

Document Understanding

評価テンプレート



Masaki Kumamoto
Sales Engineer
UiPath Canada



本資料について



- 本資料は、以前、隈元さんが公開されていた英語の資料を日本語訳したものです。
- 適宜、補足コメントや手順を追記しております。

Slide update

Date	Document Version	Description
Jun 3, 2021	0.0.1_doc0.1	Initial document
Jun 8, 2021	0.0.3_doc0.1	Update for v0.0.3
Jun 8, 2021	0.0.4_doc0.1	Update for v0.0.4
Jun 8, 2021	0.0.5_doc0.1	Update for v0.0.5
Jun 9, 2021	0.0.7_doc0.1	Update for v0.0.7
Jul 16, 2021	0.2.1_doc0.1	Update for v0.2.1
Mar 24, 2022	0.2.3_doc0.1	Update for v0.2.3
May 16, 2024	0.3.0_doc0.1	Update for v0.3.0

アジェンダ

1. 概要

- DU 評価テンプレートとは？
- 特徴
- 使用例

2. 報告ファイルの詳細

- DU_Evaluation.xlsx
- ActionList.xlsx

3. クイックスタートガイド

- 準備ステップ
- 開発ステップ
- 実行ステップ
- 改善ステップ

4. その他の便利な機能

- 複数OCRエンジンの実行
- 信頼度閾値による自動検証
- 検証アクションにOcrConfidenceを表示する
- 既存ActionList.xlsxの内容を実際の値として利用する

5. 制限事項と既知の問題

6. リリースノート

Document Understanding Evaluation Template

1

概要

DU 評価テンプレートとは？

このテンプレートプロジェクトは、Document Understandingの抽出精度データをクリーンなExcel形式で出力する効率的なワークフロー開発を実現します。

FileName	Total_Accuracy
invoice_00.pdf	100%
invoice_01.pdf	100%
invoice_02.pdf	64%
	88%



ミニマム・エフォート開発

編集するのは3つのファイルのみ！



Config.xlsx



taxonomy.json
n
(Taxonomy
Manager)



DU

以下の定義に注力するだけ...

- タクソノミー定義
- 分類／抽出ロジック

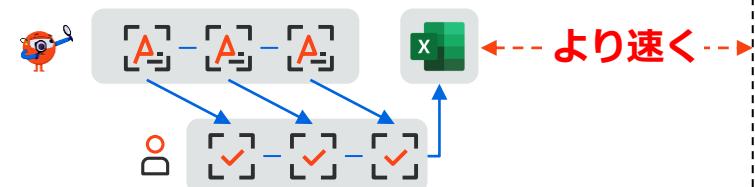
アクティビティ、変数、引数などを追加する必要はありません

より速い検証フロー

テンプレートなし

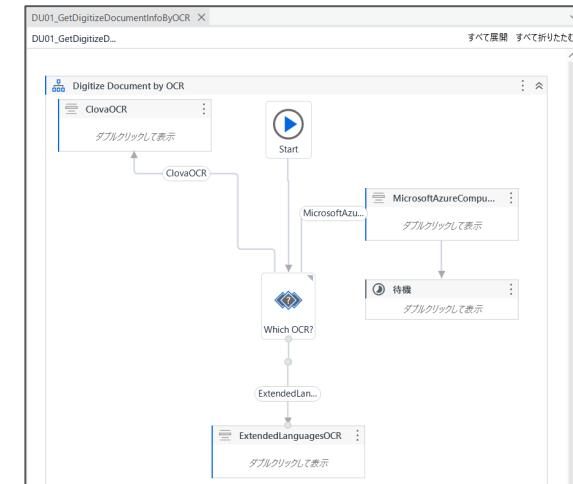


テンプレートあり



複数のドキュメントに対して一括で「ロボットによる抽出」と「人間による検証」が処理できるため検証時間の短縮が可能です

複数OCRの評価



Config.xlsxで特定のOCR定義を有効にすることで、DUに複数のOCRエンジンを適用し、各OCRの評価レポートを同時に作成することができます

使用例

DU精度報告



与えられた文書に対する
DUの抽出精度を美しい
Excel形式で自動的にレ
ンダリングし、DUの機
能評価を支援します

OCRの比較



DUに適用された各OCR
の抽出精度が比較可能

DUの精度向上



ベンチマークに基づき、
DUの文書抽出精度を向
上させるための開発プロ
セスの最適化

Action Center のデモ



Action Centerでドキュ
メント検証アクションを
構築するためのデモ

Document Understanding Evaluation Template

2

報告ファイルの詳細

このファイルには、全てのターゲット文書に対する正しい抽出の割合と詳細な抽出結果が格納されています。ファイルはConfig.xlsxで定義されたOCRの数だけ生成されます。

サマリーシート

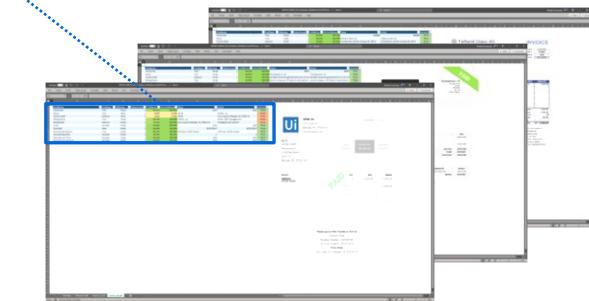
FileName	Total_Accuracy
invoice_00.pdf	100%
invoice_01.pdf	100%
invoice_02.pdf	64%
	88%

