WONDERBOTZ

DOCUMENT UNDERSTANDING

UIPATH

Version 1.0

Copyright

The information contained in this document is the proprietary and confidential information of WonderBotz LLC. It must not be disclosed to a third party without an authorized WonderBotz representative's written consent. No part of this document may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, without the written permission of WonderBotz LLC.

© 2022 WonderBotz LLC. All Rights Reserved.

All trademarks are now acknowledged and used to benefit their respective owners.

Published by:
WonderBotz LLC
821 Alexander Road
Suite 206
Princeton
NJ 08540-6352
United States
Web: www.WonderBotz.com
Email: Info@WonderBotz.com

Disclaimer

Optimum efforts have been taken to ensure the accuracy of this document. However, discrepancies may still occur, so kindly report documentation issues to the author at the earliest.

WonderBotz acknowledges no inaccuracies or errors in this document and reserves the right to update and revise the contents without prior notice.

Document Control

Version History

Author	Information contact(s)	Date
Lead Technical Writer	Senior Automation Developer	09-Feb-2022

Modified by	Version	Date	Comments
Lead Technical Writer	V 1.0	30-Dec-2022	Taxonomy Manager added.

Signoff Authority

Role	Date
Country Head	03-Mar-2022

Document Conventions

Convention	Meaning
Italics	Reference documents
Bold	The subject under the action, Emphasis
Italics, small font	Reference information and captions
`Courier New font`	User input and scripts
Pantone cool grey color	Note and caution body
Pantone cool grey color bold	Note
Highlight	Information to be updated
Pantone red bold	Caution

© 2022 WonderBotz LLC. All Rights Reserved.

Contents

1 Introduction 5
1.1 What is Document Understanding? 5
1.2 How does Document Understanding work? 5
1.3 Scope and Purpose 5
1.4 Prerequisites 6
2 Document Understanding Framework Component 8
2.1 Taxonomy 8
2.2 Digitization 8
2.3 Document Classification 8
2.4 Document Classification Validation 8
2.5 Classification Training 9
2.6 Data Extraction 9
2.7 Data Extraction Validation 9
2.8 Data Extraction Training 9
3 Practical Implementation 11
3.1 Package Creation 11
3.2 Taxonomy Manager 13
3.3 Digitize Document 15
3.4 Classification 16
3.5 Data Extraction 18
3.6 Present Validation Status 20
3.7 Export Extraction Data 20
APPENDICES 22
APPENDIX A: References 22
APPENDIX B: List of Figures 22

© 2022 WonderBotz LLC. All Rights Reserved.

1 Introduction

1.1 What is Document Understanding?

UiPath Document Understanding utilizes RPA and AI to process the documents automatically. It enables intelligent document processing within automation workflows, thus allowing the automation of complex and cognitive processes.

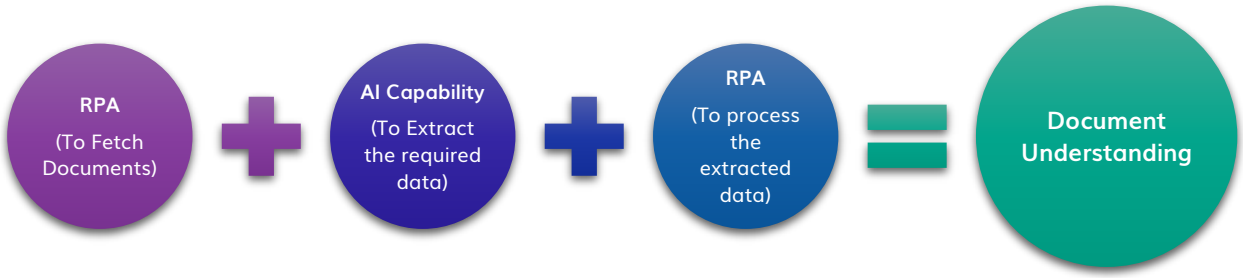


Figure 1: Document Understanding

1.2 How does Document Understanding work?

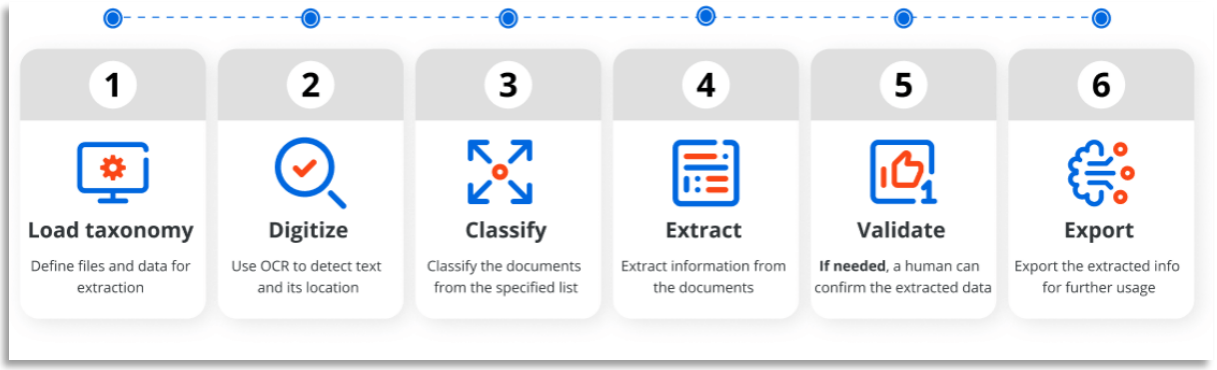


Figure 2: Document Understanding stages

Source: UiPath Forum, Document Understanding Best Practises/REFramwork, https://forum.uipath.com/

1.3 Scope and Purpose

This document aims to acquaint RPA professionals with the basic flow and terminologies involved in Document Understanding by guiding them through different stages to work on it in UiPath.

© 2022 WonderBotz LLC. All Rights Reserved.

1.4 Prerequisites

- UiPath.IntelligentOCR.Activities package.
  - UiPath.IntelligentOCR.Activities package contains a Document Understanding framework.
  - Taxonomy Manager wizard appears in the top ribbon of the UiPath Studio after the UiPath.IntelligentOCR.Activities package is installed.
  - UiPath.IntelligentOCR.Activities package contains all the core documents understanding framework activities.
  - **Navigation**: Manage Packages→All Packages

