

Document Understanding:

1. Apply Regex Extractor on Financial Document using Document Understanding (DU) Framework
2. How to Extract table from pdf using Document Understanding Framework
3. What is DOM?

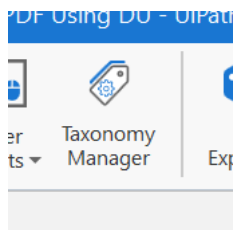
Apply Regex Extractor on Financial Document using Document Understanding (DU) Framework:

Input:

Extract text from forms and certificate using Regex Extractor

Step 1:

Click on load Taxonomy Manager

A screenshot of the 'Taxonomy Manager' application window. The window is divided into three main panels. The left panel, 'Document Types', shows a tree view of document types: Forms (Tax, PDF), Certificate (image, Certificatefiling), Syruis (pdf, invoices), W-9 (Tax, form), and W-4 (pdf, Form1). The middle panel, 'Document Type Details', shows details for 'Forms.Tax.PDF', including its Name (PDF), Group (Forms), Category (Tax), and Document Type Code (Optional value). The right panel, 'Edit Field', shows details for the 'FormType' field, including its Name (FormType), Is Multi-Value (unchecked), Requires Reference (checked), and Type (Text). The window has a title bar with standard OS controls and a bottom bar with 'Save' and 'Cancel' buttons.

Taxonomy Manager

Document Types

Any Group + Any Category

Search by name

+ Add New Document Type

Click on a document type to edit it.

- Forms
 - Tax
 - PDF
- Certificate
 - image
 - Certificatefiling**
- Syruis
 - pdf

