Intelligent Document Classifier

Installation and Configuration Guide

Version: 2.1

Disclaimer

This document contains information proprietary to Newgen Software Technologies Ltd. User may not disclose or use any proprietary information or use any part of this document without written permission from Newgen Software Technologies Ltd.

Newgen Software Technologies Ltd. makes no representations or warranties regarding any software or to the contents or use of this guide. It also specifically disclaims any express or implied warranties of merchantability, title, or fitness for any particular purpose. Even though Newgen Software Technologies Ltd. has tested the hardware and software and reviewed the documentation, it does not guarantee or imply that this document is error free or accurate regarding any particular specification. As a result, this product is sold as it is and user, the purchaser, is assuming the entire risk as to its quality and performance. Further, Newgen Software Technologies Ltd. reserves the right to revise this publication and make changes in its content without any obligation to notify any person, of such revisions or changes. Newgen Software Technologies Ltd. authorizes no Newgen agent, dealer or employee to make any modification, extension, or addition to the above statements.

Newgen Software Technologies Ltd. has attempted to supply trademark information about company names, products, and services mentioned in this document. Trademarks indicated below were derived from various sources.

Copyright © 2023 **Newgen Software Technologies Ltd.** All Rights Reserved.
No part of this publication may be reproduced and distributed without the prior permission of Newgen Software Technologies Ltd.

Newgen Software, Registered Office, New Delhi
E-44/13
Okhla Phase - II
New Delhi 110020
India
Phone: +91 1146 533 200
info@newgensoft.com

Contents

Preface **4**
Revision history 4
About this guide 4
Intended audience 4
Related documents 5
Documentation feedback 5
Initiating IDC deployment **6**
Installing IDC **7**
Prerequisites 7
Installation steps 8
Verifying successful installation 10
Post-installation steps 11
Uninstalling IDC **13**
Configuring IDC **14**
Prerequisite 14
Configuring Classifier 14
Configuring Verifier 17
Processing a document **18**
Collecting samples for each document type 18
Training the model 18

Preface

This chapter provides information about the purpose of this guide, details on the intended audience, revision history, and related documents for the *Intelligent Document Classifier (IDC) Installation and Configuration Guide*.

Revision history

Revision date	Description
August 2023	Initial publication

About this guide

This guide provides information on how to install IDC on the Windows operating system, post-installation steps, verifying the successful installation, and procedures to uninstall the product. This guide also includes information on how to configure the INI file parameters for customization. It also includes information on document processing.

Intended audience

This installation and configuration guide is intended for IT system administrators responsible for installing, deploying, and configuring IDC 2.1. The reader must be familiar with the Microsoft Windows Server. The reader must have administrator rights to install IDC 2.1.

Related documents

The following documents are related to the IDC Installation and Configuration Guide:

- Intelligent Document Classifier Release Notes
- Intelligent Document Classifier Deployment Guide
- Intelligent Document Classifier Developer Guide
- Intelligent Document Classifier Best Practices Guide

Documentation feedback

To provide feedback or any improvement suggestions on technical documentation, write an email to docs.feedback@newgen.co.in.

To help capture your feedback effectively, share the following information in your email:

- Document name
- Version
- Chapter, topic, or section
- Feedback or suggestions

Initiating IDC deployment

This section provides information on the initiating IDC deployment.

To prepare for the installation and deployment of IDC 2.1, follow the below steps:

1. Download the *IDC2.1-Development-KIT* folder containing all the necessary files and folders required for IDC deployment from the provided FTP location.
2. Verify that your system meets the given minimum software and hardware requirements as specified in the Prerequisites section.
3. Execute the *AVXcheck.exe* file to verify whether your machine's CPU supports AVX. If the status indicates support (Yes), proceed with the subsequent steps; otherwise, replace the CPU with an AVX-compatible one.
4. Verify that ports 8015, 8016, and 27017 are open before execution.
5. Initiate the installation of IDC 2.1 using the automated installer. For procedural details, refer to the Installing IDC section.
6. After a successful installation, perform the post-installation steps mentioned in the Post-installation steps section.

Installing IDC

This section provides information about the prerequisites and the installation process of IDC 2.1.