このシートは、DUによって抽出されたすべてのターゲット文書について、正しい抽出のパーセンテージを表示します

ヘッダー名	説明
Boolean_Accuracy	Boolean フィールドの抽出精度
Others_Accuracy	Text、Number、Date、Name、Address、Set フィールドの抽出精度
Total_Accuracy	全フィールドの抽出精度

抽出・実績値報告シート（ターゲット文書ごと）

FieldName	FieldType	isMissing	ValuesCount	Confidence	OcrConfidence	Value	Actual	isCorrect
Invoice No	Text	FALSE	1	85.21%	100.0%		1234	TRUE
Name	Text	TRUE	0	0.00%	0.00% [N/A]	UiPath, Inc.	1234	FALSE
Vendor Addr	Address	TRUE	0	0.00%	0.00% [N/A]	123 Long Ave Raleigh, NC 27610 US	1234	FALSE
Billing Name	Text	FALSE	1	82.41%	100.00%	UiPath, Inc.	ACME CORP Management	FALSE
Billing Addr	Address	FALSE	1	78.39%	100.00%	123 Long Ave Raleigh, NC 27610 US	1122Big Street Suite 87	FALSE
Total	Number	FALSE	1	92.26%	100.00%		1234	TRUE
Due Date	Date	FALSE	1	96.59%	100.00%		10/22/2017	10/22/2017
Items(0).Description	Text	FALSE	1	88.12%	100.00%	Software UIPath Studio	Software UIPath Studio	TRUE
Items(0).Quantity	Number	FALSE	1	86.96%	100.00%		1	TRUE
Items(0).Unit Price	Number	FALSE	1	93.01%	100.00%		1234	TRUE
Items(0).Line Amount	Number	FALSE	1	96.90%	100.00%		1234	TRUE



ヘッダー名	説明
FieldName	フィールド名
FieldType	フィールドタイプ
isMissing	抽出器がフィールドを外したかどうか
ValuesCount	抽出された値の数
Confidence	信頼度
OcrConfidence	OCR信頼度
ExtractedValue	DUが抽出した値
ActualValue	実際の値
isCorrect	抽出された値が正しいかどうか

このファイルには、生成されたドキュメント検証アクションとその値の情報が含まれています。このファイルはロボットがアクションセンターから検証結果を取得するために使用されます。

アクションシート

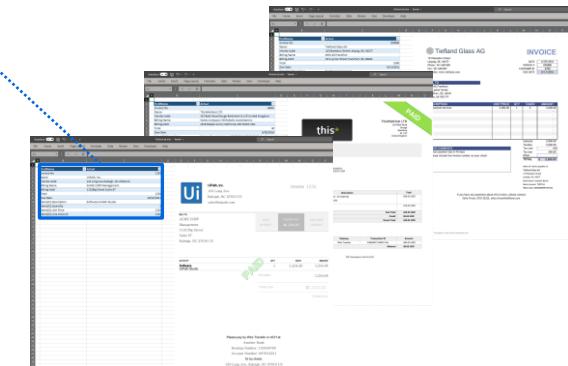
FileName	TaskId	Status	CreationTime	LastModificationTime	ActionUrl
invoice_00.pdf	58655	Pending	5/12/2021 16:12	5/12/2021 16:12	https://clo
invoice_01.pdf	58656	Pending	5/12/2021 16:13	5/12/2021 16:13	https://clo
invoice_02.pdf	58657	Pending	5/12/2021 16:13	5/12/2021 16:13	https://clo

実績値報告シート（ターゲット文書ごと）

FieldName	Actual
Invoice No	1234
Name	UiPath, Inc.
Vendor Addr	123 Long Ave Raleigh, NC 27610 US
Billing Name	ACME CORP Management
Billing Addr	1122Big Street Suite 87
Total	1234
Due Date	10/22/2017
Items(0).Description	Software UiPath Studio
Items(0).Quantity	1
Items(0).Unit Price	1234
Items(0).Line Amount	1234

ヘッダー名	説明
FileName	対象文書ファイル名
TaskId	アクションのタスクID
Status	活動状況
CreationTime	アクションの作成時間
LastModificationTime	アクションの最終更新時刻
ActionUrl	アクションのURL

ヘッダー名	説明
FieldName	フィールド名
ActualValue	検証値（実際の値）



Document Understanding Evaluation Template

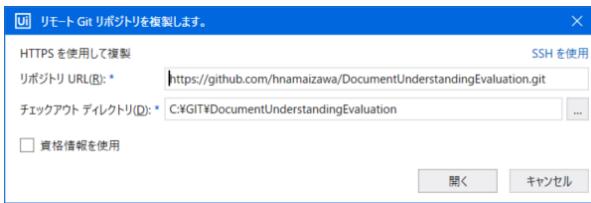
3

クイックスタートガイド

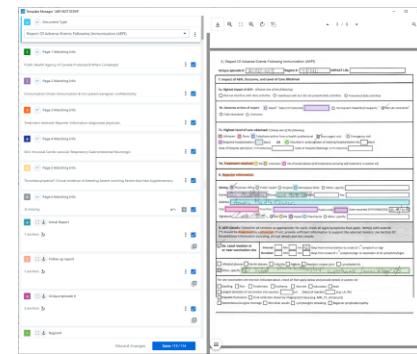
クイックスタートガイド

GitHub の DU 評価テンプレート リポジトリを複製した後、以下のステップを踏むことで、最小限の労力で目標を達成することができます。
(詳細な手順は次のページにあります)