Document Type Details

Document Type ID: Certificate.Image.Certificatefiling

Name
Certificatefiling

Group
Certificate

Category
image

Document Type Code
Optional value

Fields

- File Number
- Dated
- Effective Date

Edit Field

Field ID: Certificate.Image.Certificatefiling.FileNumber

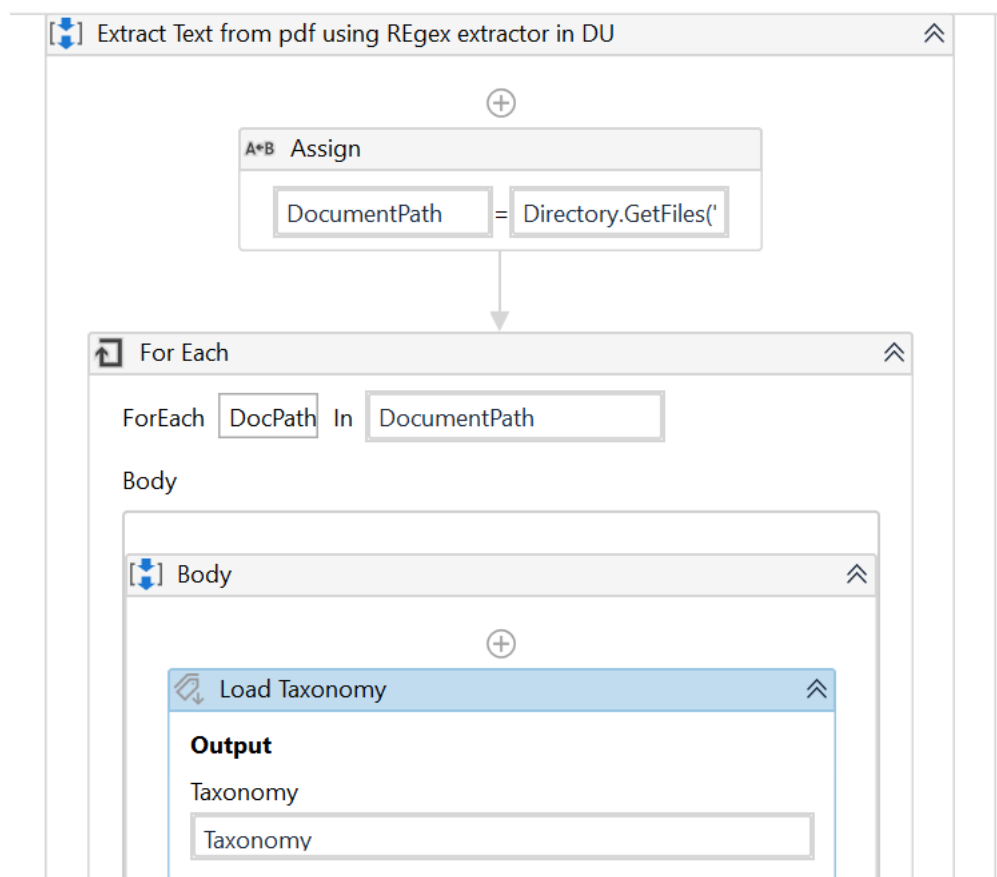
Name
File Number

☐ Is Multi-Value

☒ Requires Reference

Type
Number

Save **Cancel**



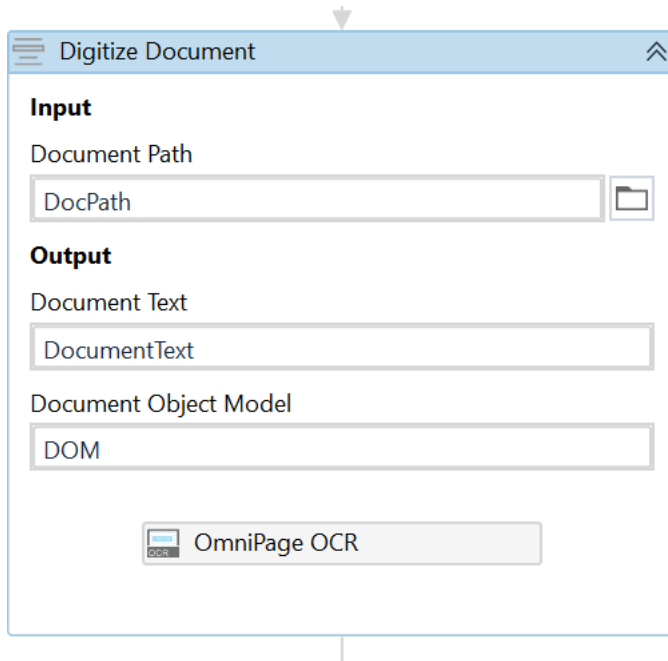
Taxonomy.json created.

Step 2:

Digitize Document:

All documents that are to be processed (native and scanned) must pass through this step-in order for the robot to understand the kind of document it's working with and what data is relevant.

The OCR engine will be used only if the incoming documents require OCR processing, and the decision gets taken on a page-by-page basis.



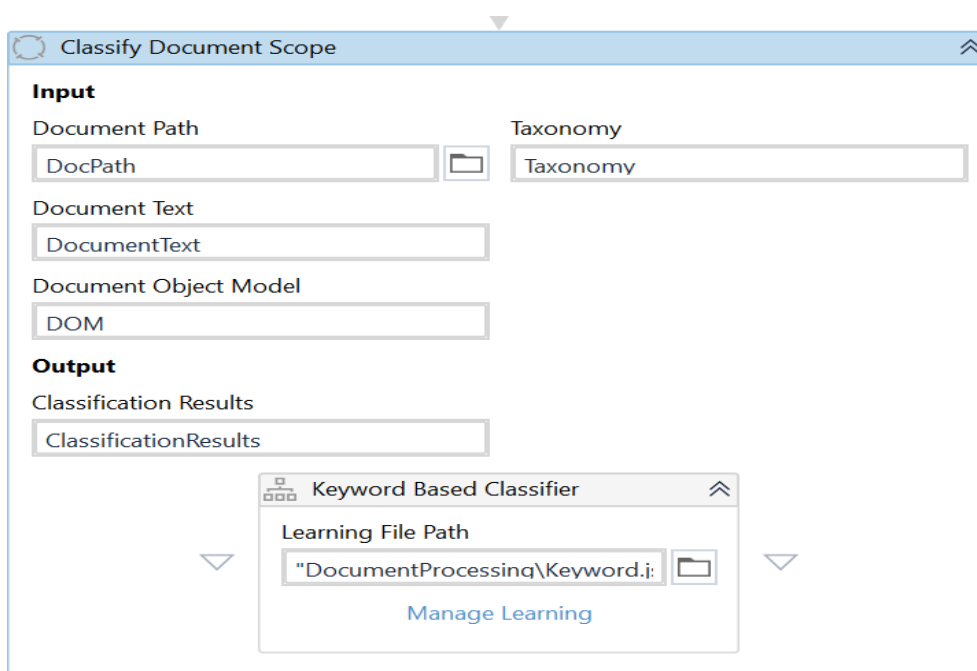
The 'Digitize Document' dialog box is shown with a blue header bar containing a menu icon and the title 'Digitize Document'. It has two main sections: 'Input' and 'Output'. In the 'Input' section, there is a 'Document Path' label above a text field containing 'DocPath' and a folder icon. In the 'Output' section, there are two labels: 'Document Text' above a text field containing 'DocumentText', and 'Document Object Model' above a text field containing 'DOM'. At the bottom of the dialog is a button labeled 'OmniPage OCR' with a small icon to its left.

Step 3:

Classification is done through the Classify Document Scope and it's performed by classifiers.