Prerequisites

The following prerequisites are required before installing IDC 2.1:

System requirement	Description
Supported JDK versions	• JDK 1.7 • JDK 1.8
Supported operating systems	• Windows Enterprise Server 2016 • Windows Enterprise Server 2019
Supported browsers	• Microsoft Edge 113 and later • Google Chrome 113 and later • Mozilla Firefox 113 and later
Supported Python version	Python 3.8.10
Supported Visual C++ Redistributable package	Visual C++ Redistributable 2015-2019 package
Supported MongoDB version	MongoDB 4.0.8
Supported Tesseract OCR version	Tesseract OCR v4.0.0.20181030
Supported Miniconda version	Miniconda3-4.5.4
Administrative rights	• You must have administrative rights to install IDC 2.1. • You must have modification rights on the IDC installation directory, JDK directory, and Windows system folders.
Supported language	English (United States)

Before proceeding with the IDC installation, ensure that you have a reliable internet connection. If you encounter any issues with internet access, we suggest using the containerized version of IDC for optimal performance.

For procedural details on deploying IDC container services, refer to the *Intelligent Document Classifier Deployment Guide*.

Installation steps

To install IDC 2.1, follow the below steps:

1. Right-click *IntelligentDocumentClassifier2.1.exe* and select **Run as administrator**. If prompted for installation confirmation, click **Yes**. The installer Wizard progress bar appears.
   Once the setup application gets fully loaded, the Introduction screen appears.
2. Click **Next**. The License Agreement screen appears. Read the terms and conditions for using the software.
3. Select the **I accept the terms of the License Agreement** checkbox to agree with the terms.
4. Click **Next**. The Prerequisites screen appears.
5. Select the respective checkboxes to confirm the following prerequisites are met:
   - Microsoft Visual C++ Redistributable for Visual Studio 2015, 2017, and 2019 are installed.
   - Miniconda3-4.5.4 is installed.
   - MongoDB Community Server v4.x is installed.
   - Tesseract-OCR v4.0.0.20181030 is installed.
6. Click **Next**. The Classifier/Verifier Configuration screen appears.
7. Specify the following classifier and verifier configuration details:

Field	Description
Classifier/Verifier IP	Specify the IP address of the machine where the classifier and verifier are running.
Classifier Port	Specify the port number where the classifier is running.
Verifier Port	Specify the port number where the verifier is running.

8. Click **Next**. The MongoDB Home Path screen appears.
9. Click **Choose** to select the path of the MongoDB Home folder. For example, *C:\Program Files\MongoDB\Server\4.0*
10. Click **Next**. The MongoDB Configuration screen appears.
11. Specify the MongoDB configurations details:

Field	Description
MongoDB Cabinet Name	Specify the name of the MongoDB cabinet. The cabinet name must be written without any spaces.
MongoDB Host	Specify the IP address of the machine where MongoDB is running.
MongoDB Port	Specify the port number where MongoDB server is running.

12. Click Next. The Tesseract OCR Path screen appears.
13. Click **Choose** to select the path of the Tesseract OCR folder.
14. Click **Next**. The Choose Install Folder screen appears.
15. Select a folder to save the IDC 2.1 installation. You can keep the default installation folder for the product or click **Choose** to browse for and select a different folder.
16. Click **Next**. The Pre-Installation Summary screen appears. Review the pre-Installation summary before you proceed to the next step.
17. Click **Install**. The Installing Intelligent Document Classifier 2.1 screen appears. It displays the progress of the files being copied to the installation folder.
    Once all the files get copied to the destination location, the Install Complete screen appears.
18. Click **Done** to complete the installation and close the installer.

For procedural details on post-installation configurations, refer to the Post-installation steps section.

Verifying successful installation

To verify the successful installation of IDC 2.1, follow the below steps:

1. Ensure that the installation process completes without any errors. Upon successful installation, the following folders and files get created in the IDC 2.1 installation directory:
   - Folders —
     - Classifier
     - Client Component
     - jre
     - Predict-Data
     - Submit-Data
     - Uninstall IntelligentDocument Classifier2.1
     - Verifier
   - Files —
     - IDC_version
     - Pre-requisite setup instructions
2. Open the *Uninstall IntelligentDocument Classifier2.1* folder and navigate to the *Logs* folder. Then, open the *IntelligentDocumentClassifier2.1_Install_<Date of installation>_<Time of installation>* file to review the installation logs and confirm that the installation was error-free.