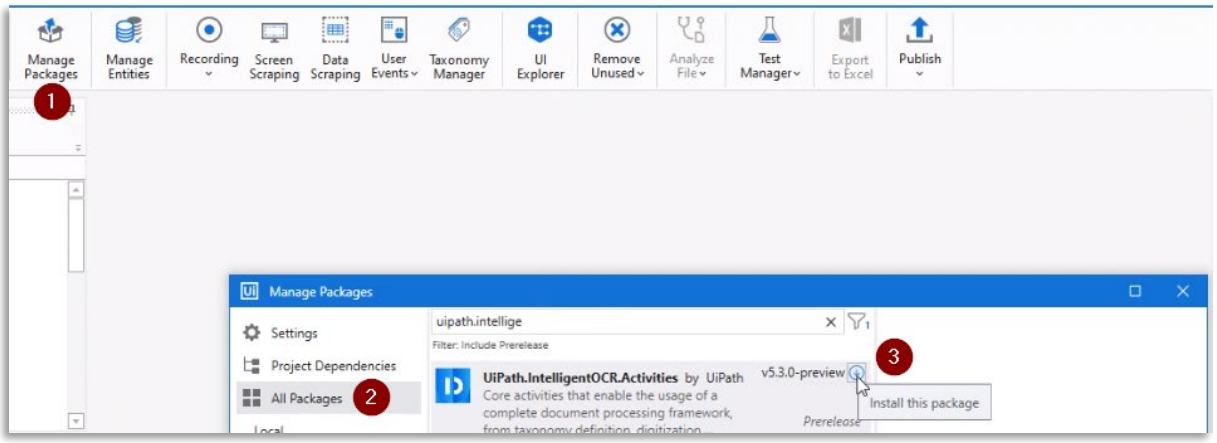


Figure 3: UiPath.IntelligentOCR.Activities package installation in UiPath

- OmniPage.Activities package to include OCR engine.

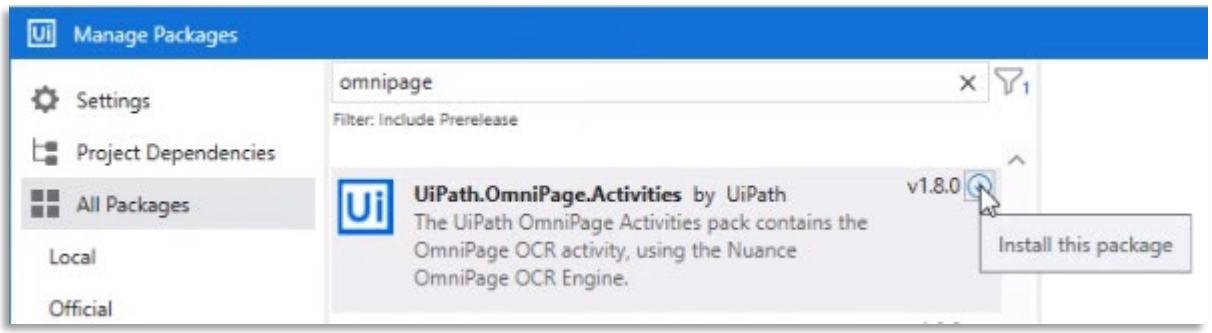


Figure 4: OmniPage.Activities package installation

© 2022 WonderBotz LLC. All Rights Reserved.

- DocumentUnderstanding.MLActivities package to include ML activities

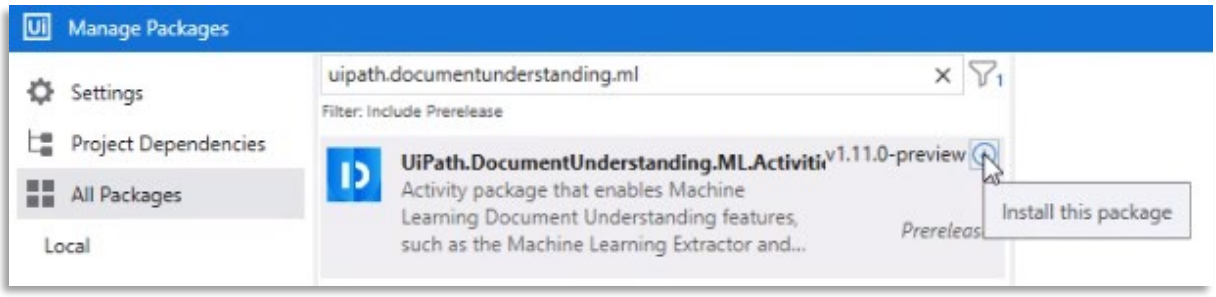


Figure 5: DocumentUnderstanding.MLActivities installation

- Document understanding API Key to use the services of document understanding.

Navigation: Orchestrator→Admin→Licenses→Robots & Services.

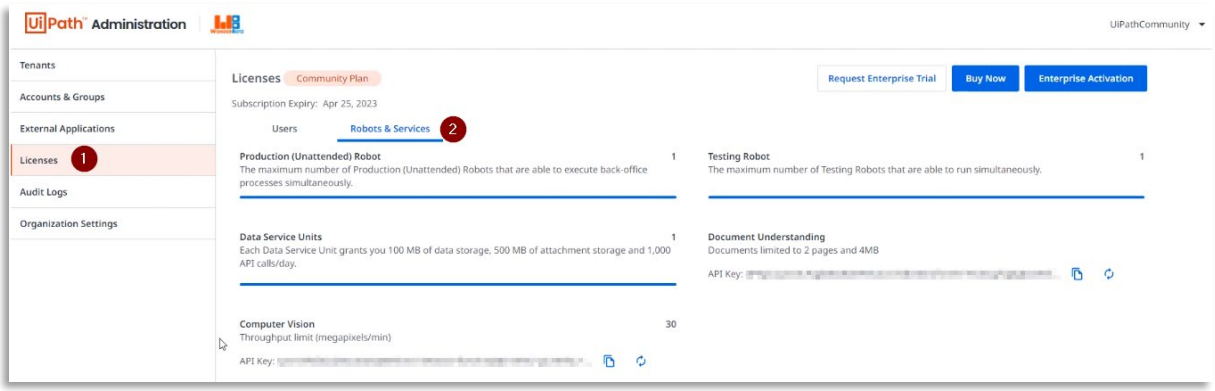


Figure 6: API Key

© 2022 WonderBotz LLC. All Rights Reserved.

2 Document Understanding Framework Component

2.1 Taxonomy

What documents must be processed? What data is required from them?

Taxonomy defines the document types and the pieces of information targeted for data extraction (fields) for each document type. It formalizes this information into a dedicated Taxonomy structure. This metadata information is managed through the Taxonomy Manager.

In this pre-processing step, you can add multiple document types and the fields you are interested in extracting. For example, you can work with Invoices to extract the vendor and the total amount and medical forms to extract the insured ID number and patient name.

2.2 Digitization

What does a file contain?

Digitization obtains the incoming document's textual content and structure, turning a file into machine-readable content for further downstream processing.

As the documents are processed, they go through the digitization process. The difference for non-digital (scanned) documents is that you must apply your choice's OCR engine. The outputs of this step are the Document Object Model and a string variable containing all the document text. They are passed down to the next steps.

2.3 Document Classification

What types of documents from the taxonomy are found in a file?

Document Classification automatically determines what document types are found within a digitized file.

After digitization, the document is classified. Suppose you are working with multiple document types in the same project. In that case, you need to know what type of document you are working with to extract data properly. The important thing is that you can use multiple classifiers in the same scope, configure them, and train them later in the framework. The classification results help in applying the right strategy in extraction.

2.4 Document Classification Validation

Is the predicted classification correct? (This is how I can review and correct it).

Document Classification Validation assists in the human validation and correction of the automatic classification and document splitting results.