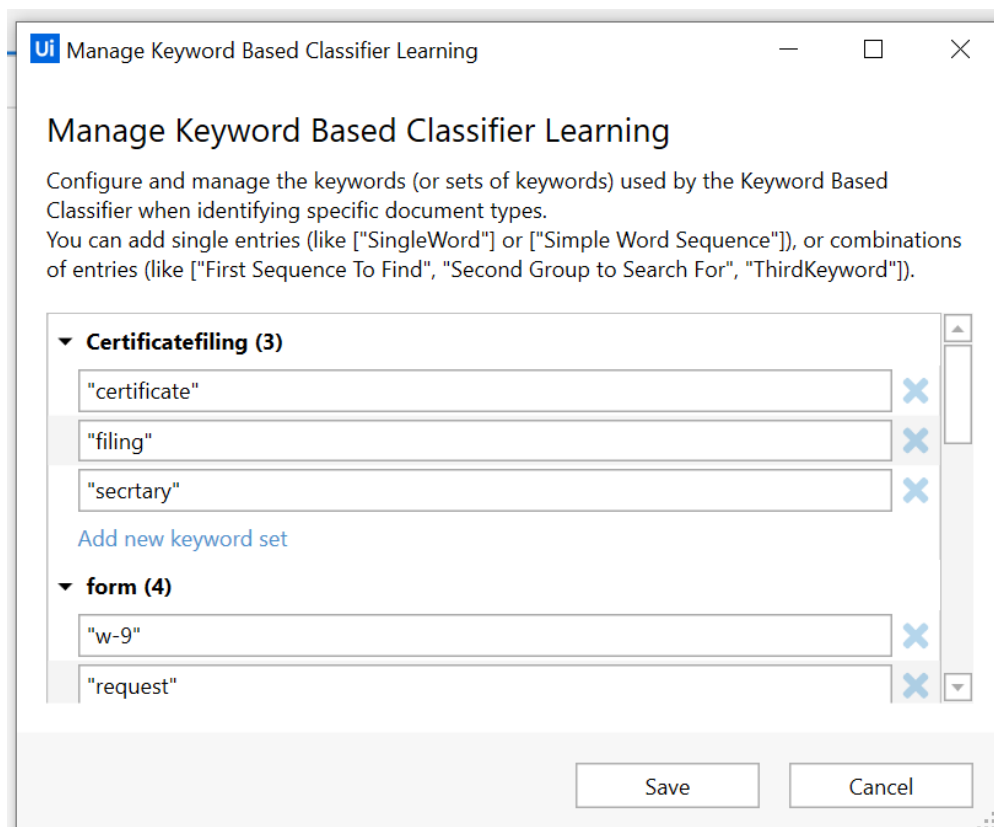
Basically, the document text and object model resulted in the digitization steps are sent to the Classifiers, which report what types they recognize within the incoming file.

The Document Classification Scope Wizard will open at selecting the **Configure Classifiers** option and it allows users to customize which classifier will be used for each individual type of document.



The 'Classify Document Scope' dialog box is shown with a blue header bar containing a gear icon and the title 'Classify Document Scope'. It has two main sections: 'Input' and 'Output'. In the 'Input' section, there are four labels: 'Document Path' above a text field with 'DocPath' and a folder icon; 'Taxonomy' above a text field with 'Taxonomy'; 'Document Text' above a text field with 'DocumentText'; and 'Document Object Model' above a text field with 'DOM'. In the 'Output' section, there is one label: 'Classification Results' above a text field with 'ClassificationResults'. A sub-dialog box titled 'Keyword Based Classifier' is open in front of the main dialog. It has a 'Learning File Path' label above a text field containing '"DocumentProcessing\Keyword.i:' and a folder icon. Below this text field is a button labeled 'Manage Learning'.