準備



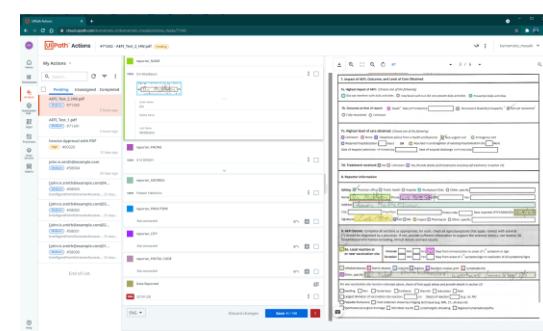
開発



1. チェックアウト用フォルダを用意する
2. GitHub のリポジトリを複製する
3. カスタムアクティビティを導入する

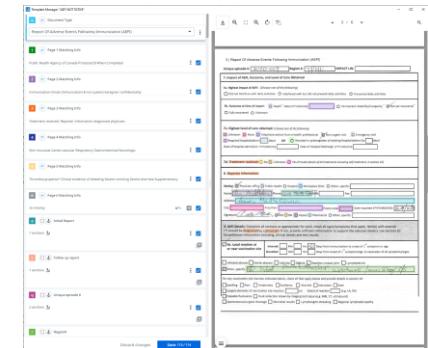
1. 対象文書をInputフォルダに入れる
2. [Config.xlsx](#) の修正
3. タクソノミーの定義
4. [DU_GetExtractionResult.xaml](#) のビルド

実行



1. [01_ExtractDocumentsData.xaml](#) を実行
2. Action Center のドキュメント検証タスクを完了させる
3. [02_CopyActualValuesToReport.xaml](#) を実行

改善



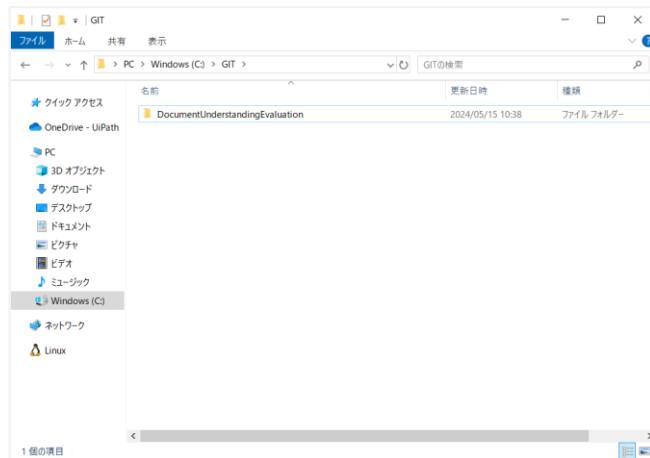
1. 作成された [ActionList.xlsx](#) を利用し DU ロジックを改善する

準備ステップ



1. チェックアウト用フォルダを用意する

- 任意の場所でフォルダを作成



2. GitHub のリポジトリを複製する

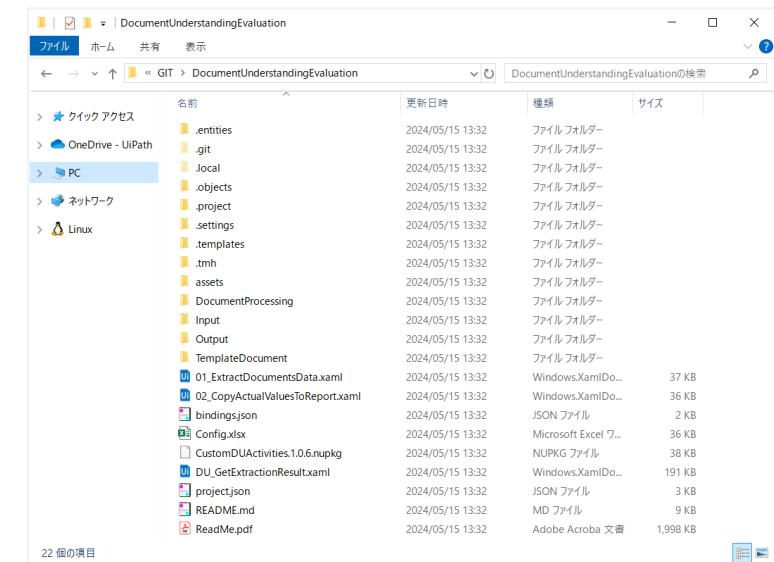
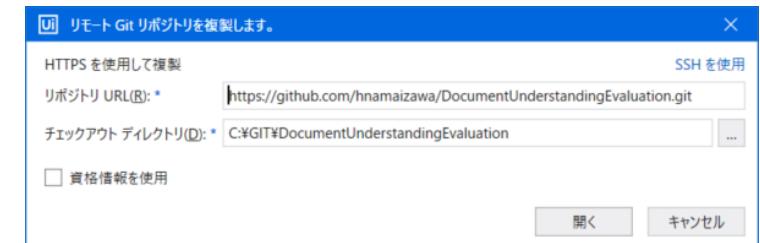
- UiPath Studio > スタート > 複製またはチェックアウト