© 2022 WonderBotz LLC. All Rights Reserved.

2.5 Classification Training

Did a human review the data? (This is how the robot can learn from it).

Classification Training passes the human-validated information back to the classifiers to improve their future predictions.

2.6 Data Extraction

What data can be found in a particular document?

Data Extraction captures the information required for the identified document type within the given input document and classification page range.

Extraction is getting just the required data. For example, extracting specific data from a 5-page document is quite troublesome if you want to do it with string manipulation. You can use different extractors in the same scope application for the different document structures in this framework. The extraction results are passed further for validation.

2.7 Data Extraction Validation

Is the extracted information correct? (This is how one can review and correct it).

Data Extraction Validation assists in the human validation and correction of the automatically extracted data results.

A human user can validate the extracted data through the Validation Station. The best practice is to build logic around the decision to add or not a human validation step, with rules depending on the specific use case. Validation results can then be exported and used in further automation activities.

2.8 Data Extraction Training

Did a human review the data? (This is how the robot can learn from it).

Data Extraction Training passes the human-validated extracted data back to the extractors to improve their extraction predictions.

Classification and extraction are as efficient as the classifiers and extractors used are. If a document was not appropriately classified, it was unknown to the active classifiers. The same way goes for incorrect data extraction. The framework provides the opportunity to train the classifiers and the extractors to improve recognition of the documents and fields.

© 2022 WonderBotz LLC. All Rights Reserved.

The figure below presents the Document Understanding Framework components and how they relate to one another:

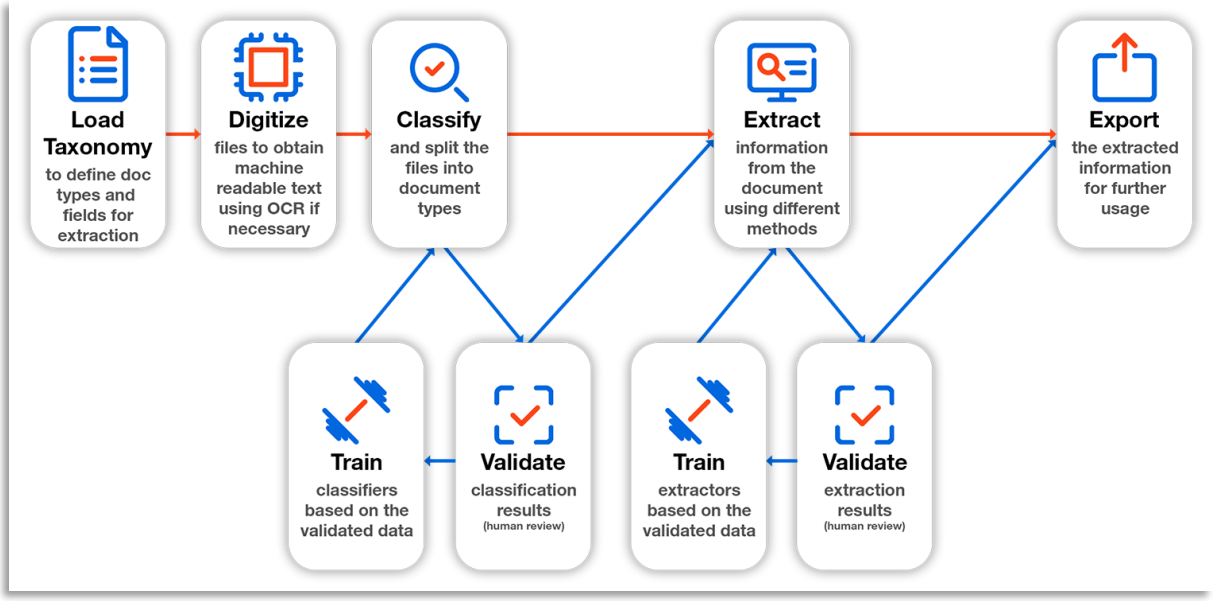


Figure 7: Document Understanding Framework Components

Source: Introduction, UiPath, https://docs.uipath.com/document-understanding/docs/introduction

© 2022 WonderBotz LLC. All Rights Reserved.

3 Practical Implementation

3.1 Package Creation

1. Create a **new project** in the AI Centre in the Orchestrator before starting the development in the UiPath Studio.

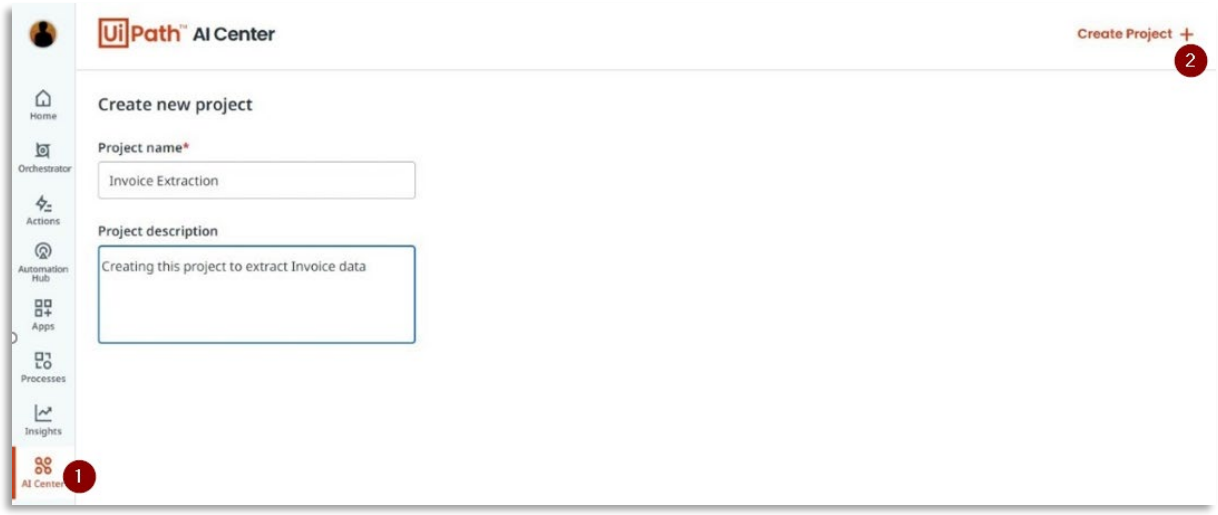


Figure 8: Creating a new project

Note AI Center is the infrastructure on which UiPath Document Understanding machine learning models run. These models can be deployed or instantiated for retraining with a few clicks.