Click on manage learning

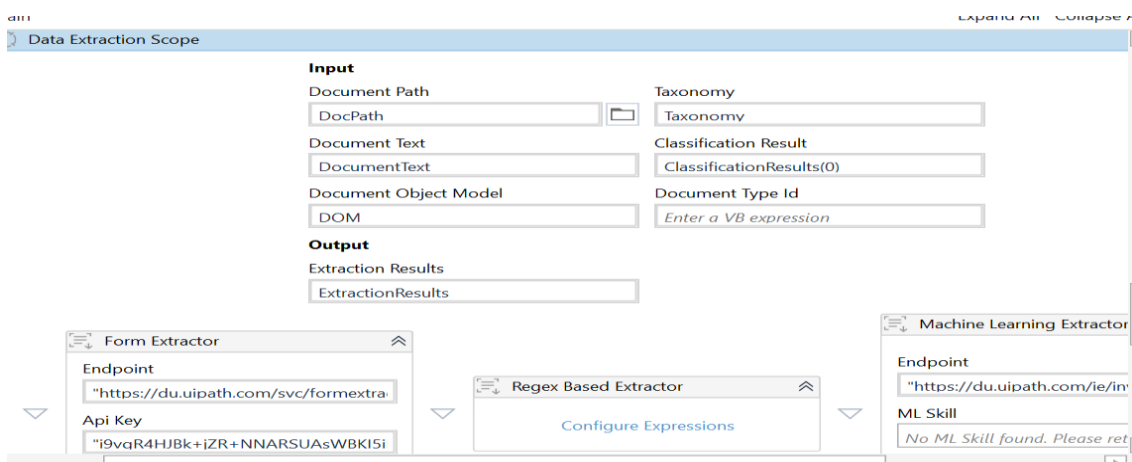


Step 4:

The input consists of the outputs of the previous Document Understanding steps, along with the document path. Based on the structure of the document, different extractors come to play.

The Data Extraction Scope Wizard will open at selecting the Configure Extractors option and it allows users to customize which extractors will be used for each individual field.

It allows users to mix and match extractors as well as use extractors in parallel based on which extractor has the highest confidence level.



Click on Configure Expressions in regex based extractor.

Configure Regular Expressions

Configure regular expressions and their options for each field of the taxonomy you wish to extract. You can either write your regular expression by hand or use our editor. In both cases you should for the information you wish to be extracted. To do this in the editor you should check the Capture checkbox next to each part of the regex that needs to be extracted.

Document Types and Fields	Expression	
▼ PDF		
FormType	FORM\s10-()	Igno
▼ Certificatefiling		
File Number	File\ Number:\s*(\d{9})	Igno
Dated	Dated:\s*(\d{2})/(\d{2})/(\d{4})	Igno
Effective Date	Effective:\s*(\d{2})/(\d{2})/(\d{4})	Igno
► invoices		

Click on Edit

Configure Regular Expressions

Configure regular expressions for the information

RegEx Builder

Test Text

RegEx	Value	Quantifiers
<div>Literal</div>	FORM	<div>Exactly</div> <div>1</div> <div>+△▽×</div> <div><input type="checkbox"/> Capture</div>
<div>Whitespace</div>	\s	<div>Exactly</div> <div>1</div> <div>+△▽×</div> <div><input type="checkbox"/> Capture</div>
<div>Literal</div>	10-	<div>Exactly</div> <div>1</div> <div>+△▽×</div> <div><input type="checkbox"/> Capture</div>
<div>Anything</div>	.	<div>Exactly</div> <div>1</div> <div>+△▽×</div> <div><input checked="" type="checkbox"/> Capture</div>

Full Expression

FORM\s10-()

Regex Options

IgnoreCase

Save

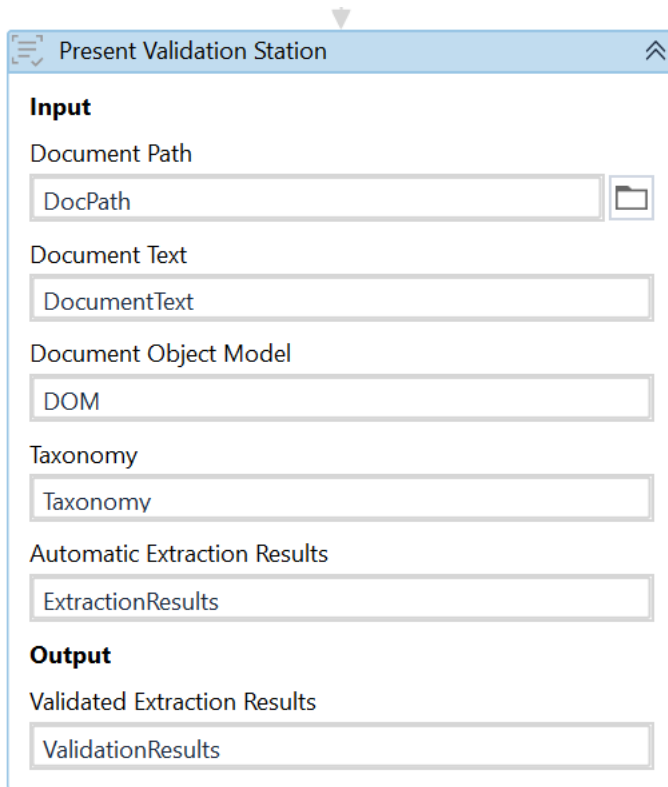
Cancel

Cancel

Step 5:


Validation:

This activity triggers the opening of the Validation station. it's the tool that allows you to review and, if necessary, correct the document classification and automatic data extraction results.