- リポジトリを複製



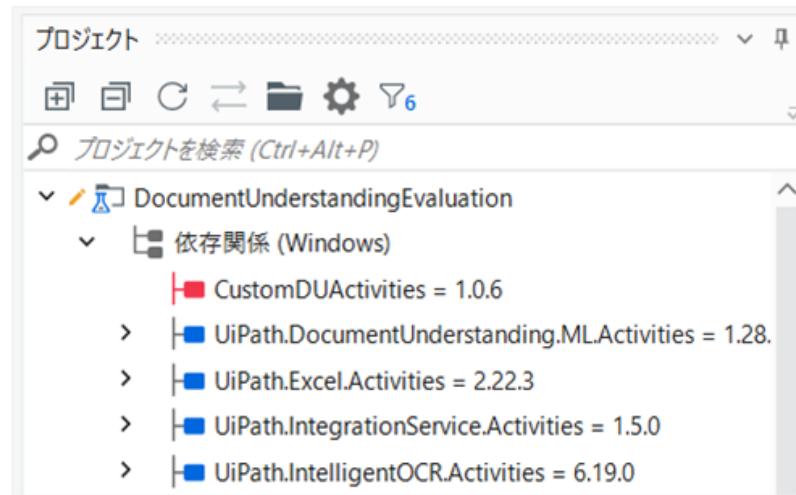
- GitHub リポジトリの URL とチェックアウト用フォルダを指定し、プロジェクトを開く
(<https://github.com/hnamaizawa/DocumentUnderstandingEvaluation.git>)



準備ステップ

3. カスタムアクティビティを導入する

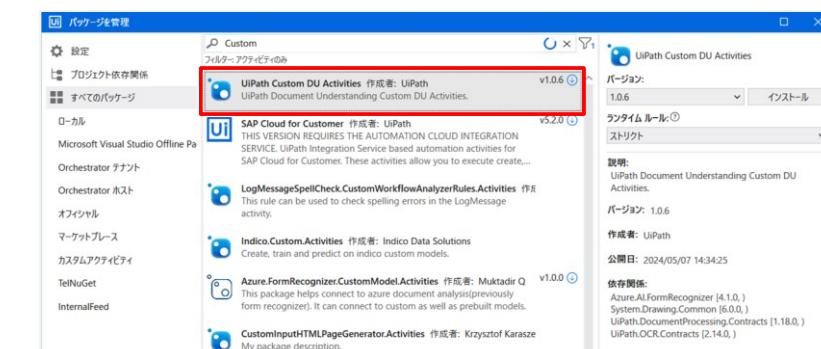
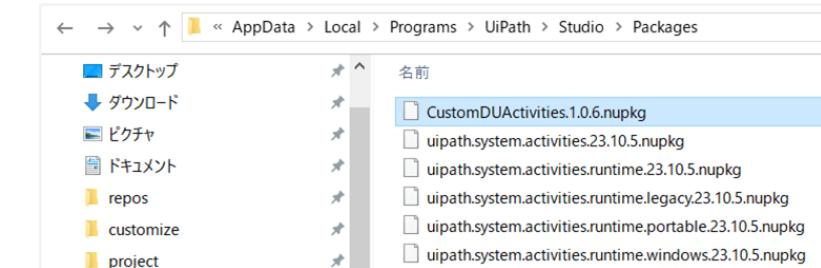
- DU 評価テンプレートは LINE WORKS OCR(Clova OCR) カスタムアクティビティを利用しているため、最初に CustomDUActivities.1.0.x.nupkgを導入します



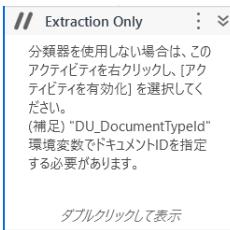
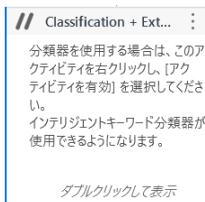
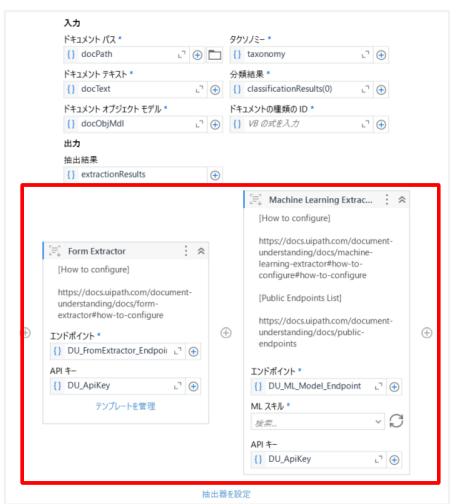
- カスタムアクティビティのソースコードは[こちら](#)ですが、事前にビルトしたnupkgファイルが含まれています