2. Select the **Out of the box Packages** option and create a **new ML package**.

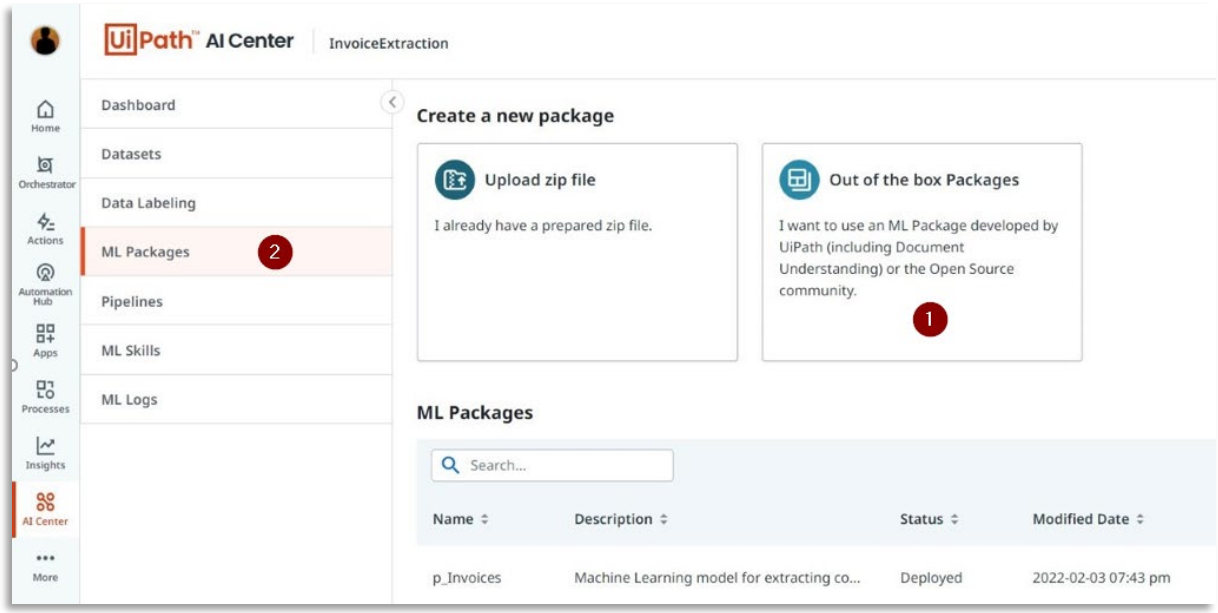


Figure 9: Creating a new ML package

© 2022 WonderBotz LLC. All Rights Reserved.

The previous action displays the document understanding package that offers multiple document options for the developer to choose based on the type of files from which the data needs to be extracted.

Document Understanding Package is a generic, retrainable model for extracting commonly occurring data points from structured or semi-structured documents, building a model from scratch. This ML Package must be trained. If deployed without training first, deployment fails with an error stating that the model is not trained.

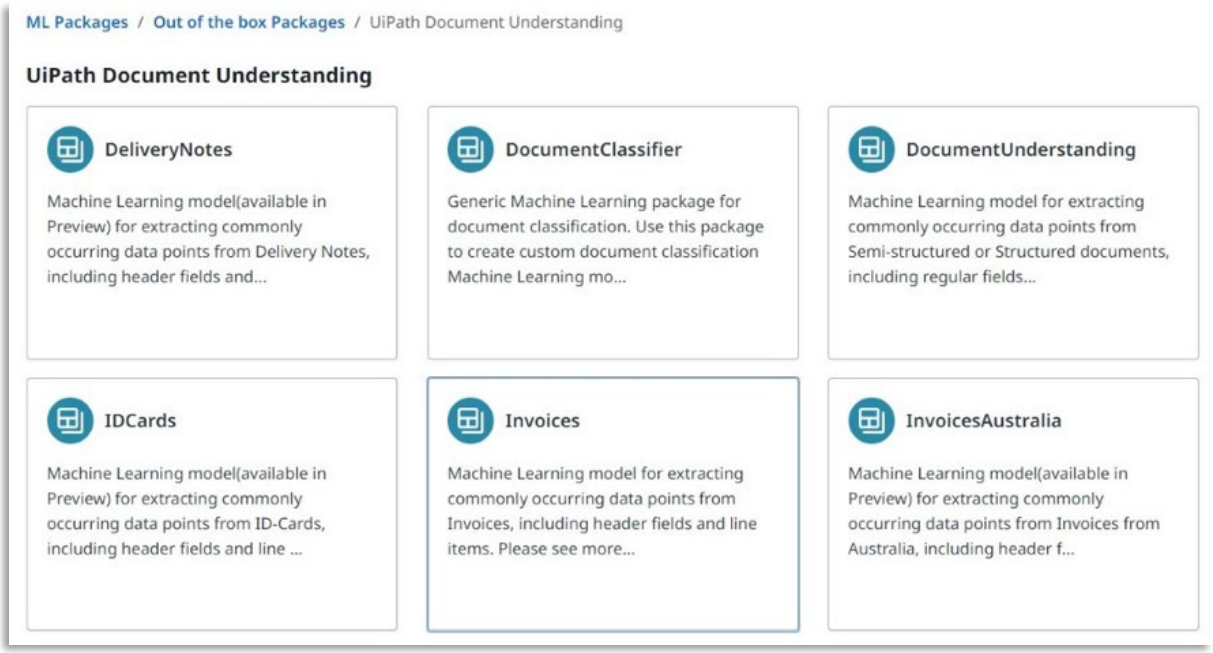


Figure 10: Document options

The package is created and is visible under the ML Packages tab of the AI Centre dashboard after the document type is specified.

© 2022 WonderBotz LLC. All Rights Reserved.

3. Create an ML skill under the ML Skills tab, and enter the `Name` and `Package` details for the new ML Skill.

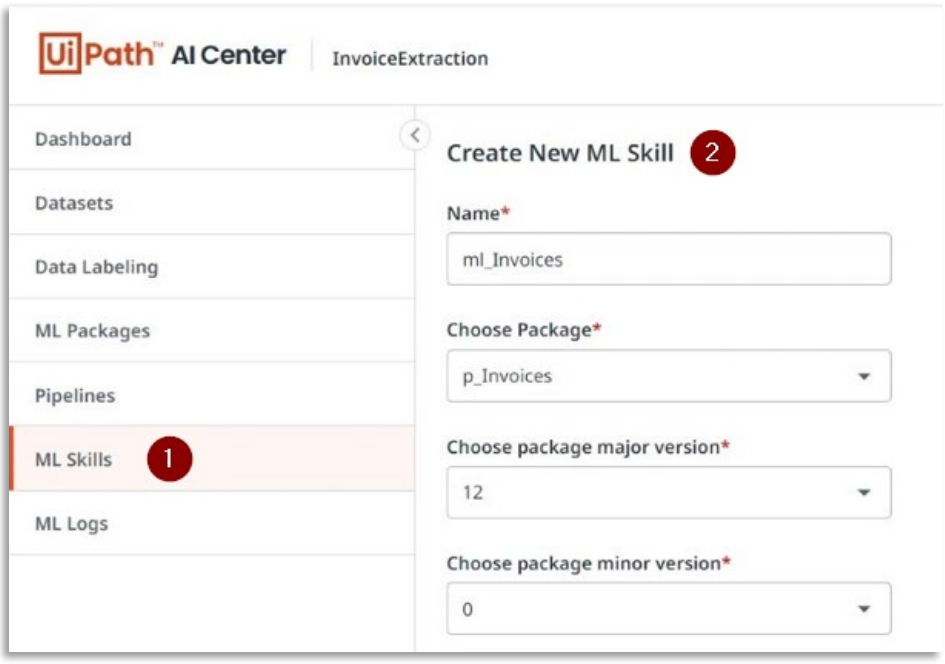


Figure 11: Creating a new ML Skill

ML skill is the actual deployment of the ML package (The ML package created in step 3). The ML skills tab of the AI Centre displays the newly created ML skills.

