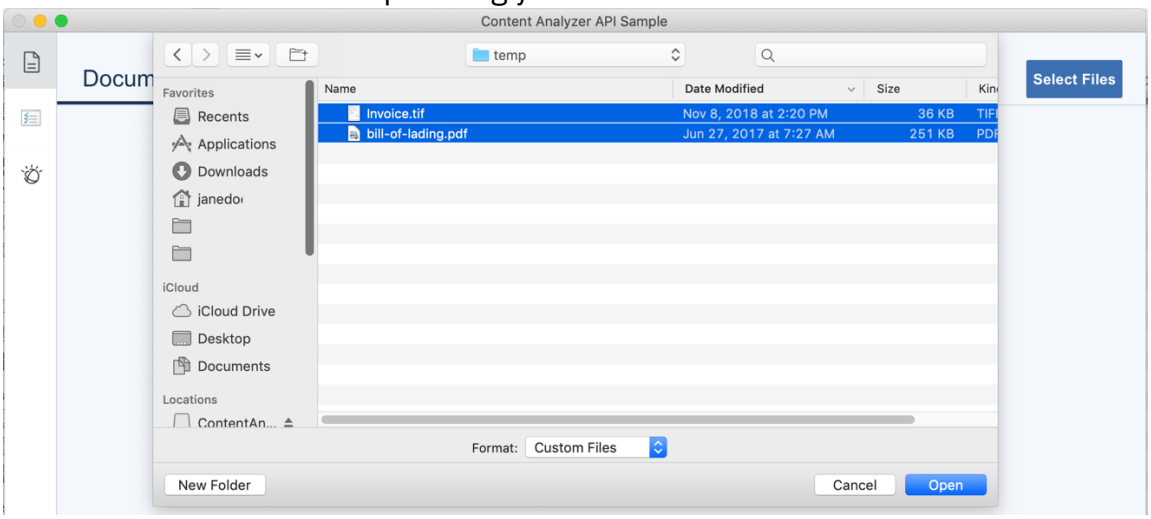
5. On the Document page, click “Load Configuration” to load your configuration JSON file. If you have not created the file earlier, you can manually enter the same information in the Settings page, and save it. If you leave the Log directory blank, the system will create a default directory in your system’s user data directory.



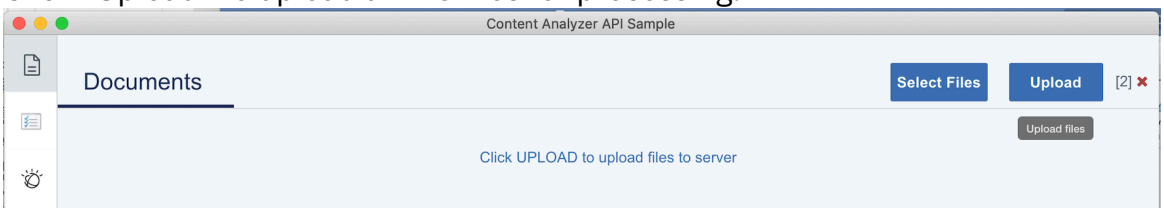
6. Once you have updated the settings, return to the Document page.



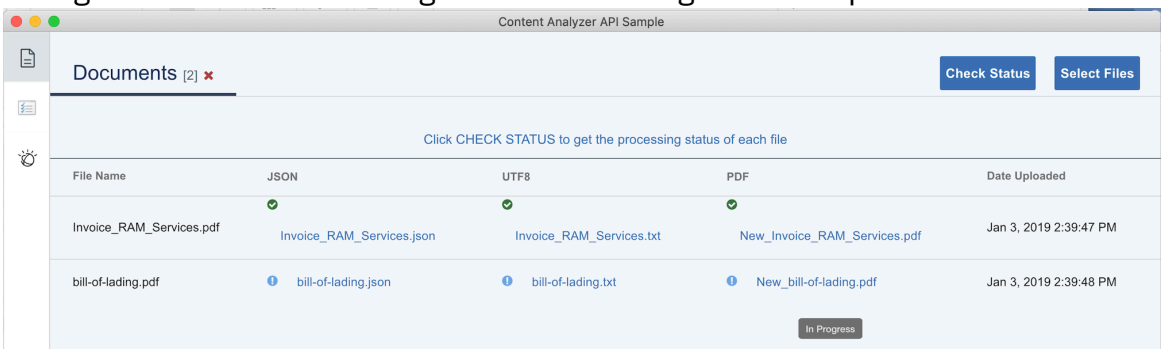
7. On the Document page, click “Select Files” and select one or more files to process. The input file types can be PDF, TIFF, JPG, or PNG. You can repeatedly select more files without uploading yet.



