Document Understanding Evaluation Template

Masaki Kumamoto – masaki.kumamoto@uipath.com – 0.2.1, July 16, 2021 | Initial project

Table of Contents

Resources
Overview
What is Document Understanding Evaluation Template?
Features
Use cases
Quick Start Guide
Preparation Steps
Development Steps
Execution Steps
Improvement Steps
Other useful features
Multi OCR execution
Auto verification confidence threshold
Show OcrConfidence in validation action
Use existing Action.xlsx as an actual value reference
Details of the reporting files
DU_Evaluation.xlsx
DU_Evaluation.xlsx

Resources

- Document.Understanding.Evaluation.Template Nupkg
  (Download from Assets)

Overview

What is Document Understanding Evaluation Template?

This template project facilitates the efficient development of workflows that output information about the extraction accuracy of Document Understanding in a beautiful Excel format automatically.

Features

Minimum Effort Development

You can just focus on…

- Taxonomy definition
- Classification/Extraction logic

You do not even need to add any additional activities, valuables and arguments!

Faster Validation Flow

Since Validation Actions will be uploaded in bulk, the lead time can be shortened because the “Extraction by Robots” and “Validation by human" can be processed in parallel for multiple documents.

Multiple OCRs Evaluation

By enabling specific OCRs in Config.xlsx, you can apply multiple OCRs to the DU and create evaluation reports for each OCR simultaneously.

Use cases

DU accuracy reporting

Rendering the extraction accuracy of a DU for given documents in a beautiful Excel format automatically to assist in evaluating the functionality of the DU

OCR Comparison

Comparing the extraction accuracy of each OCR applied to DU

Improve DU accuracy

Optimizing the development process to improve DU’s document extraction accuracy based on the benchmarks

Action Center Demo

Demonstration for building Document Validation Actions in the Action Center

Quick Start Guide

After creating a new project from the Document Understanding Evaluation Template, you can accomplish the goal with a minimum of effort by following the steps described below.

Preparation Steps

1. Place `Document.Understanding.Evaluation.Template.X.X.X.nupkg` in the template folder (Download Document.Understanding.Evaluation.Template Nupkg from here)
   - You can identify the template folder below.

     ```
     UiPath Studio > Settings > Locations > Publish project templates URL
     e.g. "C:\Users\masaki.kumamoto\Documents\UiPath\.templates"
     ```

2. Create a new project from the template

   ```
   UiPath Studio > Templates > Document Understanding Evaluation Template
   ```

3. Wait until the project will resolve the dependency
   - If Studio can't auto-resolve the dependency for `Microsoft.Office.interop.Excel`, enable nuget.org (https://api.nuget.org/v3/index.json) as a package source in package manager and install the package.

Development Steps

1. Place target documents in Input folder
   - If you use Form Extractor/Intelligent Form Extractor, you can also place template documents in "TemplateDocument" Folder).
   - The file name placed in the Input Folder should not exceed 31 characters including the file extension. If the number of characters exceeds 31, an error will occur when creating the sheet in the Report Excel file.
2. Configure `Config.xlsx`
   - "DUSettings" Sheet::
     - DU_ApiKey
     - DU_DocumentTypeId
       (If you use Classification activity, you do NOT need to specify this field. You can find the value in Taxonomy Manager once you create a taxonom in step 3.)
   - "ActionSettings" Sheet
     - AC_AssignUserEmail
     - OC_FolderPath

       Orchestrator Folder name that your Studio/Robot is deployed in
     - SB_BacketName

       Set the same name Storage Bucket in Orchestrator
   - "OcrSettings" Sheet
     - Set TRUE for the OCR to be applied to DU. You can enable multiple OCRs to be applied.
       (UiPath Document Understanding OCR will always be performed)

3. Define Taxonomy
    - Set the definition of the field information to be extracted from Ribbon>Design>Taxonomy Manager.
      (How to use Taxonomy Manager)
4. Build `DU_GetExtractionResult.xaml`
    - If you want to use Classification, Enable “Classification + Extraction” sequence
      If not, Enable “Extraction Only” sequence
    - Delete Classification/Extraction Activities which you do not use
    - For more information on how to develop DU Classification/Extraction, you can refer to the following links.
        - Classify Document Scope
            - Keyword Based Classifier
            - Intelligent Keyword Classifier
        - Data Extraction Scope
            - Regex Based Extractor
            - Form Extractor
            - Intelligent Form Extractor
            - Machine Learning Extractor
                - Public Endpoints List

Execution Steps

1. Run `01_ExtractDocumentsData.xaml`
    - You should stop the OneDrive sync function while the process is running otherwise an error may occur.
    - It takes about 1-2 minutes to process each document.
      (It would take more with “Debug” so “Run” is recommended)
    - After the execution is complete, Excel reports for each OCR set in Config.xlsx and the Document Validation Action in Action Center will be generated.
2. Complete the Document Validation Action task in Action Center
3. Run `02_CopyActualValuesToReport.xaml`
    - Immediately after the execution, the robot will prompt the user to select a folder where the DU evaluation reports are located.
    - After the execution is completed, the results of the Document Validation Actions will be pasted to the `ActionList.xlsx` and the `DU Evaluation Reports` for each OCR.
    - If there are documents that have not yet been validated by Document Validation Actions when Step 3 is completed, complete the validation in Validation Action and then execute Step 3 again to complete DU evaluation reports.