Uipath.DocumentUnderstading.ML.activities must be installed in the studio to work with the Document Understanding activities after creating the ML package and ML skill.

3.2 Taxonomy Manager

The Taxonomy Manager option appears in the studio toolbar after the Uipath.DocumentUnderstading.ML.activities is downloaded.

1. Click the **Taxonomy Manager** option.
2. Enter `Document Type details` (Name, Group, Category, Document Type Code, and Fields to extract.
3. Click the **Save** button; this creates a Taxonomy.txt file in the project folder.

© 2022 WonderBotz LLC. All Rights Reserved.

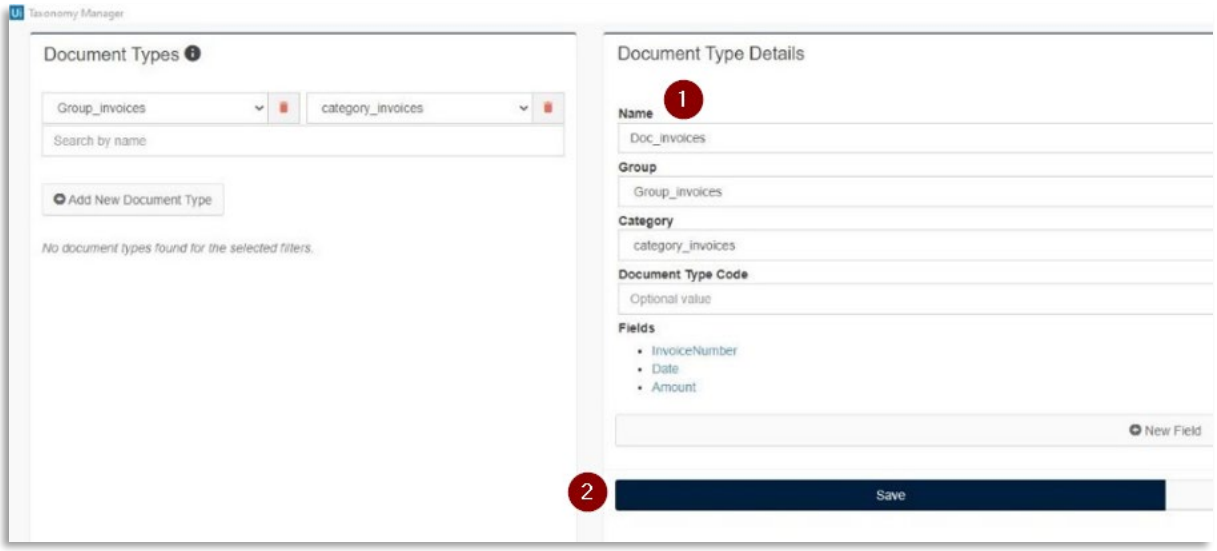


Figure 12: Creating the Taxonomy text file

The previous action makes the Load Taxonomy activity visible.



Figure 13: Load Taxonomy

Input: No Input

Configuration: Grouping of Documents that we are going to use.

Output: Taxonomy of data variable type "Taxonomy."

© 2022 WonderBotz LLC. All Rights Reserved.

3.3 Digitize Document

In this stage, the document is digitized, and the output variable is used in the following steps.

Drag the Digitize document activity from the activities panel. The document with its full path is given as input, and a Document object model is created.

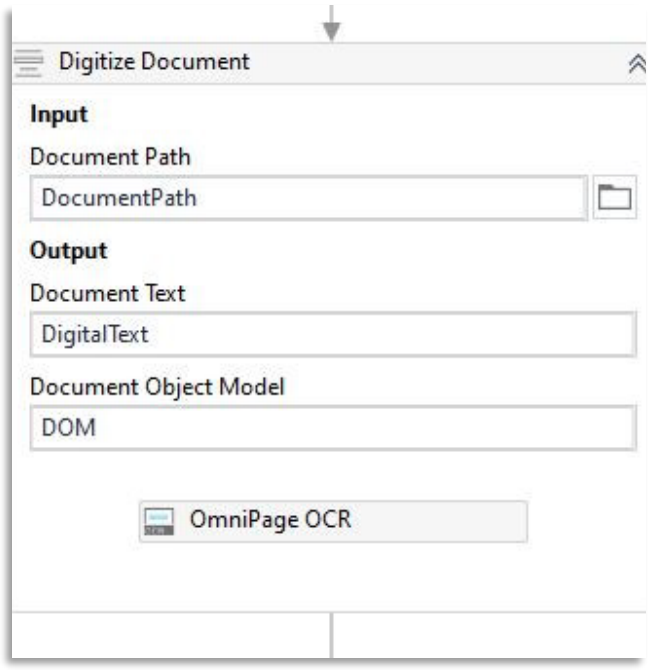


Figure 14: Digitizing document

Input: The path of the document (For multiple documents, iterative activities may be used)

Configuration: Drop the OCR Machine (Eg. OmniPage OCR).

Output: Document Text "DigitalText" of type "String" and Document Object Modelling DOM of type "DOM".

© 2022 WonderBotz LLC. All Rights Reserved.

3.4 Classification

- Classification is used when the user extracts data from distinct types of documents. Example: Invoices, Bills, Reports.
- A classifier establishes the relationship between the taxonomy and the documents.
- The Document Path, taxonomy, DOM variable, and the output from the digitization steps are input.
- Types of classifiers:
  - Intelligent keyword classifier
  - Keyword-based Classifier
  - Machine learning Classifier.
- For each Classifier, a Keyword.txt is created in the project folder, denoted as the learning file.