Present Validation Station

Input

Document Path
 

Document Text

Document Object Model

Taxonomy

Automatic Extraction Results

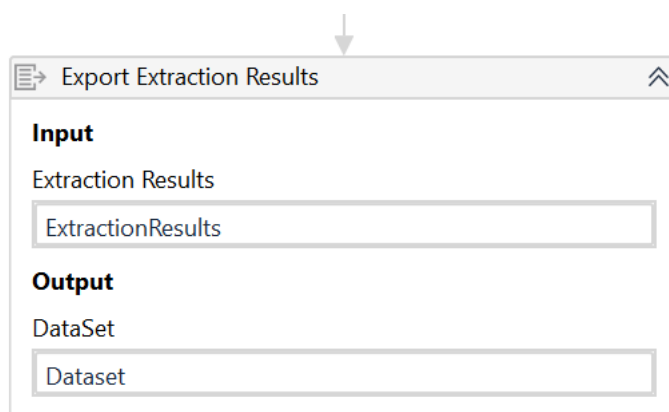
Output

Validated Extraction Results

Step 6:

Export

Add the Export Extraction Results activity, provide the validated extraction results as input, and create a new variable for the Output.



Export Extraction Results

Input

Extraction Results

Output

DataSet

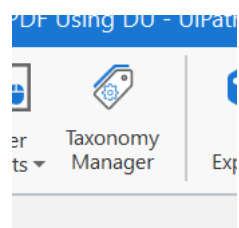
How to Extract table from pdf using Document Understanding Framework

Input Document:

INVOICE			
Sirius Cybernetics Corp. 4592 Bell Street New York, NY 10018			
Bill To	Ship To	Invoice #	890127
CHOAM	CHOAM	Invoice Date	27/01/2016
27 Shield Wall Ave,	27 Shield Wall Ave,	P.O.#	16012633
Carthag, CH 1965	Carthag, CH 1965	Due Date	26/02/2016
Arrakis	Arrakis		
Qty	Description	Unit Price	Amount
10	Nutrimatic Drinks Dispenser	4,200.00	42,000.00
17	Shipboard Computer "Eddie"	8,402.00	142,834.00
3	Happy Vertical People Transporters	21,000.00	63,000.00
Subtotal			247,834.00
VAT 19.0%			47,088.46
Total			\$294,922.46
Terms & Conditions Payment is due within 30 days		Thank you	

Step 1:

Click on Taxonomy Manager.



Document Types 1

Any Group +
Any Category +

Search by name

Add New Document Type

Click on a document type to edit it.

- Forms
 - Tax
 - PDF
- Certificate
 - image
 - Certificatefiling
- Syruis
 - pdf
 - invoices**
- W-9
 - Tax
 - Form

Document Type Details

Document Type ID: Syruis.pdf.invoices

Name
invoices

Group
Syruis

Category
pdf

Document Type Code
Optional value

Fields

- invoice no
- invoice date
- Total
- item**

New Field

Edit Field

Field ID: Syruis.pdf.invoices.item

Name
item

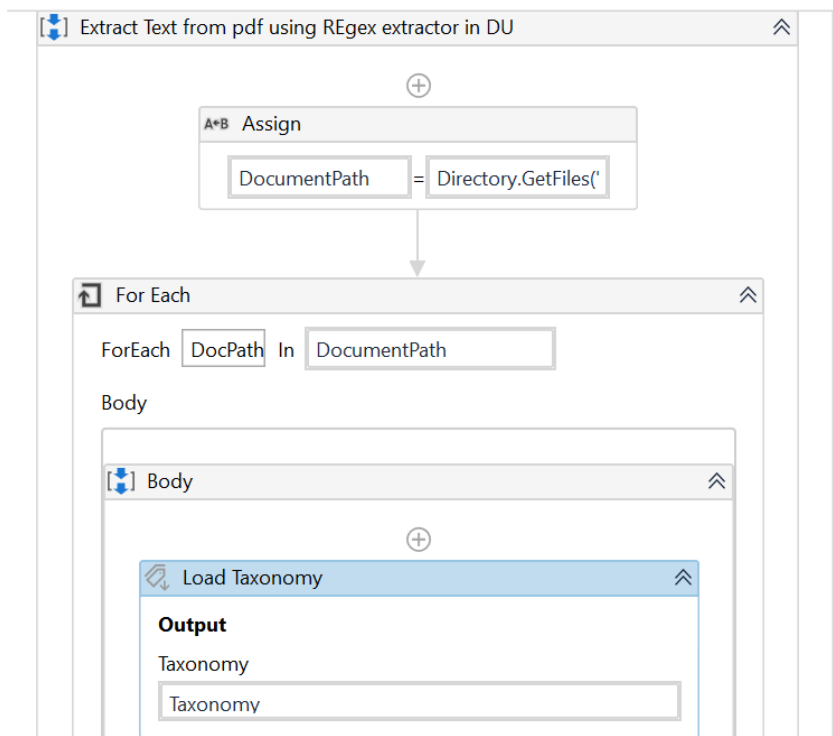
Type
Table

Columns

- Quantity
- Description
- unit price
- Amount

Add Column

Save Cancel



Step 2:

Digitize Document:

All documents that are to be processed (native and scanned) must pass through this step-in order for the robot to understand the kind of document it's working with and what data is relevant.

