## Document Understanding - Estimate

## Document Understanding

## Different types of Files

Documents are classified into three main formats:

* **Structured**: documents designed to collect information in a specific format. For example, surveys, tax forms, passports, or licenses are all structured documents.
* **Semi-structured**: documents that do not follow a strict format and are not bound to specified data fields. Semi-structured documents include invoices, receipts, utility bills, bank statements, and others.
* **Unstructured**: documents that do not follow a specific or organized model. For example, contracts, leases, or news articles are all unstructured documents.

## What is a classifier?

It’s a component in DU, which helps to identify the type of files the robot is processing.

A file can be classified into one or more document types, depending on its content and the classification methods used:

* If a file contains **a single logical document type** (e.g., it is an Invoice or a Medical Record in all its entirety), then the classification component should be configured accordingly **and return a single classification result**.
* If a file contains **multiple logical document types** (e.g., it contains an Invoice from page 1 to page 5, and a Medical Record for the next 10 pages, and an Insurance Agreement from page 16 to the end), then the classification component should **return multiple classification results**, each **corresponding to the right page range from the input file.**

## When document classification should be used

If a project needs to process files that are **all of the same document type** and are **always present** as **one instance per file** (e.g., one invoice in one file), then **classification is not necessary** and can be skipped in its entirety.

 If the project is dealing with **two or more document types** (e.g., the workflow must process Invoices and Medical Records which cannot be distinguished before processing), **or files** are sometimes expected to contain **two or more distinct document types** within them (e.g., one file contains 3 Invoices), **then classification is strongly recommended**.

## How to use the document classification component

[Classify Document Scope](https://docs.uipath.com/activities/other/latest/document-understanding/classify-document-scope) activity to classify the documents. We can use one or more classifiers.

## What the **Classify Document Scope** does:

* Provides all Classifiers the necessary configurations for them to run.
* Accepts one or more classifiers.
* Allows for document type filtering, taxonomy mapping, and minimum confidence threshold settings at classifier level.
* Reports classification information in a unified manner, irrespective of the source of classification.

The **Classify Document Scope** allows you to configure it by using the [Configure Classifiers](https://docs.uipath.com/document-understanding/automation-cloud/latest/classic-user-guide/configure-classifiers-wizard-classify-document-scope) wizard.

* Which document types are accepted from which classifier,
* What is the minimum confidence threshold for a given result that is acceptable for each classifier,

**The order of the classifiers in the Classify Document Scope** is important:

* classifiers are executed with priority, from left to right.

The available classifiers are:

* [Keyword Based Classifier](https://docs.uipath.com/document-understanding/automation-cloud/latest/classic-user-guide/keyword-based-classifier)
* [Intelligent Keyword Classifier](https://docs.uipath.com/document-understanding/automation-cloud/latest/classic-user-guide/intelligent-keyword-classifier)
* [Machine Learning Classifier](https://docs.uipath.com/document-understanding/automation-cloud/latest/classic-user-guide/machine-learning-classifier)
* [Generative Classifier](https://docs.uipath.com/activities/other/latest/document-understanding/document-understanding-ml-generative-classifier)

## [Intelligent Keyword Classifier](https://docs.uipath.com/document-understanding/automation-cloud/latest/classic-user-guide/intelligent-keyword-classifier)

Utilizes **word vector** it learns from files of certain document types to perform document classification.

* Concept of repeating content for the same document type
* Vector similarity computation.

## When to use

You should consider using this classifier if:

* your files contain one or more document types within a single file
* your document types are relatively easy to differentiate as far as content goes.

## If Unable to Classify, then what?

If we are using a **classifier** that we have trained with specific document types, the classifier will try to match the document to the trained categories. If it can't match the document to any of the trained classes, it will typically **classify it as an unknown document** and could either:

* Return a **low-confidence classification** (indicating it is unsure).
* Return a **default class**, indicating that the document doesn't fit any of the trained categories.

**Fallback Mechanisms:**

* In both cases, the **fallback mechanism** might be triggered. You can configure fallback logic to handle documents that do not fit the trained categories by either:  
  + Classifying them as "Unknown."
  + Using a *generic category* for *unclassified* documents.
  + Escalating them for human review.

4. Extractors:

There were invoices and receipts

where did you use intelligent and where keyword classifier?

9. If you reject the document, what is the expected behaviour?

10. what is the identifier , how will you know the human rejected the document?

variable name: failed exception list...?

11. Have you used DU framework?

12. What is the output of Digitize activity or other activities...?

DOM, Document

13. Any experience in working with CV based projects?

How did you utilize the CV and why?

Computer Vision Scope Activity...

Citrix or type of application

14. Computer Vision Vs Remote Runtime? Which one would you use? Why?

15. There are challenges / constraints while installing the remote runtime.. therefore sometimes we need to use Computer Vision based activities.

16. How do you estimate DU based projects?

Step by step estinmation..

how do you derive the estimates...

17. How do you identify the complexity.. How do you quantify if a document is complex or not?

- Martix / Comparision Sheet

------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

1. What type of insurance?

Healthcare?

- Eligibility Verification in terms of insurance

2. DU and CV..

Revenue cycle management - get it in the form of Citrix...?

What are the challenges do you forsee while developing in the prod?

1. Sensitive Information regarding customers

2. Any Update during the process?

3. How do you cover the test scenarios?

Send messages to users

4. How do you complete the development if the scenario is not present?

- We have to work with what we have?

5. How you ensure the team completes the development?

3. Differnet type of documents you have extracted in insurance domain.

4. What is structured, semi, unstructured documnet?

5. You have to estimate DU project timeline..

Number of templates / Type of templates / Layout of templates

Number of pages

6. I have a doc in which I have a invoice, purchase order, correspondance

How do you classify the document?

Unique Words, Document Type

7. A new document? what's the behaviour?

classification threshold

send to action center

We have the comment form Action Center

based on this we can process further

8. I have docs... In those doc, 20 fields are same. and some are different

Need classifier - to differenciate based on the fields (Doc 1 or 2 or 3)

Form Extractor or ML Extractor (Need to mention the unique fields)

9. Then why do we need to classify if I need to use only 1 skill?

Suppose we want to insert the validation based on document type...

What type of validations?

-- If mandatory fields are not there send to Action Center

10. I have a process, I am splitting the process into stages.... same as above.