Improvement Steps

Use existing ActionList.xlsx to improve the DU logic

If you have performed the "Execution Steps" and generated ActionList.xlsx for the same list of documents using the same taxonomy in the past, from next time, you can skip step 2 & 3 by following the steps below. You can also disable to create Document Validation Action so the process can run faster.

This capability is useful to modify the workflow based on the accuracy rate from previous execution result report, so you can improve the DU’s classification/extraction logic.

- Configure `Config.xlsx`
  - "BasicSettings" Sheet
    - AL_UseExistingActionListExcel (= TRUE)
    - AL_ExistingActionListExcelPath
  - "ActionSettings" Sheet
    - AC_DocumentValidationAction_Use (= False)

Other useful features

Multi OCR execution

By enabling specific OCRs in Config.xlsx, you can apply multiple OCRs to the DU and create evaluation reports for each OCR simultaneously. (UiPath Document OCR will be always used)

e.g. Use "TesseractOCR" and "OmniPageOCR"

Config.xlsx(OcrSettings sheet)

Header name	Description
OCR_TesseractOCR_Use	TRUE
OCR_TesseractOCR_Language	eng
OCR_OmniPageOCR_Use	TRUE
OCR_OmniPageOCR_Language	eng

Auto verification confidence threshold

You can use extraction Confidence and OcrOonfidence as a threshold for auto verification.
If both of them are above or equal to thresholds, the fields will be automatically verified by Robots.

e.g. Confidence Threshold = 99.98%, OcrConfidence Threshold = 95.99%

Config.xlsx(OcrSettings sheet)

Header name	Description
DU_AutoVerifyMinimumThreshold_Confidence	99.98%
DU_AutoVerifyMinimumThreshold_OcrConfidence	95.99%

Show OcrConfidence in validation action

You can select Confidence or OcrConfidence as the value to be displayed in the Action Center.
Depends on the documents set you deal with, chose the proper one.

e.g. Show OcrConfidence in Document Validation Actions instead of Confidence

Config.xlsx(DuSettings sheet)

Header name	Description
DU_ValidationConfidenceType	OcrConfidence

Use existing Action.xlsx as an actual value reference

If you have performed the "Execution Steps" and generated ActionList.xlsx for the same list of documents using the same taxonomy in the past, from next time, you can skip step validation in Action Center and execution of 02_CopyActualValuesToReport.xaml.

e.g. Use existing Action.xlsx(Output/20210601/ActionList.xlsx) as an actual value reference

Config.xlsx(DuSettings sheet)

Header name	Description
AL_UseExistingActionListExcel	TRUE
AL_ExistingActionListExcelPath	Output/20210601/ActionList.xlsx

Details of the reporting files

DU_Evaluation.xlsx

This file contains the percentage of correct extractions for all target documents and detailed extraction results. Files will be generated for the number of OCRs defined in Config.xlsx.

Summary sheet

This sheet renders the percentage of correct extractions for all target documents extracted by DU.

Extraction/Actual value report sheets (per target documents)

DU_Evaluation.xlsx (Extraction/Actual value report sheets)

Header name	Description
FieldName	Field name
FieldType	Field type
isMissing	If extractor missed the field or not
ValuesCount	Numbers of values which was extracted
Confidence	Confidence level for location
OcrConfidence	Confidence level for OCR
ExtractedValue	Extracted value by DU
ExtractedPage	Page number which include extracted value
ActualValue	Validated value
ActualPage	Page number which include validated value
isCorrect	If the extracted value is correct or not

DU_Evaluation.xlsx

This file contains information of the generated Document Validation Actions and the values.
This file will be used by Robots to get the validation results.

Actions sheet

Action.xlsx (Actions sheet)

Header name	Description
File Name	Target document file name
TaskId	Task Id of the Action
Status	Status of the Action
CreationTime	Creation time of the Action

Header name	Description
LastModificationTime	Last modification time of the Action
ActionUrl	URL of the Action

Actual value report sheets (per target documents)

Action.xlsx (Actions sheet)

Header name	Description
FieldName	Field name
ActualValue	Validated value
ActualPage	Page number which include validated value

0.2.1
Last updated 2021-07-16 11:25:35 -0700