The OCR engine will be used only if the incoming documents require OCR processing, and the decision gets taken on a page-by-page basis.

Digitize Document

Input

Document Path

Output

Document Text

Document Object Model

OmniPage OCR

Step 3:

Classification is done through the Classify Document Scope and it's performed by classifiers.

Basically, the document text and object model resulted in the digitization steps are sent to the Classifiers, which report what types they recognize within the incoming file.

The Document Classification Scope Wizard will open at selecting the **Configure Classifiers** option and it allows users to customize which classifier will be used for each individual type of document.

Classify Document Scope

Input

Document Path

Taxonomy

Document Text

Document Object Model

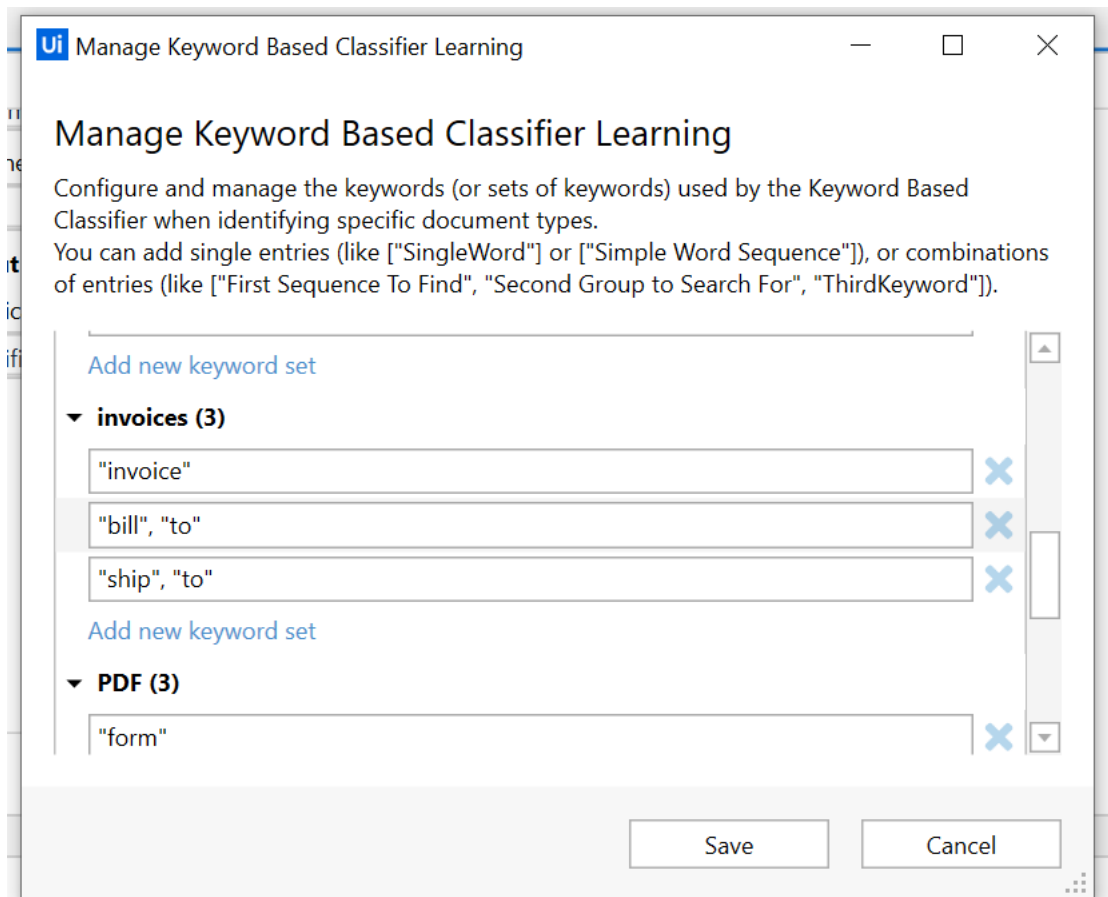
Output

Classification Results

Keyword Based Classifier

Learning File Path

[Manage Learning](#)



Step 4:

The input consists of the outputs of the previous Document Understanding steps, along with the document path. Based on the structure of the document, different extractors come to play.