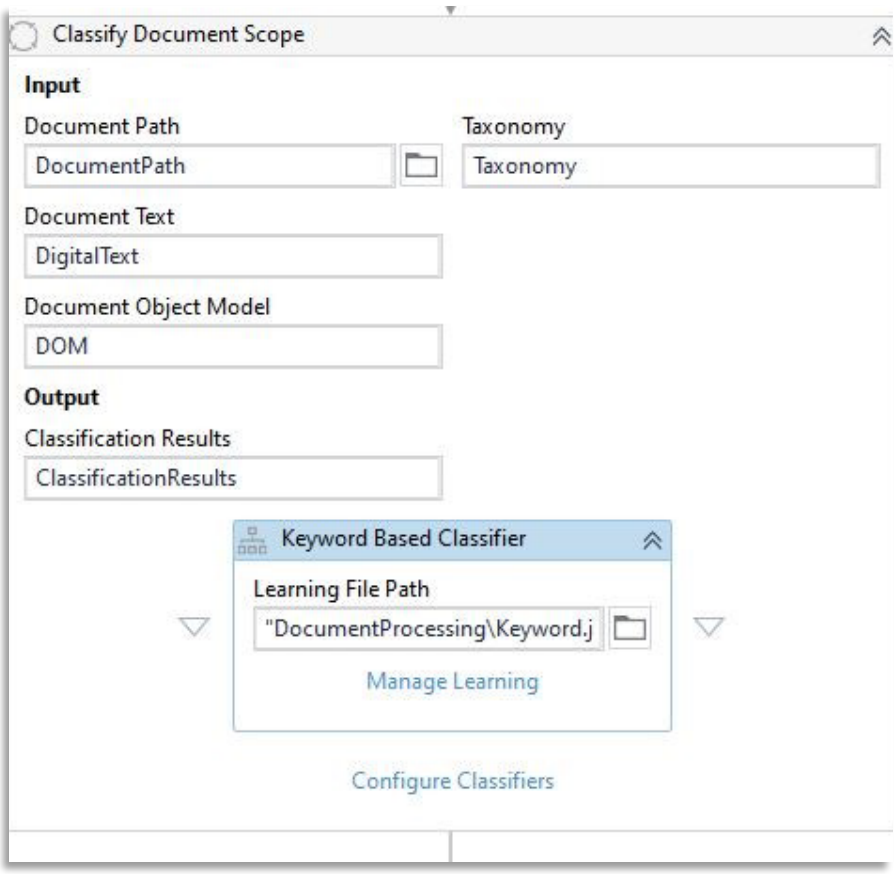


Figure 15: Classification

© 2022 WonderBotz LLC. All Rights Reserved.

- The Manage Classifier window helps configure the Document types and their respective fields.

Input: The outputs of the last two stages are the input (DocumentPath, Taxonomy, DOM, DigitalText)

Configuration: The Classifier must be added; this Classifier helps UiPath classify the document based on specific keywords. For example, "Invoice/Bill" identifies the documents for specific invoices.

Classifier Types:

- Keyword based Classifier
- Intelligent Keyword Classifier
- Machine Learning Classifier

Output: Classification results of type Array of 'ClassificationResult.'

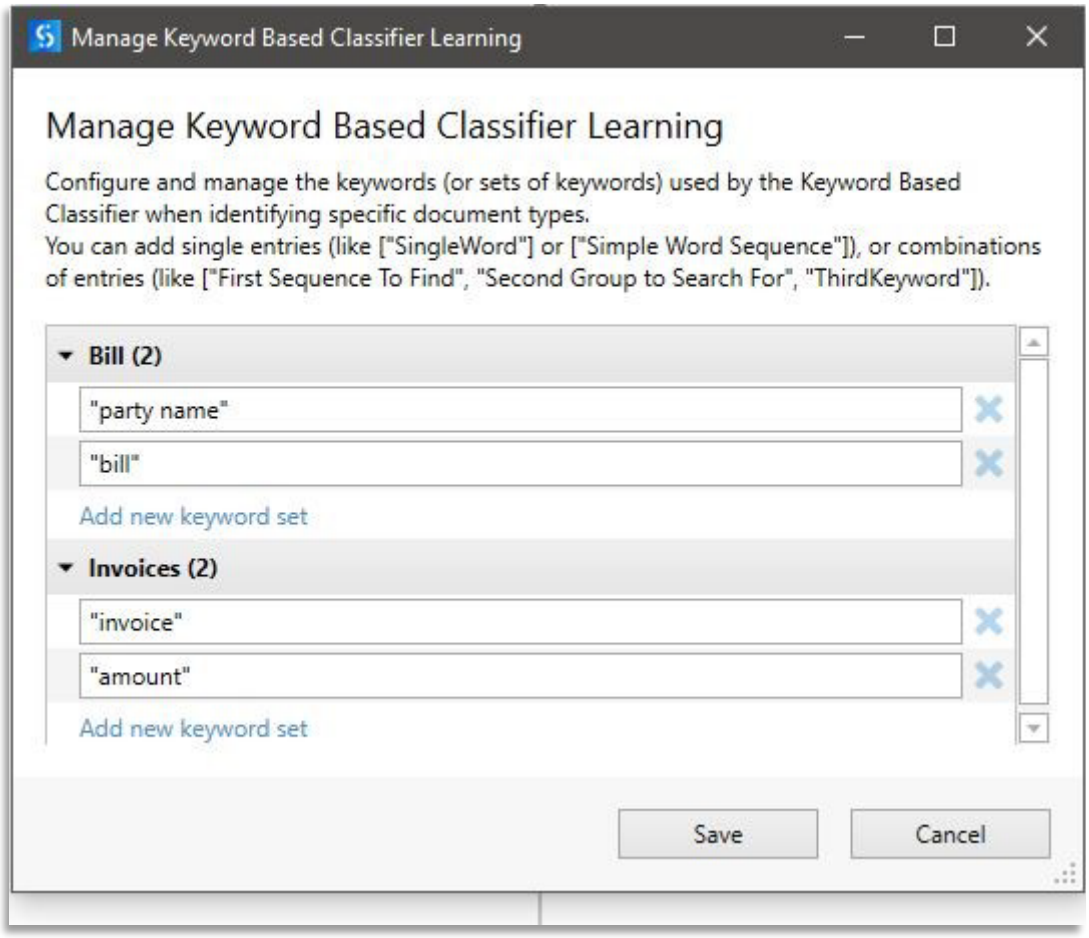
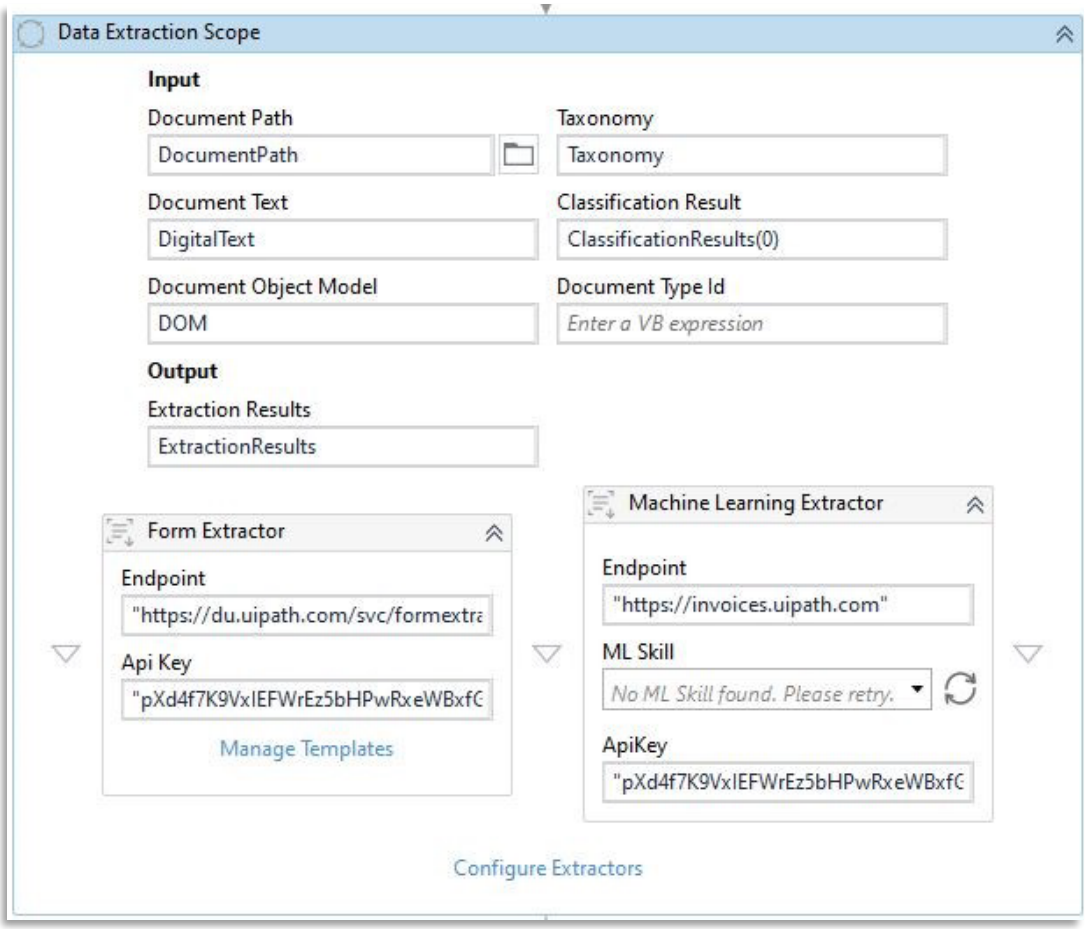