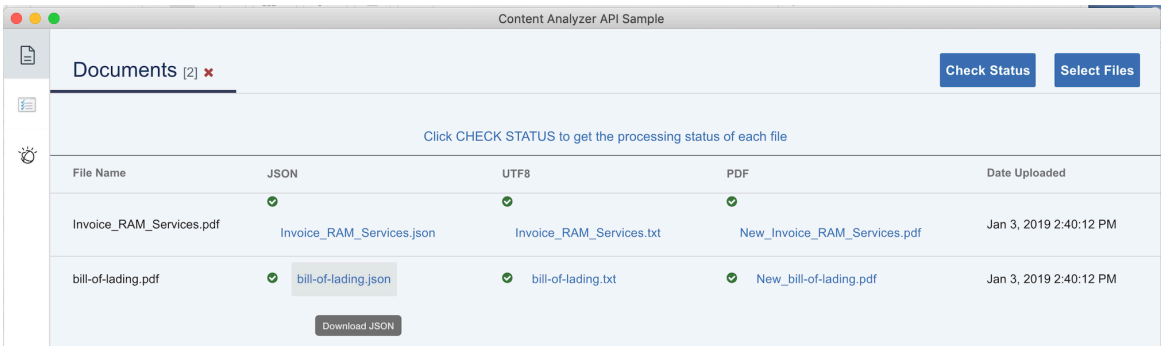
8. Click “Upload” to upload all the files for processing.



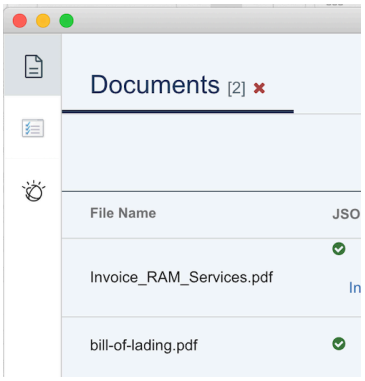
9. Click the “Check Status” to periodically monitor the status. Wait until your files change from the blue “In Progress” icon to the green “Completed” icon.



10. Once completed, you can click on each of the output file names to download the files.



11. When you no longer need the file outputs, you can click the red Delete icon to remove the resources from the server.



12. These output files are purged from the Content Analyzer server on a daily basis. Once you exit the API Sample tool, the list of processed files from this session will not be saved. But you can always check the Sample tool's log file for your record.

Logging and Troubleshooting

The log file, ContentAnalyzer.log, will show the timestamps of the actions done by the Sample client and any errors. Check the log file location in the Settings page. Here is an example log showing uploading 1 file, checking status, downloading the output files, and deleting all processed files on the server.


```
{"uploadProcessing":"2019-01-03T23:04:31.520Z","filesForUpload":1}

{"status":{"code":202,"messageId":"CIWCA12001","message":"Content Analyzer
request was created"},"data":{"message":"json,pdf,utf8 processing request was
created successful","fileNameIn":"bill-of-lading.pdf","analyzerId":"ee995040-
0fab-11e9-afda-8f436bdeba74","type":["json","pdf","utf8"]}}

{"analyzerId":"ee995040-0fab-11e9-afda-
8f436bdeba74","uniqueId":"","creationDate":"2019-01-
03T23:04:32.912Z","fileName":"bill-of-
```

lading.pdf","numPages":1,"statusDetails":[{"type":"JSON","status":"InProgress"
,"completedPages":0,"progress":8},{"type":"PDF","status":"InProgress"},{"type"
:"UTF8","status":"InProgress"}]}

{"analyzerId":"ee995040-0fab-11e9-afda-
8f436bdeba74","uniqueId":"","creationDate":"2019-01-
03T23:04:53.124Z","fileName":"bill-of-
lading.pdf","numPages":1,"statusDetails":[{"type":"JSON","status":"Completed",
"completedPages":1,"progress":100},{"type":"PDF","status":"Completed"},{"type"
:"UTF8","status":"Completed"}]}

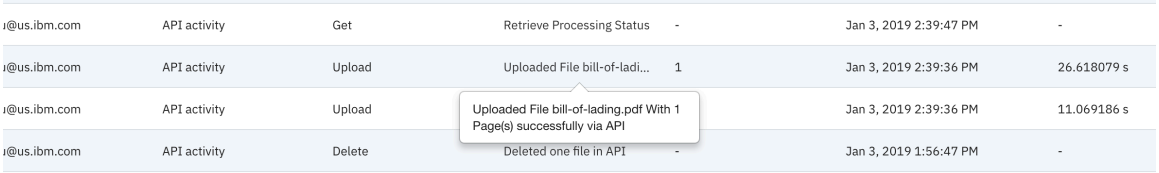
{"downloadJSON":"2019-01-03T23:04:59.772Z","transactionID":"ee995040-0fab-
11e9-afda-8f436bdeba74"}