The Data Extraction Scope Wizard will open at selecting the Configure Extractors option and it allows users to customize which extractors will be used for each individual field.

It allows users to mix and match extractors as well as use extractors in parallel based on which extractor has the highest confidence level.

Data Extraction Scope

Input

Document Path:

Document Text:

Document Object Model:

Taxonomy:

Classification Result:

Document Type Id:

Output

Extraction Results:

Form Extractor

Endpoint:

Api Key:

Regex Based Extractor

[Configure Expressions](#)

Machine Learning Extractor

Endpoint:

ML Skill:

Use Machine learning extractor to extract table from pdf.

Configure Extractors

Configure which extractors you want to apply to each document type and field.
To activate extractors for certain fields, check the appropriate boxes in the configurator.
For extractors that use your defined taxonomy for configuration and data extraction, entering the unique IDs for document types and fields is optional.
For extractors that have their own internal taxonomy, provide the internal taxonomy unique IDs for both document types and fields for which the extractor will be activated.

Document Types and Fields	Form Extractor Framework Alias: <input type="text"/> Minimum Confidence %: <input type="text" value="0"/>	Regex Based Extractor Framework Alias: <input type="text"/> Minimum Confidence %: <input type="text" value="0"/>	Machine Learning Extractor for invoices Framework Alias: <input type="text"/> Minimum Confidence %: <input type="text" value="0"/>
PDF	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/> N/A
Certificatefilling	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/> N/A
invoises	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/> N/A
invoice no	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/> invoice-no
invoice date	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/> due-date
Total	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/> total
item	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/> items
Quantity	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/> quantity
Description	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/> description
unit price	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/> unit-price
Amount	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/> line-amount
form	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/> N/A

Step 5:

Validation:

This activity triggers the opening of the Validation station. it's the tool that allows you to review and, if necessary, correct the document classification and automatic data extraction results.

Present Validation Station

Input

Document Path
DocPath

Document Text
DocumentText

Document Object Model
DOM

Taxonomy
Taxonomy

Automatic Extraction Results
ExtractionResults

Output

Validated Extraction Results
ValidationResults

Step 6:

Export

Add the Export Extraction Results activity, provide the validated extraction results as input, and create a new variable for the Output.

Export Extraction Results

Input

Extraction Results
ExtractionResults

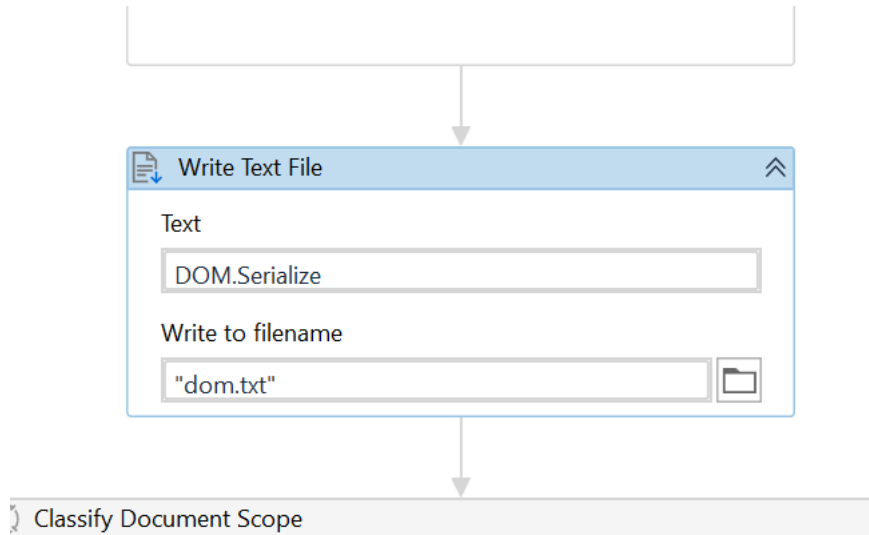
Output

DataSet
Dataset

What is DOM?

Document Object Model contains basic information regarding the processed document such as name, content type, text length, and information about each page. It is passed down to the next activities using a document variable.

The other outcome is all the text from the document, stored in a string variable.