Figure 16: Manage Keyword-Based Classifier Learning

© 2022 WonderBotz LLC. All Rights Reserved.

3.5 Data Extraction

- The Data Extraction Scope activity extracts the data from the documents using various extractors available.
- Types of Extractors:
  - Form based extractor
  - RegEx based extractor
  - Intelligent form extractor
  - Machine learning extractor
- The inputs for the Data extraction activity are:
  - Document Path
  - Digitized text, DOM variable
  - Taxonomy variable
  - Classification results



© 2022 WonderBotz LLC. All Rights Reserved.

Figure 17: Data Extraction Scope

- Multiple extractors can also be configured.
- Each extractor holds parameters respective to the document type and the fields to be extracted from them, and they are configured in the Configure Extractor window.

Input: The outputs of the last three stages are the input (DocumentPath, Taxonomy, DOM, DigitalText, ClassificationResults).

Configuration: Import the extractor from the activities panel.

Extractor Types:

- Intelligent form extractor
- Machine learning extractor
- RegEx based extractor

Output: Extraction Results of data type "ExtractionResults."

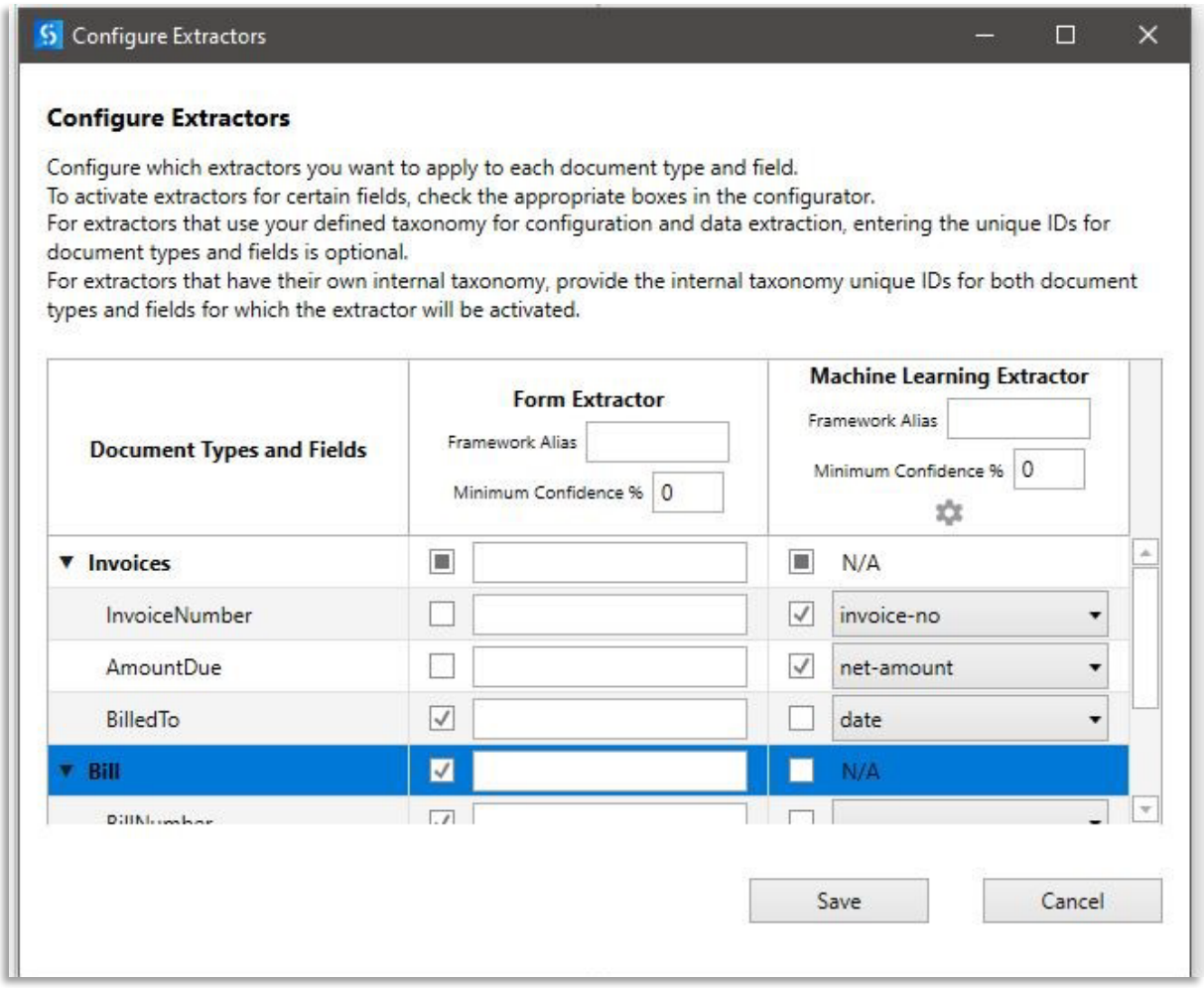
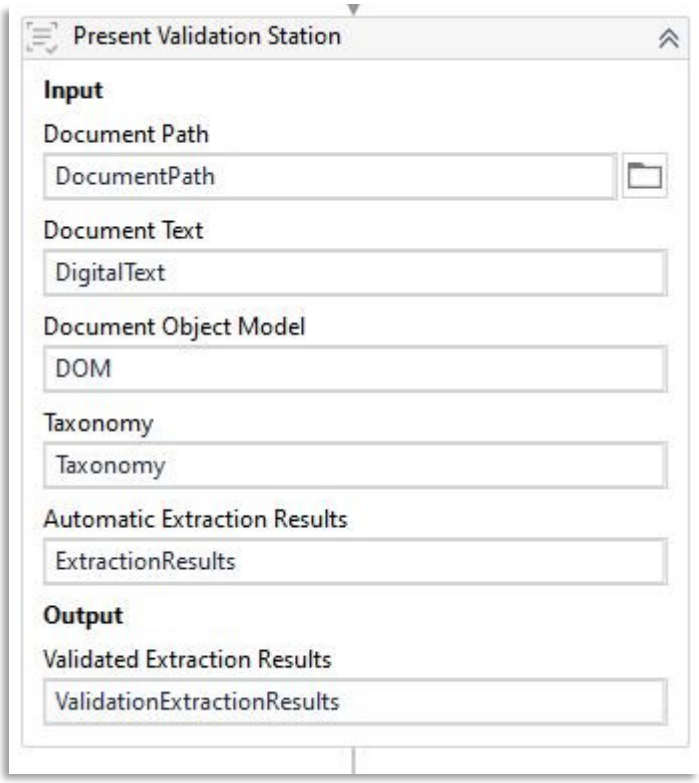