Input	2024/05/13 18:11
Output	2024/05/13 18:11
TemplateDocument	2024/05/13 18:11
01_ExtractDocumentsData.xaml	2024/05/13 18:11
02_CopyActualValuesToReport.xaml	2024/05/13 18:11
bindings.json	2024/05/13 18:11
Config.xlsx	2024/05/13 18:11
CustomDUActivities.1.0.6.nupkg	2024/05/13 18:11
DU_GetExtractionResult.xaml	2024/05/13 18:11
project.json	2024/05/13 18:33
README.md	2024/05/13 18:11
ReadMe.pdf	2024/05/13 18:11

- nupkgファイルを C:\Users\{ユーザー名}\AppData\Local\Programs\UiPath\Studio\Packagesへコピーし、[パッケージ管理]より導入します。



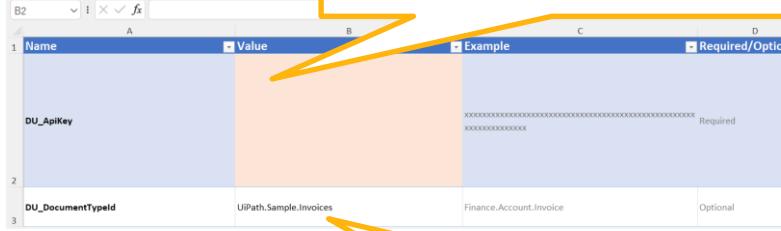
開発ステップ[®]

- 対象文書を **Input** フォルダに入れる
- Config.xlsx** を変更する
 - DUSettings** シート
 - DU_ApiKey
 - DU_DocumentTypeId (分類器を使用する場合、このファイルを指定する必要はありません。タクソノミーを作成すればタクソノミーマネージャからIDが確認可能)
 - ActionSettings** シート
 - AC_AssignUserEmail
 - OC_FolderPath (スタジオ/ロボットがデプロイされる Orchestrator フォルダ名)
 - SB_BucketName (Orchestrator のストレージバケットと同じ名前を設定する)
 - OcrSettings** シート
 - DU に適用する OCR には TRUE を設定します。複数の OCR が適用可能です (UiPath Extended Languages OCR は常に利用されます)
- タクソノミーを定義する
 - リボン > デザイン > タクソノミーマネージャーから抽出するフィールド定義を設定します。
[\(タクソノミーマネージャー\)](#)
- DU_GetExtractionResult.xaml** をビルドする
 - 分類器を使用する場合 "Classification + Extraction" を有効にする
そうでない場合は "Extraction Only" を有効にする
 - 使用しない分類器、抽出器のアクティビティを削除する

(補足) DU 評価テンプレートには、Clova OCR と Microsoft Azure Computer Vision OCR (MS Read OCR) に対応した定義が含まれます

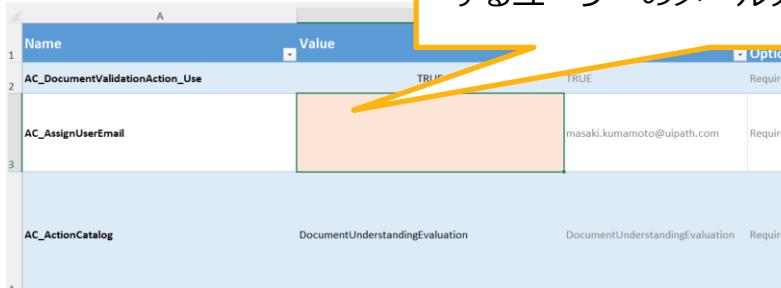
開発ステップ ～ Config.xlsx の補足説明

DU_ApiKey へ DU を利用するための API Key を指定します



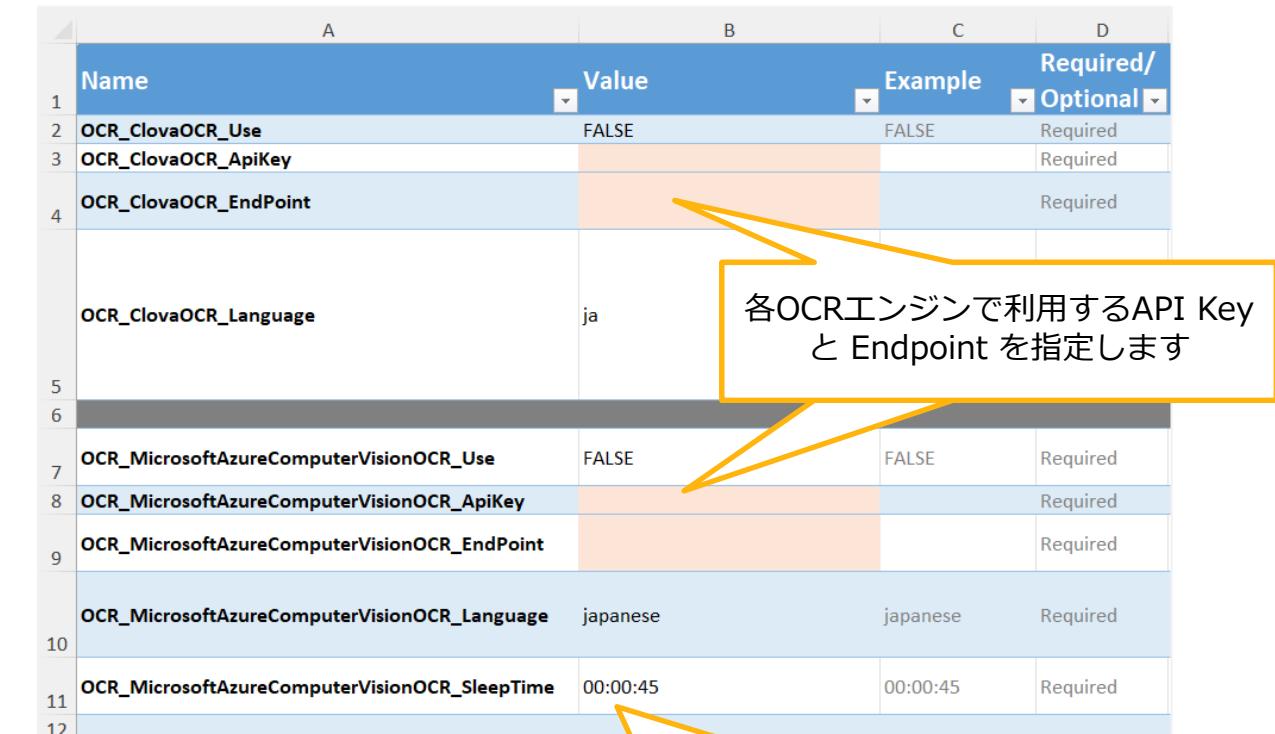
Name	Value	Example	Required/Optional
DU_ApiKey	xxxxxxxxxxxxxx		Required
DU_DocumentTypeId	UiPath.Sample.Invoices	Finance.Account.Invoice	Optional