Output Of DOM:

```
dom - Notepad
File Edit Format View Help
{"DocumentId":"10q-apple-2018-3.pdf","ContentType":"application/pdf","Length":8838,"Pages":[{"PageIndex":0,"Size":
[0.0,0.0,612.0,792.0],"Sections":
[{"IndexInText":0,"Language":"eng","Length":10,"Rotation":"None","SkewAngle":0.0,"Type":"Paragraph","WordGroups":
[{"IndexInText":0,"Length":10,"Type":"Other","Words":[{"Box":[120.7563,17.55,18.9412,10.1296],"Polygon":
[17.55,120.7563,36.4912,120.7563,36.4912,130.8859,17.55,130.8859],"IndexInText":0,"OcrConfidence":1.0,"Text":"(Mark","VisuallineNumber":4},
{"Box":[120.7563,38.5405,16.4835,10.1296],"Polygon":
[38.5405,120.7563,55.024,120.7563,55.024,130.8859,38.5405,130.8859],"IndexInText":6,"OcrConfidence":1.0,"Text":"One)","VisuallineNumber":4}}]}]
{"IndexInText":12,"Language":"eng","Length":71,"Rotation":"None","SkewAngle":0.0,"Type":"Paragraph","WordGroups":
[{"IndexInText":12,"Length":71,"Type":"Other","Words":[{"Box":[31.3113,259.8539,42.7099,16.4337],"Polygon":
[259.8539,31.3113,302.5638,31.3113,302.5638,47.745,259.8539,47.745],"IndexInText":12,"OcrConfidence":1.0,"Text":"UNITED","VisuallineNumber":0},
{"Box":[31.3113,305.7539,45.2689,16.4337],"Polygon":
[305.7539,31.3113,351.0228,31.3113,351.0228,47.745,305.7539,47.745],"IndexInText":19,"OcrConfidence":1.0,"Text":"STATES","VisuallineNumber":0},
{"Box":[44.8113,183.6633,68.8614,16.4337],"Polygon":
[183.6633,44.8113,252.5247,44.8113,252.5247,61.245,183.6633,61.245],"IndexInText":26,"OcrConfidence":1.0,"Text":"SECURITIES","VisuallineNumber":
1},{"Box":[44.8113,255.7148,24.8548,16.4337],"Polygon":
[255.7148,44.8113,280.5696,44.8113,280.5696,61.245,255.7148,61.245],"IndexInText":37,"OcrConfidence":1.0,"Text":"AND","VisuallineNumber":1},
{"Box":[44.8113,283.7597,65.0288,16.4337],"Polygon":
[283.7597,44.8113,348.7885,44.8113,348.7885,61.245,283.7597,61.245],"IndexInText":41,"OcrConfidence":1.0,"Text":"EXCHANGE","VisuallineNumber":1},
{"Box":[44.8113,351.9786,75.23,16.4337],"Polygon":
[351.9786,44.8113,427.2086,44.8113,427.2086,61.245,351.9786,61.245],"IndexInText":50,"OcrConfidence":1.0,"Text":"COMMISSION","VisuallineNumber":
1},{"Box":[59.0614,248.0836,60.75,14.5003],"Polygon":
[248.0836,59.0614,308.8336,59.0614,308.8336,73.5617,248.0836,73.5617],"IndexInText":61,"OcrConfidence":1.0,"Text":"Washington","VisuallineNumt
r":2},{"Box":[59.0614,311.6483,20.2501,14.5003],"Polygon":
[311.6483,59.0614,331.8984,59.0614,331.8984,73.5617,311.6483,73.5617],"IndexInText":73,"OcrConfidence":1.0,"Text":"D.C.","VisuallineNumber":2},
{"Box":[59.0614,334.7131,28.1475,14.5003],"Polygon":
[334.7131,59.0614,362.8606,59.0614,362.8606,73.5617,334.7131,73.5617],"IndexInText":78,"OcrConfidence":1.0,"Text":"20549","VisuallineNumber":2}
]}],{"IndexInText":85,"Language":"eng","Length":9,"Rotation":"None","SkewAngle":0.0,"Type":"Paragraph","WordGroups":
[{"IndexInText":85,"Length":9,"Type":"Heading","Words":[{"Box":[86.4358,263.1867,45.659,22.2338],"Polygon":
[263.1867,86.4358,308.8457,86.4358,308.8457,108.6696,263.1867,108.6696],"IndexInText":85,"OcrConfidence":1.0,"Text":"FORM","VisuallineNumber":3},
{"Box":[86.4358,313.1306,34.4655,22.2338],"Polygon":
[313.1306,86.4358,347.5961,86.4358,347.5961,108.6696,313.1306,108.6696],"IndexInText":90,"OcrConfidence":1.0,"Text":"10-
Q","VisuallineNumber":3}}]}],{"IndexInText":96,"Language":"eng","Length":89,"Rotation":"None","SkewAngle":0.0,"Type":"Paragraph","WordGroups":
[{"IndexInText":96,"Length":89,"Type":"Other","Words":[{"Box":[133.9488,90.7453,10.125,10.0854],"Polygon":
```