Figure 18: Configure Extractors

© 2022 WonderBotz LLC. All Rights Reserved.

3.6 Present Validation Status

The extracted results are verified in this stage, and the user validates if the extracted results are as per the target fields.



Input: The outputs of the last four stages are the input (DocumentPath, Taxonomy, DOM, DigitalText, ClassificationResults, ExtractionResult)

Configuration: Provide inputs.

Output: Validation Extraction Results of data type "ExtractionResults."

3.7 Export Extraction Data

The final stage saves the extracted results and is of the type dataset. A dataset is a collection of Tables, and each table contains the data extracted from a specific document.

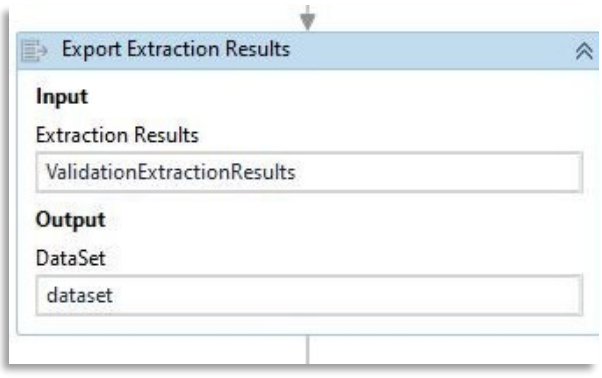


Figure 19: Export Extraction Results

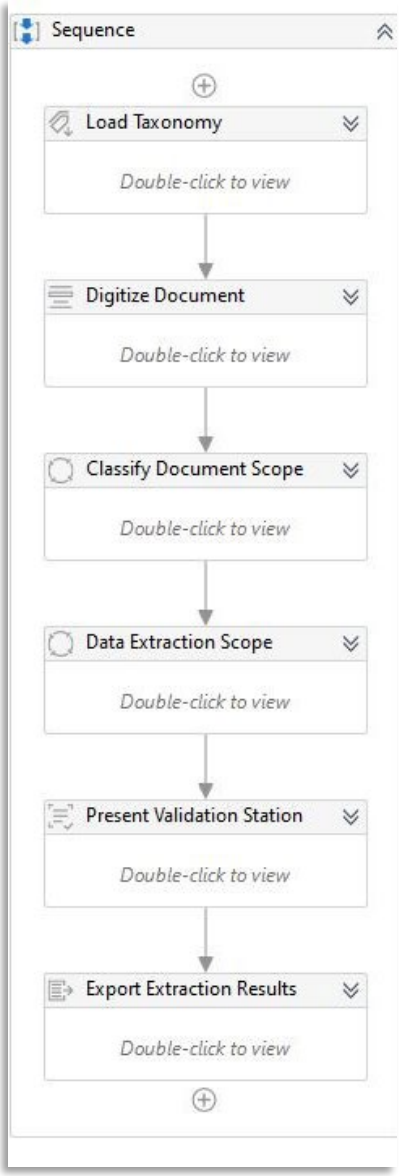
© 2022 WonderBotz LLC. All Rights Reserved.

Input: The outputs of the last 5 stages would be the input (DocumentPath, Taxonomy, DOM, DigitalText, ClassificationResults, ExtractionResult, ValidationExtractionResult)

Configuration: Provide inputs.

Output: A data set of type "DataSet" creates a data structure for extracted results.

The workflow looks as follows after all the stages are complete.



.

Figure 20: Workflow after the completion of the Document Understanding stages

© 2022 WonderBotz LLC. All Rights Reserved.

APPENDICES

APPENDIX A: References

Overview. (n.d.). Retrieved from Basic Commands:
https://www1.cmc.edu/pages/faculty/aaksoy/latex/latextwo.html

Suraj, G. S. (2020, October 23). *UiPath Document Understanding Framework*. Retrieved from SOAIS:
https://www.soais.com/uipath-document-understanding-framework/

UiPath Community Forum. (2021, October). *Introduction*. Retrieved from UiPath:
https://docs.uipath.com/document-understanding/docs/introduction

APPENDIX B: List of Figures

Figure 1: Document Understanding 5
Figure 2: Document Understanding stages 5
Figure 3: UiPath.IntelligentOCR.Activities package installation in UiPath 6
Figure 4: OmniPage.Activities package installation 6
Figure 5: DocumentUnderstanding.MLActivities installation 7
Figure 6: API Key 7
Figure 7: Document Understanding Framework Components 10
Figure 8: Creating a new project 11
Figure 9: Creating a new ML package 11
Figure 10: Document options 12
Figure 11: Creating a new ML Skill 13
Figure 12: Creating the Taxonomy text file 14
Figure 13: Load Taxonomy 14
Figure 14: Digitizing document 15
Figure 15: Classification 16
Figure 16: Manage Keyword-Based Classifier Learning 17
Figure 17: Data Extraction Scope 19
Figure 18: Configure Extractors 19
Figure 19: Export Extraction Results 20
Figure 20: Workflow after the completion of the Document Understanding stages 21

© 2022 WonderBotz LLC. All Rights Reserved.

Thank you!

Copyright ©2022 WonderBotz LLC. All rights reserved. No parts of this document may be reproduced, transmitted, or stored electronically without WonderBotz's prior written permission.

This document contains WonderBotz's confidential or proprietary information. By accepting this document, you agree that: (A) (1) if a pre-existing contract containing the disclosure and use restrictions exists between your company and WonderBotz, you and your company will use this information subject to the terms of the pre-existing contract; or (2) if no such pre-existing contract exists, you and your company agree to protect this information and not reproduce or disclose the information in any way; and (B) WonderBotz makes no warranties, express or implied, in this document, and WonderBotz shall not be liable for damages of any kind arising out of the use of this document.