Post-installation steps

This section provides information on the steps to be followed after the installation process, including the activation of the license key.

After the successful installation of IDC 2.1, follow the below steps:

1. Obtain the *Model* directory generated through model training and replace the *Model* folder present inside the IDC 2.1 installation home path with the trained Model.

   Ensure to turn off the Python server (if currently running) before replacing the Model.

2. Execute the *setup.bat* file present inside both *Verifier* and *Classifier* directories in the IDC 2.1 installation home path to complete the IDC 2.1 server requirement installation.

   - The default path of the classifier directory is *C:\ProgramFiles\IntelligentDocumentClassifier2.1\Classifier*
   - The default path of the verifier directory is *C:\ProgramFiles\IntelligentDocumentClassifier2.1\Verifier*.

3. Execute the *RunMongoDB.bat* file present inside the IDC 2.1 installation home path to start the MongoDB server.

   Before executing the *RunMongoDB.bat* file, ensure to terminate any existing MongoDB instances. You might need to stop it from services. To do this, search for *services.msc* file using the Windows search bar, locate the running MongoDB server from the list. Right-click the MognoDB server name and select Stop.

4. Execute the *server.bat* file present inside both *Verifier* and *Classifier* directories in the IDC 2.1 installation home path to start the Python server.
5. Launch your web browser and enter the Intelligent Document Classifier Admin Module URL in the browser's address bar.
   The format of the URL is *http://<ip_address>:<port_number>* where the *ip_addresss* and *port_number* represent the IP address and port number of the machine on which IDC is installed.
6. Sign in to the admin module using the default credentials:
   • Username — admin

- Password — system123#
  Upon successful sign-in, the Intelligent Document Classifier - Admin Module home page appears.

7. From the left pane, select **Subscription**. The Current Subscription Details page appears. It displays the comprehensive details of the currently active subscription, if any.
8. Click **Add license** and provide the license key that you received during the IDC software purchase.
9. Click **Activate License** to initiate the license key activation process. Upon successful activation, a pop-up confirming the activation of the subscription license appears. Click **OK** to close it.
   The Current Subscription Details section now displays the details of active subscriptions.

   If an add-on is required, you can specify a new license in the **Add New License** field. This license can be used until it expires.

Within the Intelligent Document Classifier - Admin Module, you can access the following information through different tabs:

- The Subscription tab displays details of the currently active subscriptions and also provides the option to a add new license key.
- The Model Info tab provides information about the currently deployed Model.

  If the model is not available, you must provide the trained model. For procedural details on providing a trained model, follow step 1.

- The Usage Analysis tab allows you to view various reports, including:
  - Hourly Load
  - Daily Load
  - Response Time
  - Classwise Request
  - Classwise Confidence
  - Classification Response Accepted
  - Classification Response Not Accepted
  - Model Usage
  - Classification Response
  - Engine Response Status
- The Feedback tab allows you to view the classification feedback. This tab displays the confusion matrix where the rows represent the Predicted categories, and the columns represent the Actual categories.

Uninstalling IDC

To uninstall IDC 2.1, follow the below steps:

1. Go to the *IntelligentDocumentClassifier2.1* install directory.
2. Open the *Uninstall IntelligentDocumentClassifer2.1* directory.
3. Right-click the *Uninstall IntelligentDocumentClassifier2.1.exe* file and select **Run as administrator**.
   Depending on your Windows settings, you might be asked to confirm that you want to uninstall the software; if so, confirm that you wish to do so. The Prerequisites screen appears.
4. Select the respective checkbox to confirm that the following prerequisite is met: Intelligent Document Classifier services must be in stop mode.
5. Click **Next**. The Uninstall IntelligentDocumentClassifier2.1 screen appears.
6. Click **Uninstall**. The uninstallation begins.
   The Uninstall Complete screen appears upon the successful uninstallation of Intelligent Document Classifier 2.1.
7. Click **Done** to complete the uninstallation and close the uninstaller.