分類器を利用しない場合、
DU_DocumentTypeId へドキュメントタイプIDを指定します



Name	Value	Example	Required/Optional
AC_DocumentValidationAction_Use	TRUE	TRUE	Required
AC_AssignUserEmail	masaki.kumamoto@uipath.com		Required
AC_ActionCatalog	DocumentUnderstandingEvaluation	DocumentUnderstandingEvaluation	Required

各OCRエンジンで利用するAPI Key と Endpoint を指定します

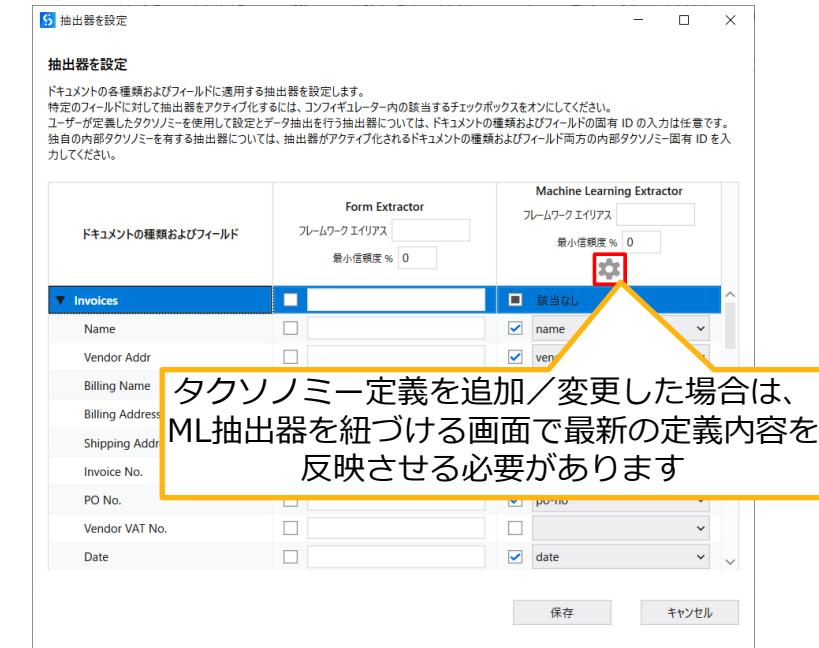
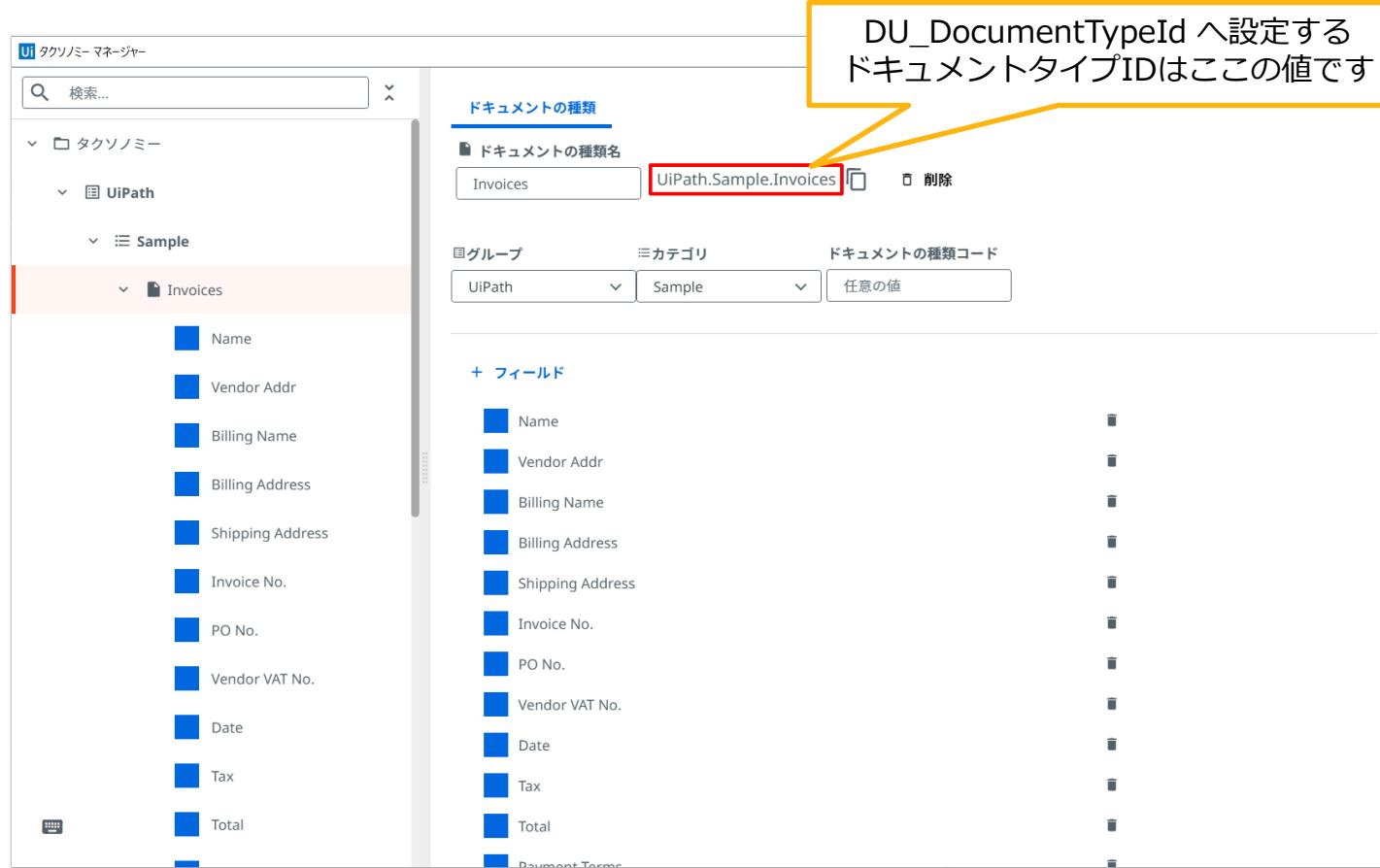


