UiPath The Foundation of Innovation™

Document Understanding Framework Activities

Mastering the best practices for utilizing Document Understanding Studio activities to optimize your document processing workflows.

This guide equips you with tips for leveraging Document Understanding Studio activities to maximize efficiency and accuracy in document processing workflows.

Document Understanding Framework Activities

A Document Understanding framework is the steps/layers through which a document goes through when processing it. Through the UiPath Document Understanding framework, you can digitize files, validate extracted data, and process incoming files all in a versatile, extensible, and open environment.

Typically documents go through six fundamental steps namely, pre-processing (load taxonomy), digitization, classification, extraction, validation, and post-processing. The Document Understanding framework is found in the **UiPath.IntelligentOCR.Activities** package**.** Following are the tips for the scope activities and the wizards:

Taxonomy Manager

- Group the document types into meaningful groups and categories based on the business units.
- Always use meaningful names in naming the fields required from a document type.
- Avoid using special characters when naming the groups, categories, and fields.
- Provide appropriate data types for the fields depending on the values it holds. The data type also helps clear out unwanted characters from the value. (For example: The type of Number helps remove thousand separators and other special characters captured).
- Business Rules play an important role in the Document Understanding workflows. Building validation rules for simple verifications and cleansing may require a lot of effort. Using the Field-level Rules in Taxonomy Manager can optimize the validation effort. Use the field-level rules in the Taxonomy Manger to perform the following:
    - Check if a mandatory field is empty.
    - Check if a field contain only the values that are known (possible values)
    - Check if a value starts or ends with a specific string/ character.
    - Check if a value matches a specific string pattern using RegEx (Ex: Email)
    - Check if a field is empty.
- Use the Rules feature in Taxonomy Manager to build business rules that need to be fulfilled by the extraction result. The Rules option contains multiple options to use as needed to build effective validations.

Digitize Document

- Use the activity inside a Retry Scope activity that is placed within a Try Catch activity. This combination enables the bot to retry the Digitization in case of an error. Please pay attention that each retry attempt consumes a license.
- Select the OCR engine based on the documents, content, and required features. All OCR engines may not support all scenarios. For example, use UiPath Document OCR and OmniPage OCR if you are required to extract handwritten data.
- Perform OCR Benchmark testing when selecting the best OCR for your scenario. OCR Benchmark testing refers to testing multiple OCR engines on a set of real documents covering all variations. Analyze the test results and identify which OCR performed best in accurately extracting data.

- Make sure the same OCR engine is used when configuring all activities that imply digitization. For example: Intelligent Keyword Classifier, Form Extractor template definitions, etc. Changing the OCR engine requires you to reconfigure all these activities for the new OCR engine.

Classify Document Scope

- It is important to select the classification method according to the documents and the activities you perform. Classification is not a mandatory step.
- Check the classificaton results generated by the activity according to the business rules defined in the use case.
- Consider using multiple classification methods to address complex tasks that cannot be performed with a single classifier. For example, let's assume that the document contains highly unstructured data and requires splitting into different classification types. These documents may be a best fit for the combination of Machine Learning Classifier (classifying unstructured documents) and Intelligent Keyword Classifier (for splitting).
- Go through all the document variations and decide which classifier or combination of classifiers will generate the best result.

Keyword-based Classifier

- The list of keywords used for each category should be distinctive. These keywords should cover all variations of keywords and phrases that might appear in the documents.

Intelligent Keyword Classifier Training

- Consider the following best practices if a file has multiple classification types, and each type has varying layouts.
    - Split the original document into single page files.
    - The activity generates the page range as the output for each classification which can be used with UiPath PDF activities to split and create a new file.
    - Feed the single page files into Intelligent Keyword Classifier training and train the classification type with the variations.
- Use multiple training documents for each document type to train and get better classification results. The higher the number of variations it sees, the higher the accuracy levels are.

  Note: This method helps the Intelligent Keyword Classifier to understand the variations of each page that fall into a specific classification type. It is much more effective than uploading a full document at once for training.

Classification Station and Classification Action

- Use Classification Station only when manual classification or classification validation is required.
- Always use the **Present Classification Station** activity within a Try Catch activity to handle exceptions. This also helps to handle user defined exceptions to exclude a document from the Document Understanding flow.
- Consider all user defined exceptions as Business Rule Exceptions.

Document Data Extraction Scope

- Use the extraction confidence only as a reference and not as the only source to check the accuracy. Sometimes a wrong value may provide a higher confidence and a correct value may provide a lower confidence. Hence, it is important to always cross validate the extracted data to ensure the accuracy of the data.
- Consider checking the OCR confidence when extracting data from scanned documents.
- Place the Document Data Extraction Scope activity within a Try Catch activity to handle unexpected errors (example: Timeout errors) during runtime.

RegEx Based Extractor

- Ideal for scenarios where you have limited variations in documents with fixed fields.
- Regular Expressions work with the patterns identified in the Document Text (output of Digitize Document activity). Perform the following to build a reliable regular expression.
  - Perform the Digitization for a sample document that you need to build the regex.
  - Extract the digitized document text and write it into a temporary text file.
  - Copy the text from the file and use it in the RegEx Builder as the Test Text. The patterns you build should match the text.
  - Perform some test runs with several similar types of documents to validate and optimize your Regular Expression.

Form Extractor

- Ideal for scenarios when processing fixed form documents.
- Defining the Custom Range for the value to be extracted for a field should cover the entire space available for the field. This allows extracting all the text inside the specified range.
- Form Extractor works with a template. Hence, slight changes in the document may cause the template to fail in extracting values. For example, let's assume that the document template is created on a native PDF. However, the document received is scanned, tilted and smaller in size. This may cause the values not to fit into the regions defined in the template. It can be resolved by using anchors when defining the regions. Anchors help identify the region based on a fixed value next to the field (like Anchors used in Ui Automation) allowing accurate extraction in exceptional scenarios.
- Always use native PDFs for template definitions, if possible. If native PDF is not available, use the same OCR engine that is used for Digitization.
- Pick document & page matching words from all areas of the page, not just from the top.
- The extraction areas should be defined as large as possible, but without any overlapping.

Machine Learning Extractor

- Ideal for processing semi-structured documents such as invoices, purchase orders, receipts etc.

Validation Station and Validation Action

- Always wrap the Validation Station activity within a Try Catch activity. This allows handling unexpected errors (if any) and user defined document exceptions.
- The user defined exceptions should be handled as a Business Rule Exception
- Use the Validation Station only when manual verification is required. Ideally this step should come in after performing the extraction and the post processing of extracted data.

Train Extractor Scope

- Use the Train Extractor Scope and Machine Learning Extractor Trainer only on the documents that got verified manually.

Train Classifier Scope

- Use the Train Classifier Scope and its Trainer activities only on the documents that get classified manually.
- Use all the respective Trainer activities for each classifier or combination of classifiers used in the Classification Scope activity.

Export Extraction Results

- Store extracted data in a persistent storage location so it can be retrieved without re-executing the DU process.