{"downloadUTF8":"2019-01-03T23:05:03.924Z","transactionID":"ee995040-0fab-
11e9-afda-8f436bdeba74"}

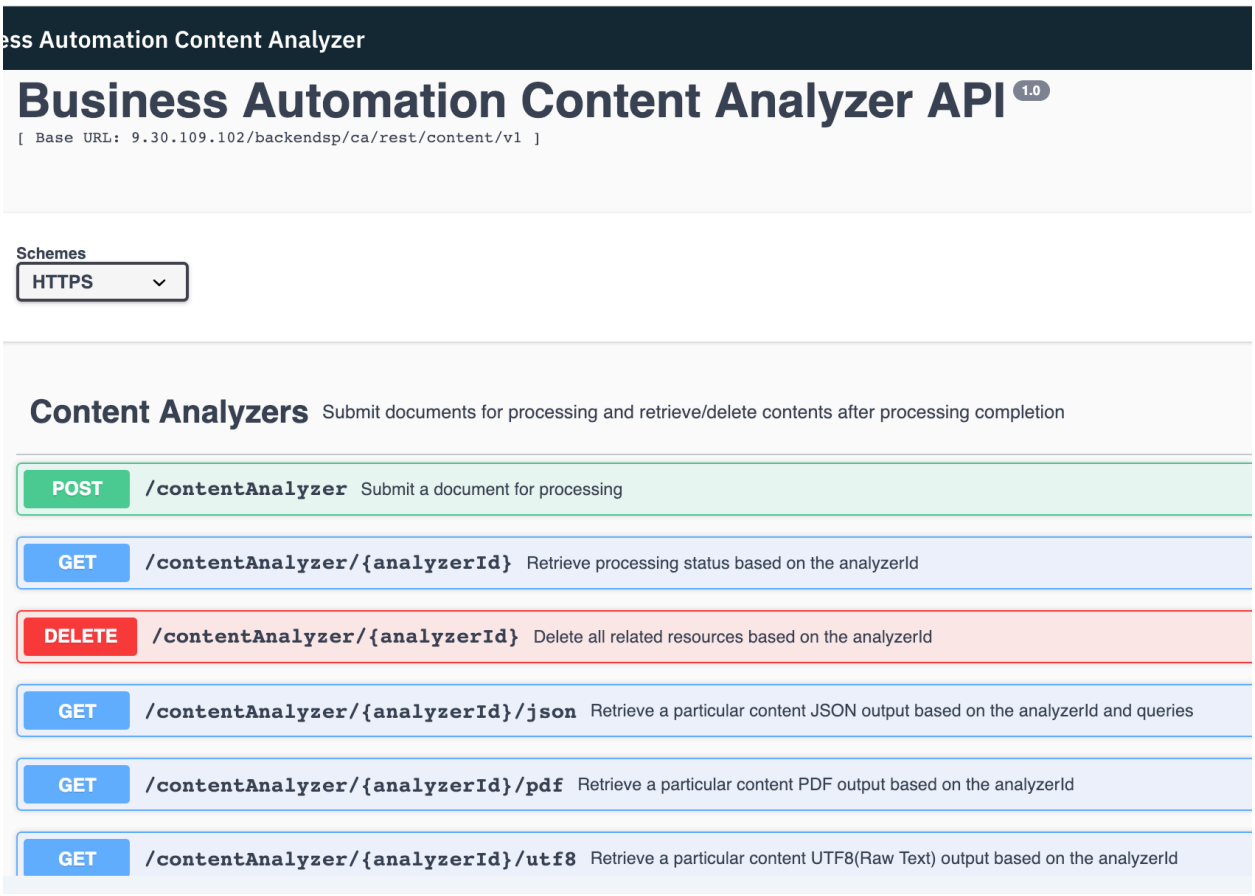
{"downloadPDF":"2019-01-03T23:05:09.598Z","transactionID":"ee995040-0fab-11e9-
afda-8f436bdeba74"}

{"deleteProcessing":"2019-01-03T23:05:13.718Z","filesForDelete":1}

The ContentAnalyzer web interface will also show the API activities in the History page (Filter on ‘API Activities’).



Content Analyzer APIs



Related links

1. https://www.ibm.com/support/knowledgecenter/SSUM7G/com.ibm.bacanalyzertoc.doc/bacanalyzer_1.0.html
2. The Content Analyzer web interface also has the API documentation in the API page.