Name	Value	Example	Required/Optional
OCR_ClovaOCR_Use	FALSE	FALSE	Required
OCR_ClovaOCR_ApiKey			Required
OCR_ClovaOCR_EndPoint			Required
OCR_ClovaOCR_Language	ja		
OCR_MicrosoftAzureComputerVisionOCR_Use	FALSE	FALSE	Required
OCR_MicrosoftAzureComputerVisionOCR_ApiKey			Required
OCR_MicrosoftAzureComputerVisionOCR_EndPoint			Required
OCR_MicrosoftAzureComputerVisionOCR_Language	japanese	japanese	Required
OCR_MicrosoftAzureComputerVisionOCR_SleepTime	00:00:45	00:00:45	Required

MS Read OCR 無償環境の場合は API 実行数に制限が設定されているため、
APIを実行後にスリープする時間(秒)を指定します
(制限がない環境では 0 を設定してください)

開発ステップ ～タクソノミー定義の補足説明

事前に幾つかのドキュメントタイプに対応したタクソノミー定義が用意されておりますが、必要に応じて定義を変更する、もしくは定義を追加してください。



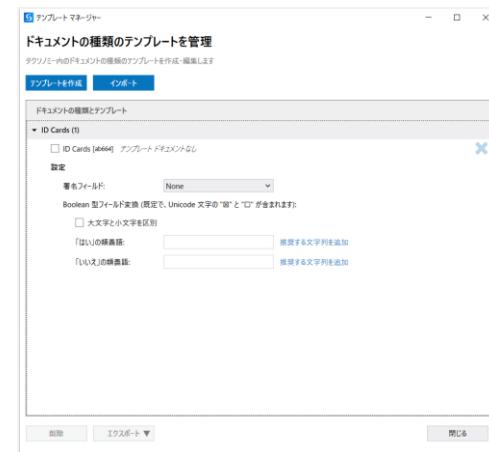
開発ステップ ～ DU_GetExtractionResult.xaml の補足説明

分類器を使用する “Classification + Extraction” では、サンプルデータに関する設定が用意されているため、そのまま動作させることができます。

