Different Format Invoice Data Extraction

This is a complete solution workflow in which you can input your documents, create a taxonomy of different formats and extract data using document understanding.

Packages to be Installed:

1)Intelligent OCR activities

2)Omni Page OCR

3)Machine Learning Extractor

Processing and extracting data from different types of documents using Document Understanding

- This workflow explains how you can use document understanding and process multiple documents of different formats and extract data.
- The UiPath Document Understanding Framework is designed to help users combine different approaches to extract information from multiple documents, not necessarily with the same structure

Steps to be followed:

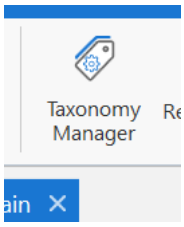
1. Get Files
2. Load Taxonomy
3. Digitize Document
4. Classify Document Scope
5. Data Extraction Scope
6. Present Validation Station
7. Export Extraction Results

Load Taxonomy:

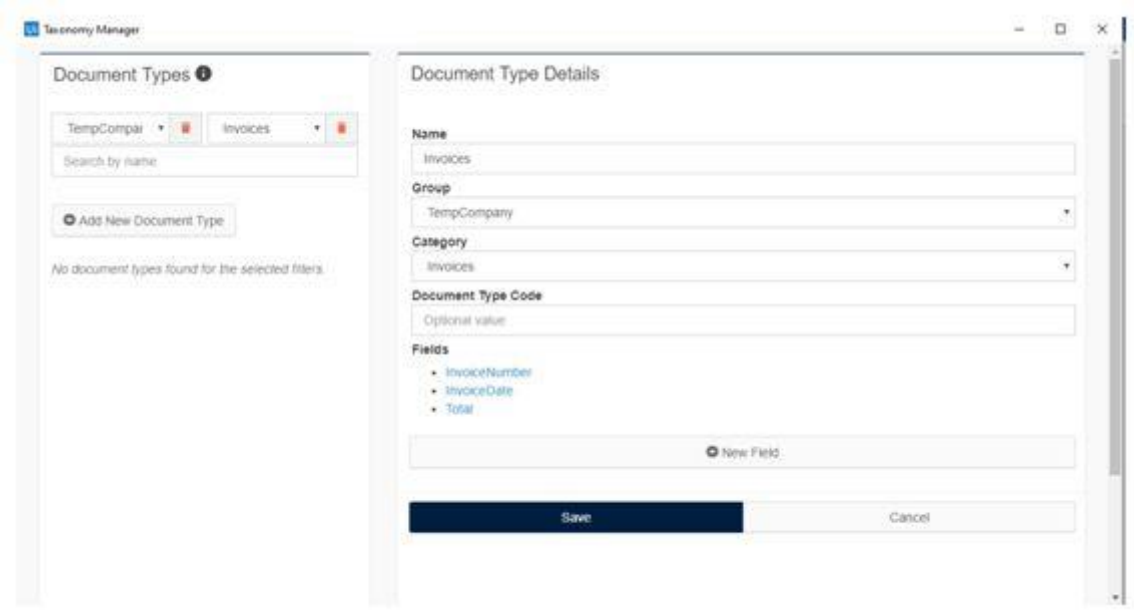
In this pre-processing step, you can add multiple document types and the fields you are interested in extracting. For example, you can work with Invoices, wanting to extract the Invoice Number, Invoice Date and Due Date.