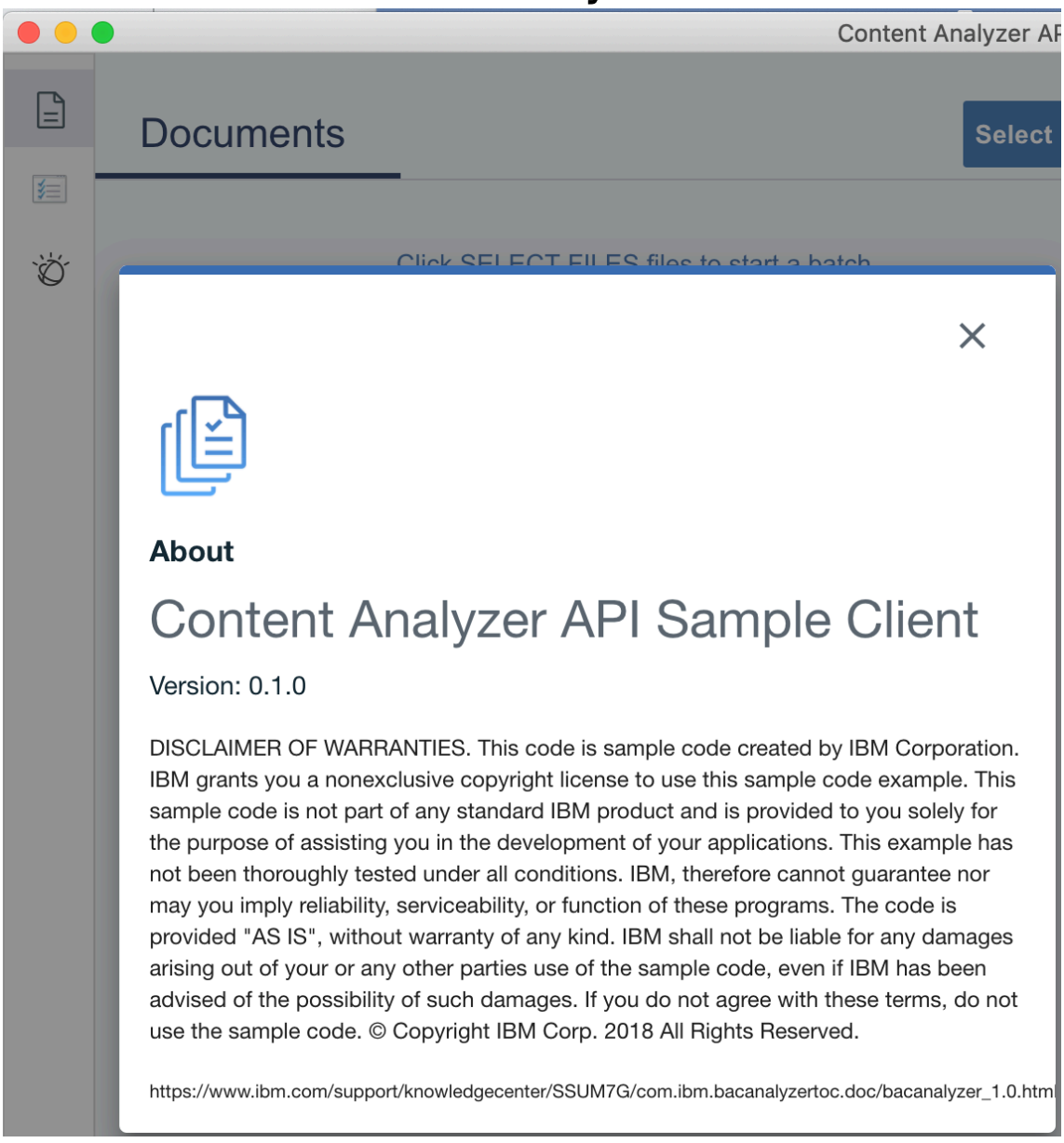
Downloadable resources

ContentAnalyzer API Sample.dmg for MacOS. (64MB)
ContentAnalyzer API Sample.exe for Windows (47MB)

Summary

You should now be familiar with the Content Analyzer API Sample tool and can easily demonstrate the classification and extraction capabilities of the Content Analyzer APIs. This API Sample tool is provided ‘AS-IS’.

Disclaimer of Warranty



This code is sample code created by IBM Corporation. IBM grants you a nonexclusive copyright license to use this sample code example. This sample code is not part of any standard IBM product and is provided to you solely for the purpose of assisting you in the development of your applications. This example has not been thoroughly tested under all conditions. IBM, therefore cannot guarantee nor may you imply reliability, serviceability, or function of these programs. The code is provided "AS IS", without warranty of any kind. IBM shall not be liable for any damages arising out of your or any

other parties use of the sample code, even if IBM has been advised of the possibility of such damages. If you do not agree with these terms, do not use the sample code.