Configuring IDC

This section provides information about the prerequisites and configuration files for the Classifier and Verifier components in IDC 2.1. The configuration files are available at the following locations:

- **Configuration file (config.ini) for the Classifier** — *C:\ProgramFiles\IntelligentDocumentClassifier2.1\Classifier*
- **Configuration file (config.ini) for the Verifier** — *C:\ProgramFiles\IntelligentDocumentClassifier2.1\Verifier*

System administrators and other admin users are advised not to modify these configuration files until it is specifically required.

Prerequisite

IDC 2.1 must be successfully installed along with the necessary prerequisites.

Configuring Classifier

The following table provides information on the different tags available for customization in the *config.ini* file of the classifier:

Parameter	Description	Default value	Possible value
[Service]			
MODEL_IMAGE_PATH	This parameter contains the path of the trained image model.	Not applicable	Not applicable
MACHINE_IP	This parameter contains the IP address of the machine.	Not applicable	Not applicable

Parameter	Description	Default value	Possible value
MODEL_PORT	This parameter contains the port number on which the Classifier is running.	8016	
VERIFICATION_PORT	This parameter contains the port number on which Verifier is running	8015	
VERIFICATION_URL	This parameter contains the URL of the image model verification API.	Not applicable	Not applicable
MODEL_THREAD_POOL	This parameter determines the maximum number of threads of the web server.	10	1 to 10
LOG_IMAGE	This parameter enables or disables the storage of incoming request images.	no	yes or no
MODEL_PATH	This parameter contains the path of the trained Model folder.	Not applicable	Not applicable
MODEL_OBJECT_TYPE	This parameter determines the type of Object Model trained.	tiny	tiny
MODEL_OBJECT_PATH	This parameter contains the path of the trained Object Model.	Not applicable	Not applicable
SUBMIT_DATA_PATH	This parameter contains the path to save the feedback data.	Not applicable	Not applicable
PREDICT_DATA_PATH	This parameter contains the path to save the log images.	Not applicable	Not applicable
MODEL_TYPE	This parameter determines the type of Model to be used.	Content	Image, Content, Object, or Smart

Parameter	Description	Default value	Possible value
MODEL_HEADING	This parameter enables or disables the use of the headings-based content model.	N	Y or N
DEBUG	This parameter enables or disables logging.	Y	Y or N
[IMAGE]			
ORIENTATION	This parameter enables or disables the usage of different orientations for the image model.	N	Y or N
SEGMENTATION	This parameter enables or disables the usage of segmentation for the image model.	N	Y or N
[WRAPPER]			
DB_NAME	This parameter contains the name of the associated database cabinet.	Not applicable	Not applicable
DB_IP	This parameter contains the IP address on which the database is available.	Not applicable	Not applicable
DB_PORT	This parameter contains the port number of the database.	27017	

Configuring Verifier

The following table provides information on the different tags available for customization in the *config.ini* file of the verifier:

Parameter	Description	Default value	Possible values
[Service]			
MODEL_PORT	This parameter contains the port number on which the Classifier is running.	8016	
VERIFICATION_PORT	This parameter contains the port number on which the Verifier is running.	8015	
VERIFICATION_DATA_PATH	This parameter contains the path of the data to be used for verification.	Not applicable	Not applicable
MODEL_THREAD_POOL	This parameter determines the maximum number of threads of the web server.	10	1 to 10
LOG_IMAGE	This parameter enables or disables the storage of incoming request images.	no	yes or no
CONFIG	This parameter contains the path of the verification threshold file.	Not applicable	Not applicable

Processing a document

This section provides information on the steps involved in processing a document:

- Collecting samples for each type of document
- Training the model
- Configuring IDC (For more information on configuring IDC, refer to the Configuring IDC section)

Collecting samples for each document type

IDC 2.1 offers the capability to train the model for identifying various types of documents. Each of these document types represents a distinct class within the model. For successful training of the model, a minimum of 100 samples is required for each document type.

The accuracy of the model depends on several factors, including the quality of the document images and the sufficiency of training samples. To achieve higher accuracy, it is essential to have a substantial amount of sample training data for each category, and it must be equally distributed among all categories. Proper and accurate labeling of samples helps the model in more effective self-training.

Training the model

To train the model for better document classification, gather the dataset and ensure that the different data classes are kept separated in class-wise folders. For identity card-type documents, data labeling must be done using the LabelImg tool, defining various crucial regions within each document. The labeled data is then fed into the training platform.