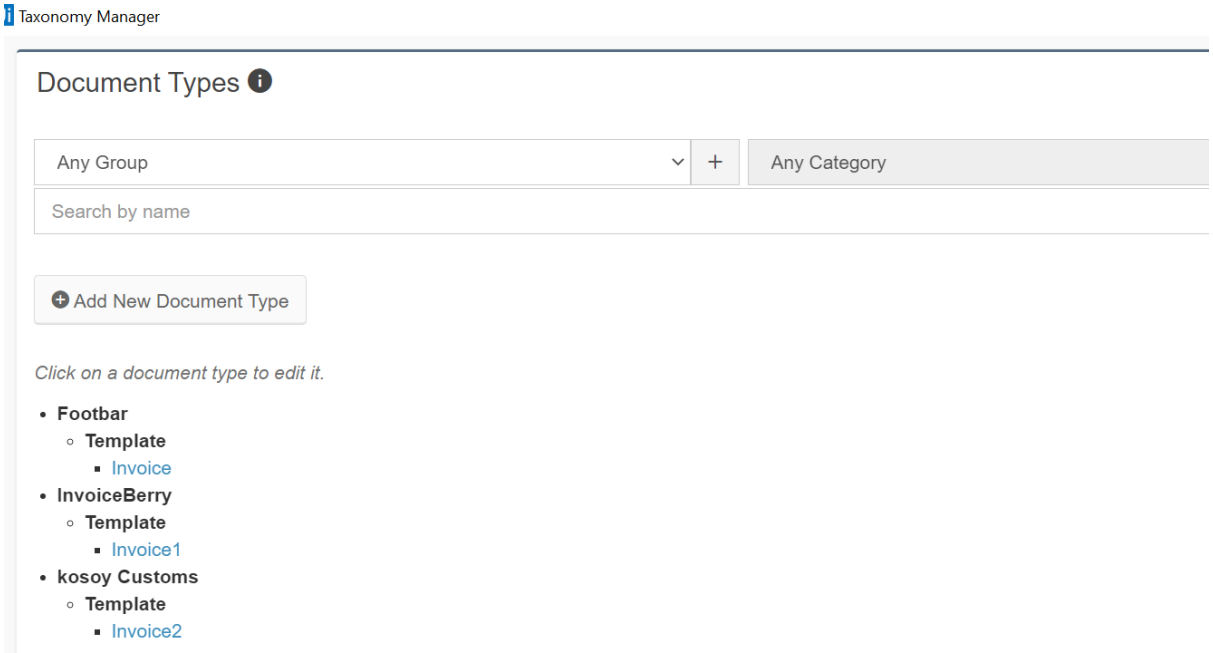
Step 2:

Click on Taxonomy Manager.



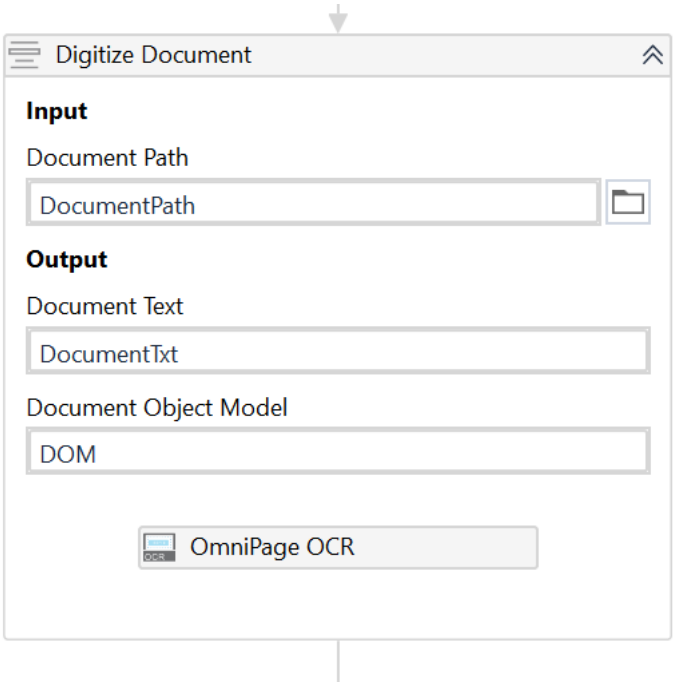
Step 3:





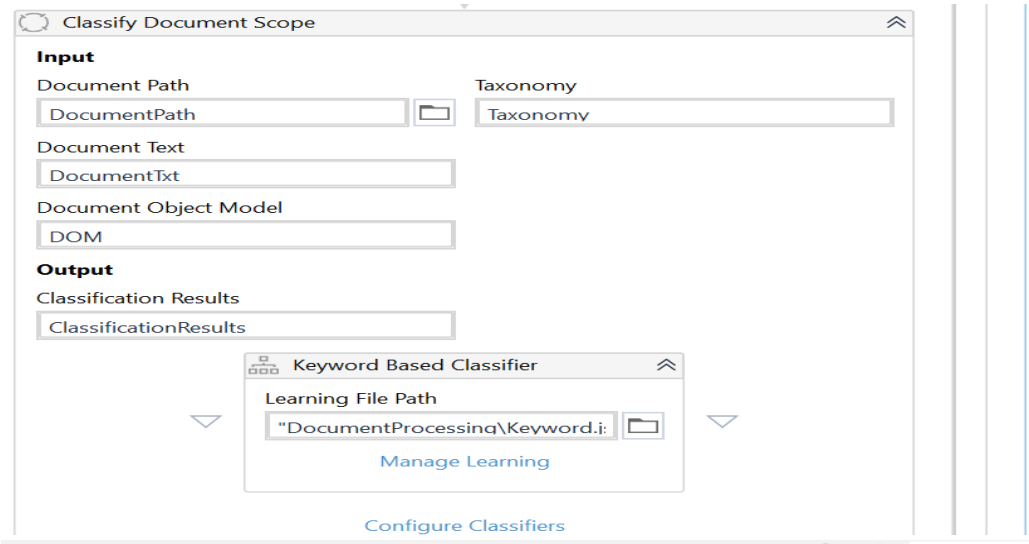
Digitize Document:

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



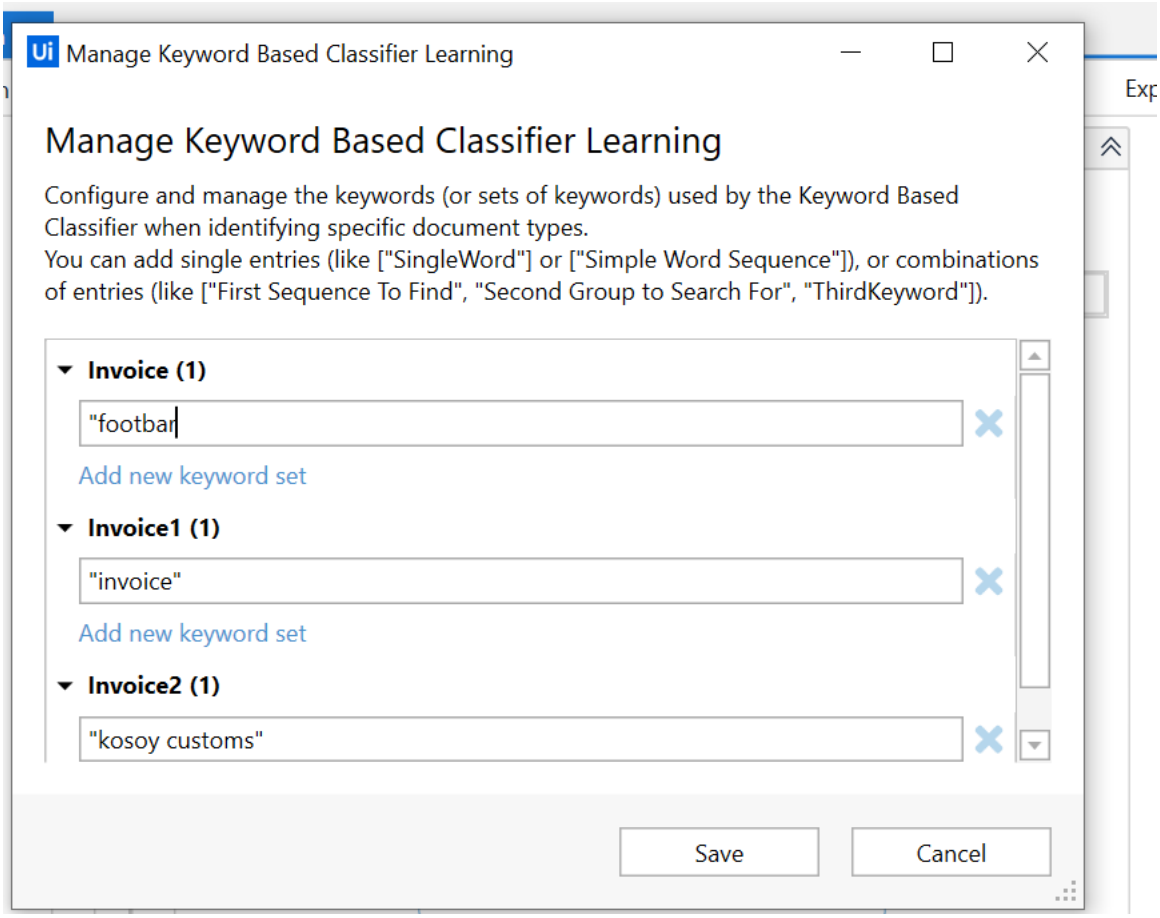
Classification:

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



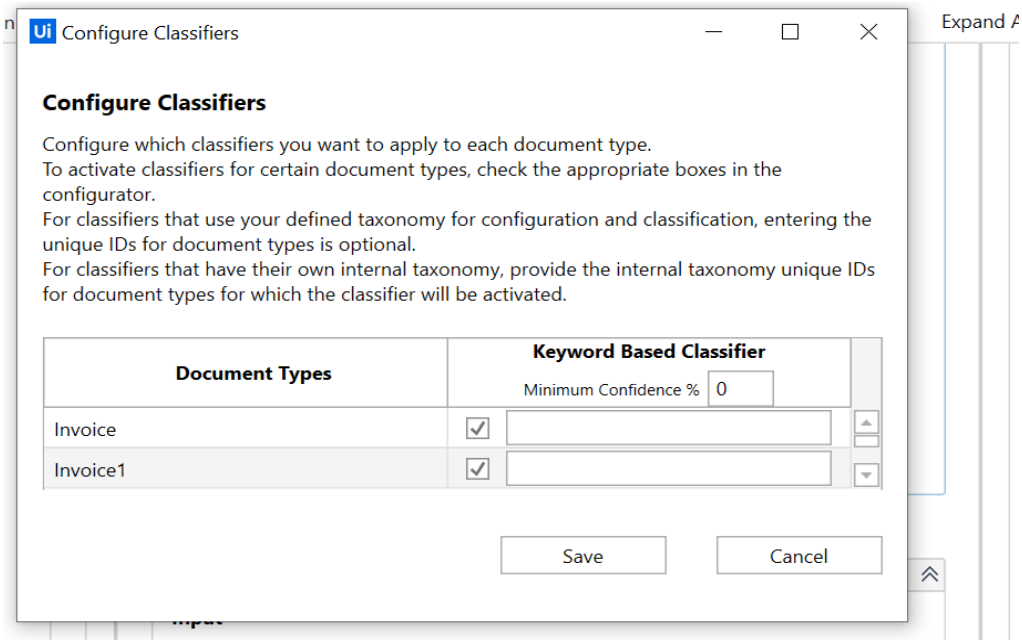
Step 2:

Click on ***Manage Learning.***



Step 3:

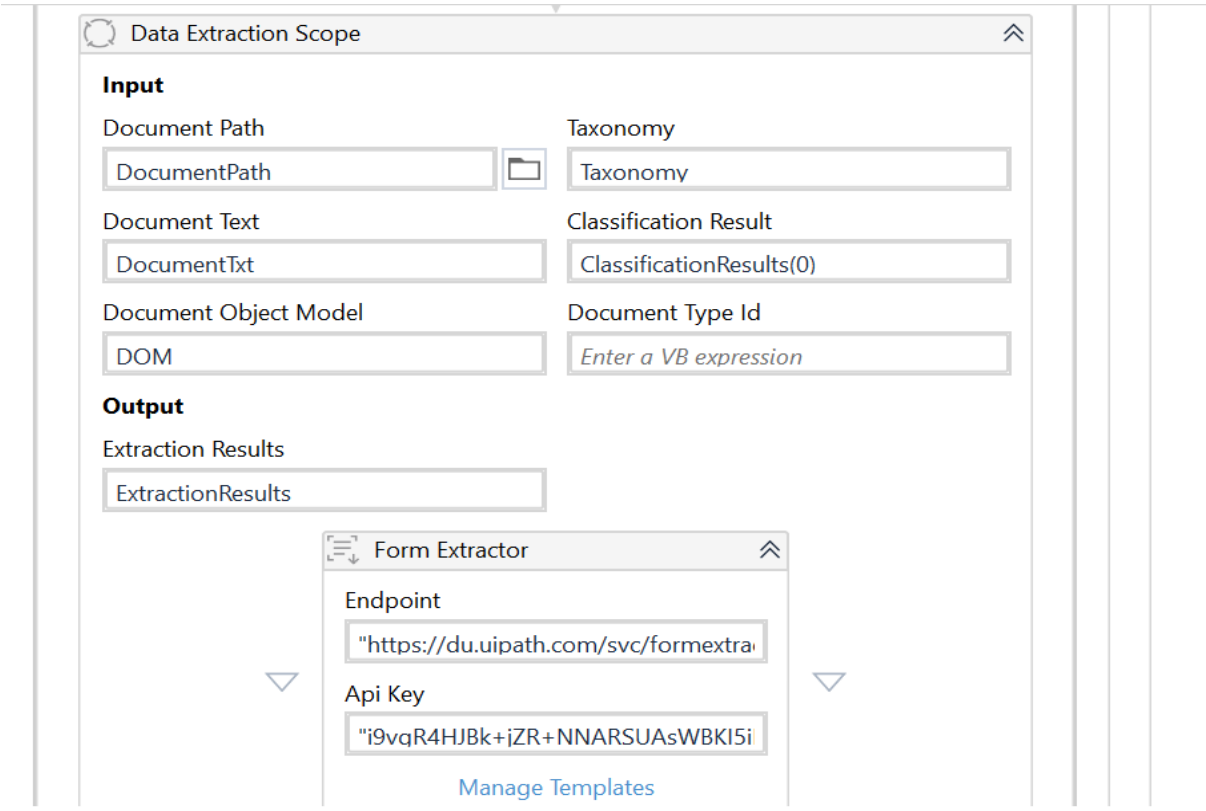
Click on Configure Classifiers



Click on check boxs and save.

Data Extraction Scope:

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



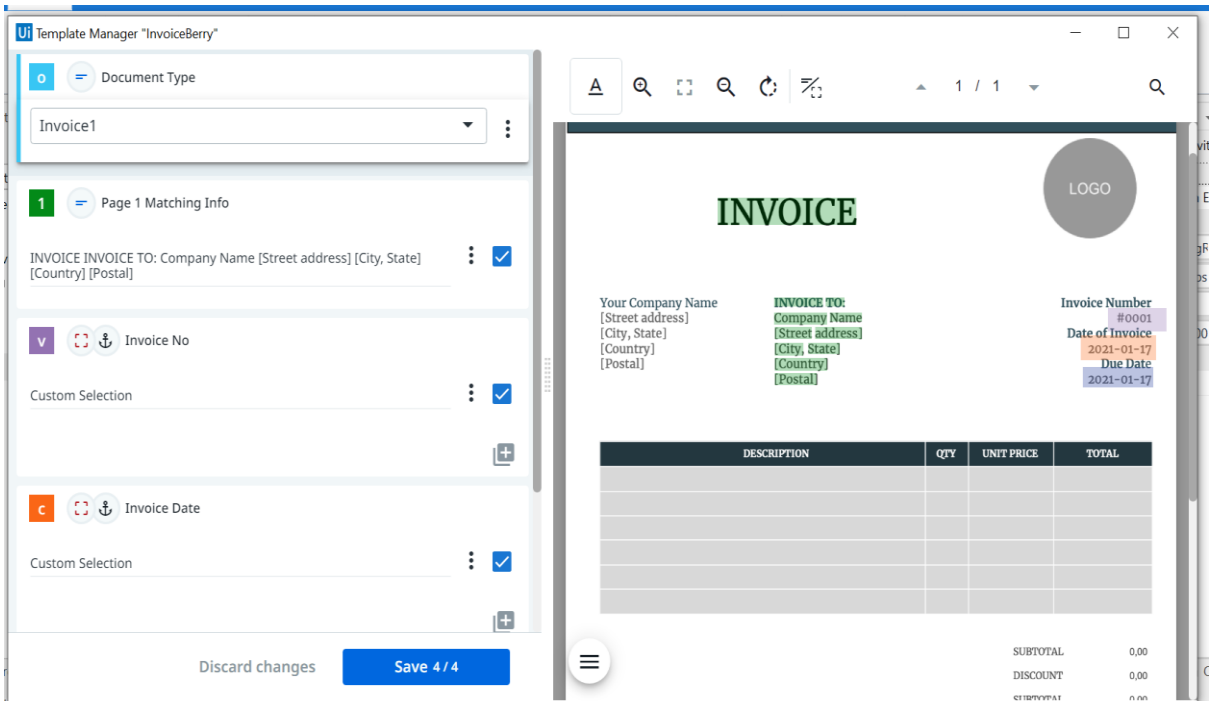
There are different extractors,

- Form Extractor
- Machine Learning Extractor
- Regex Based Extractor

To get API Key go to Orchestrator click Admin→ Licences→Robots and Services→Generate API->Copy API

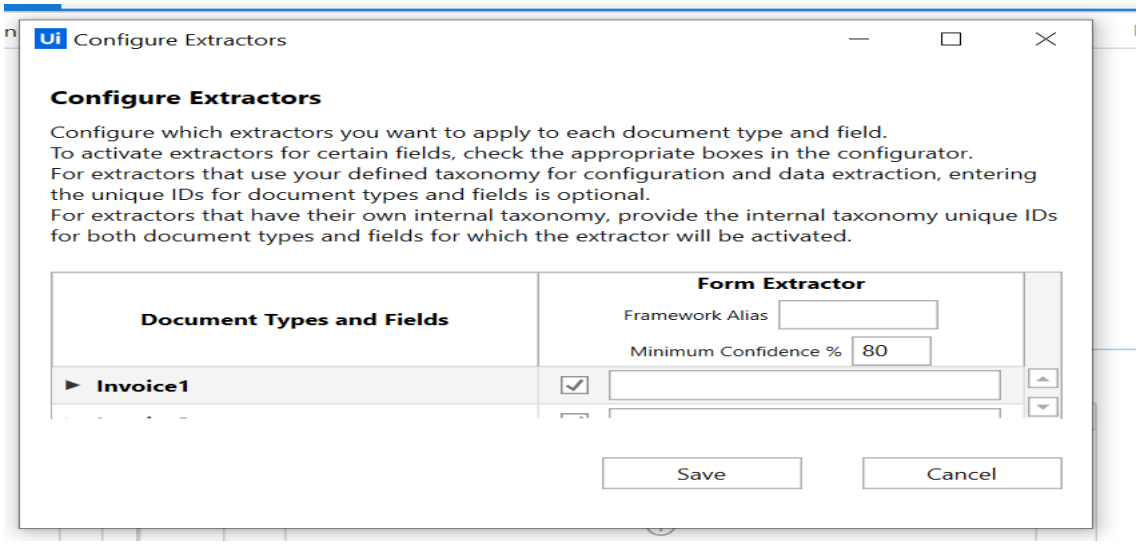
Step 2:

Click on Manage Template



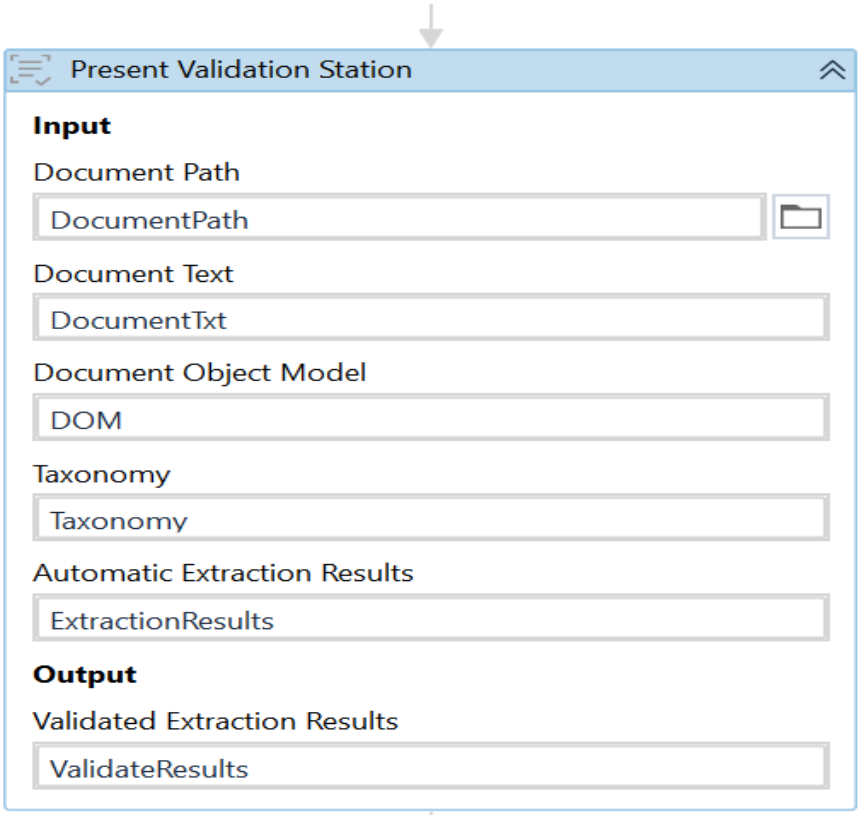
Step 3:

Click on Configure Classifiers



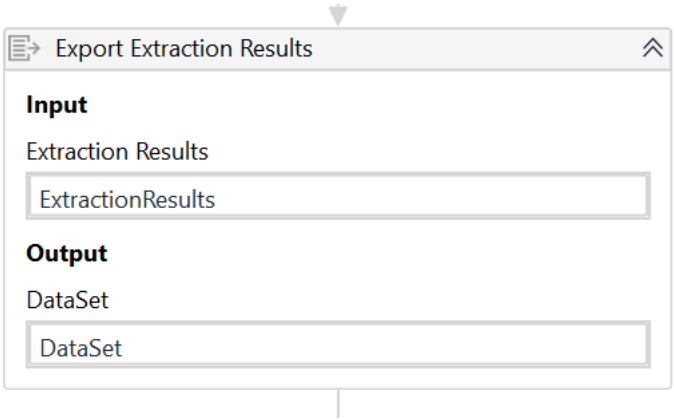
Present Validation Station:

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



Export Extraction Result:

Once you have your validated information, you can use it as it is, or save it in a DataTable format that can be converted very easy into an Excel file.



Example:

ExtractUnstructuredDocuments

Input file path

FilePath = Directory.GetFiles('

Assign

Counter = 2

For Each

ForEach Docume in FilePath

Body

Body

Load Taxonomy

Output

Taxonomy

Taxonomy

Digitize Document

Input

Document Path

DocumentPath

Output

Document Text

DocumentTxt

Document Object Model

DOM

Document Object Model

DOM

OmniPage OCR

Classify Document Scope

Input

Document Path

DocumentPath

Taxonomy

Taxonomy

Document Text

DocumentTxt

Document Object Model

DOM

Output

Output

Classification Results

ClassificationResults

Keyword Based Classifier

Learning File Path

"DocumentProcessing\Keyword.j

Manage Learning

Configure Classifiers

Data Extraction Scope

Input

Document Path

DocumentPath

Taxonomy

Taxonomy

Document Text

DocumentTxt

Classification Result

ClassificationResults(0)

Document Text

DocumentTxt

Classification Result

ClassificationResults(0)

Document Object Model

DOM

Document Type Id

Enter a VB expression

Output

Extraction Results

ExtractionResults

Form Extractor

Endpoint

"https://du.uipath.com/svc/formextra

Api Key

"i9vqR4HJBk+jZR+NNARSUAsWBKI5i

Manage Templates

Configure Extractors

Present Validation Station

Input

Document Path

DocumentPath

Document Text

DocumentTxt

Document Object Model

DOM

Taxonomy

Taxonomy

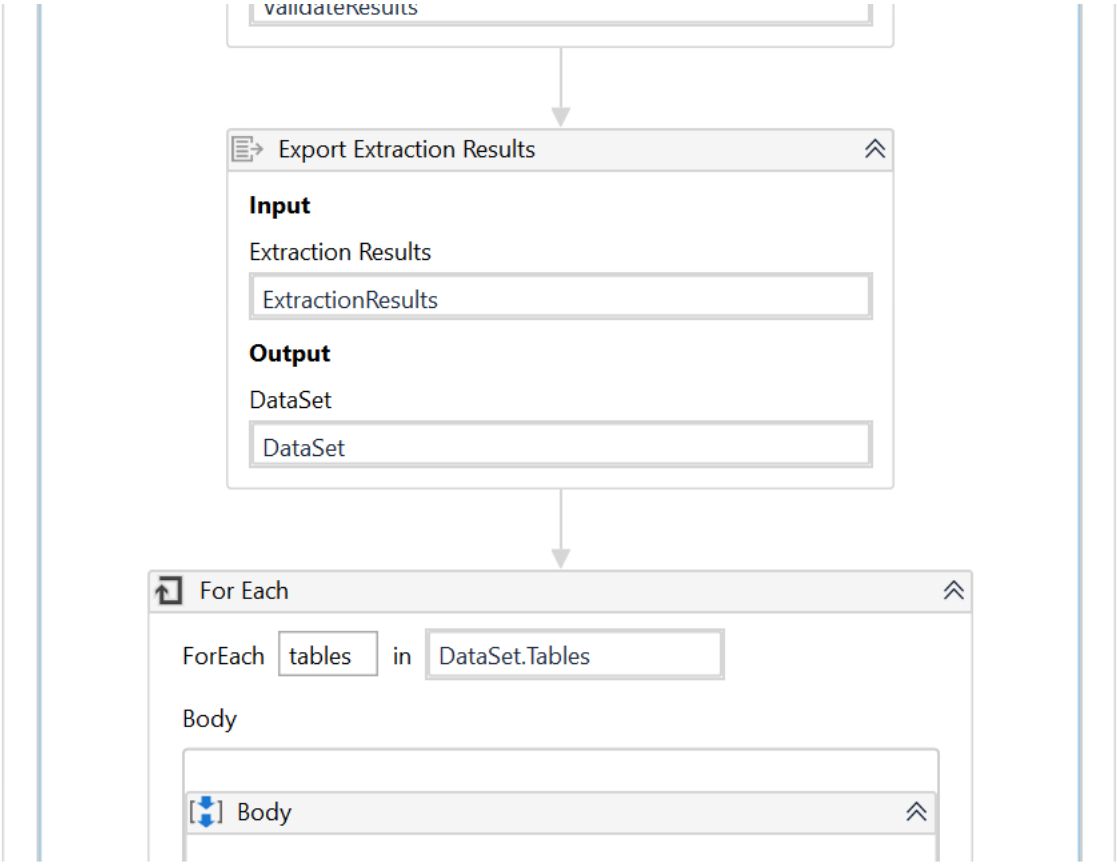
Automatic Extraction Results

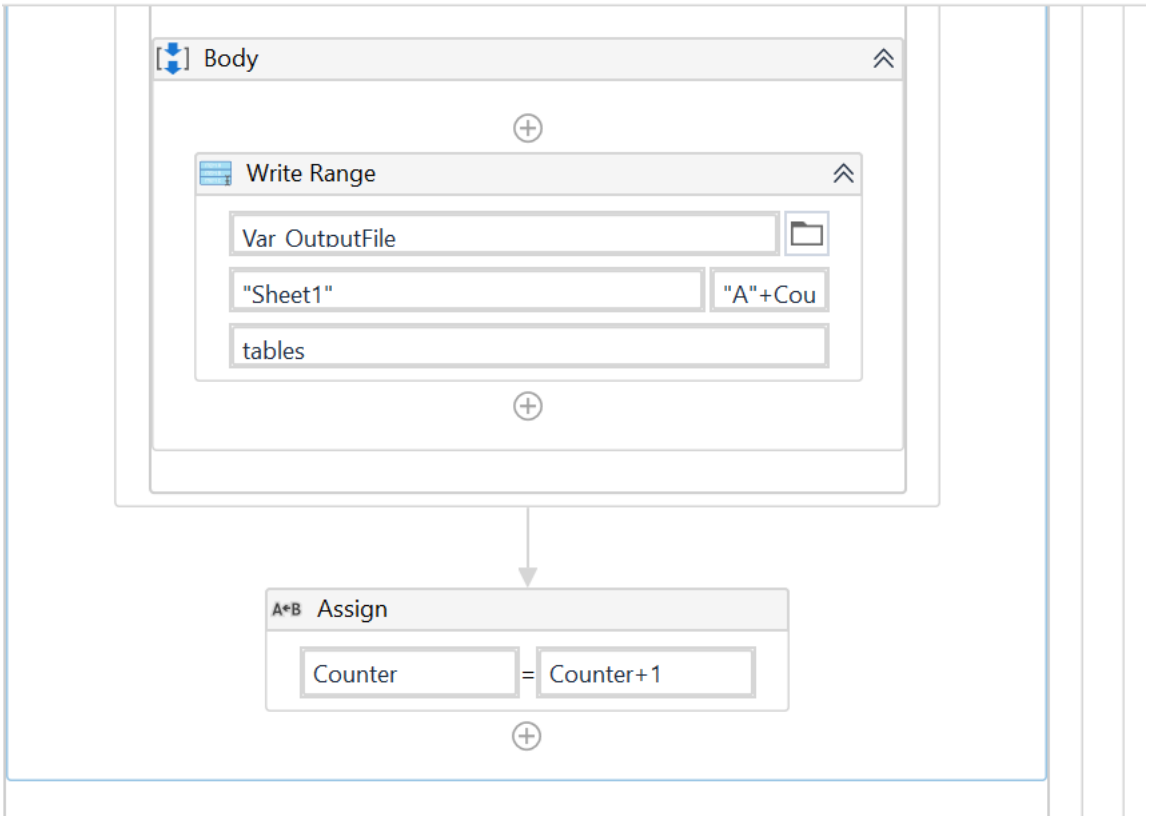
ExtractionResults

Output

Validated Extraction Results

ValidateResults


ValidateResults
Export Extraction Results
Input
Extraction Results
ExtractionResults
Output
DataSet
DataSet
For Each
ForEach
tables
in
DataSet.Tables
Body
Body


Body
Write Range
Var_OutputFile
"Sheet1"
"A"+Cou
tables
Assign
Counter
=
Counter+1