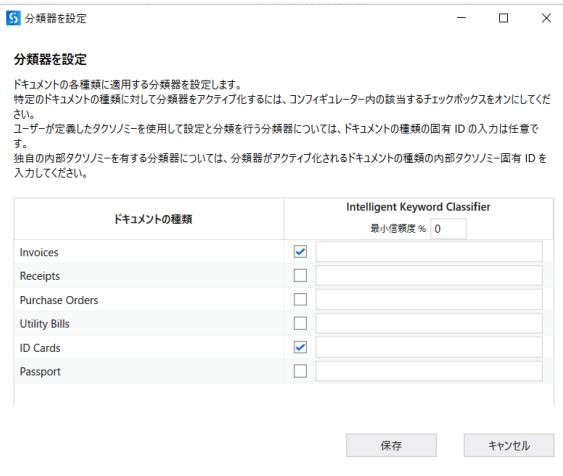
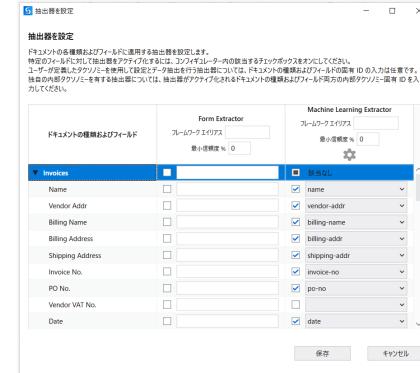
分類器の設定



フォーム抽出器のテンプレート定義

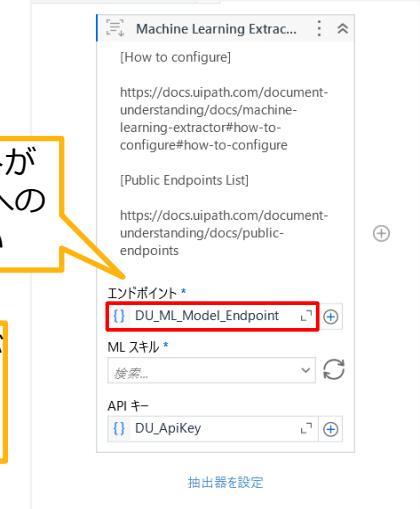


抽出器の設定



ML抽出器には請求書モデルへのパブリックエンドポイントが設定されていますが、必要に応じて個別に用意したモデルへのエンドポイント、もしくはMLスキルを指定してください

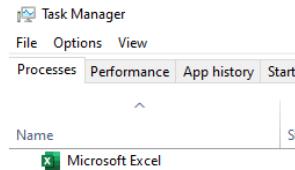
分類器を使用しない “Extraction Only” も同様にサンプルデータに関する設定が用意されていますが、分類器がないため、実行する前に使用しない抽出器のアクティビティや不要なサンプルデータを削除する必要があります



実行ステップ[®]

1. 01_ExtractDocumentsData.xaml を実行する

- 実行中にOneDriveの同期機能を停止すると、エラーが発生する可能性があります。
- Microsoft Excel がバックグラウンドでも実行されていないことを確認してください。



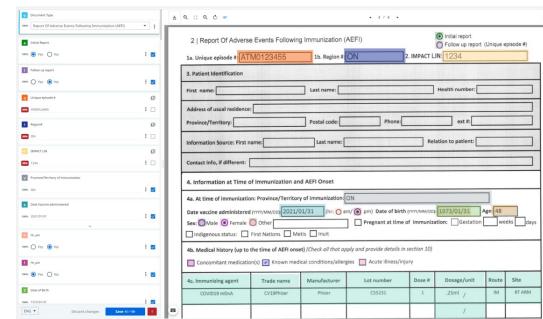
- 各ドキュメントの処理には約1~2分かかります。([ファイルをデバッグ] だともっと時間がかかるため、[ファイルを実行] をお勧めします)
- 実行が完了すると、Config.xlsx の各OCRセットと Action Center のドキュメント検証アクションの Excel レポートが生成されます。



01_ExtractDocumentsData.xaml

2. Action Center のドキュメント検証アクションタスクを完了する

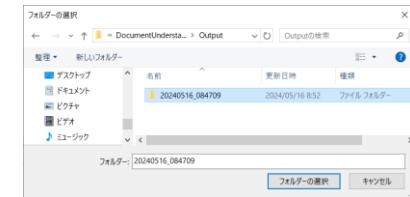
- ここで設定した内容がOCR読み取り処理後の期待される正しいデータとして後続の [02_CopyActualValuesToReport.xaml](#) で利用されるため、厳しくチェックします



UiPath™ Action Center

3. 02_CopyActualValuesToReport.xaml を実行する

- 実行直後に、ユーザーにDU評価レポートが保存されているフォルダを選択するダイアログが出力されますのでご注意ください。



- 実行完了後、ドキュメント検証アクションの結果は、各OCRの [ActionList.xlsx](#) と [DU Evaluation Reports](#) に貼り付けられます。
- ステップ3が完了した時点で、未だドキュメント検証アクションで検証されていない文書がある場合は、ドキュメント検証アクションで検証を完了させてから、再度ステップ3を実行し、DU Evaluation Reports を完成させます。



02_CopyActualValuesToReport.xaml

改善ステップ[®]

既存の ActionList.xlsx を使用して、DU ロジックを改善

過去に「実行ステップ」を実行し、同じタクソノミーを使用した同じドキュメントのリストに対して [ActionList.xlsx](#) を生成したことがある場合は、次回から以下の手順でステップ2と3をスキップすることができます。また、ドキュメント検証アクションの作成を無効にすることで、プロセスをより速く実行することができます。

この機能は、前回の実行結果レポートの精度に基づいてワークフローを修正し、DU の分類／抽出ロジックを改善する際に活用いただけます。

• Config.xlsx 設定

- “BasicSettings” シート
 - AL_UseExistingActionListExcel (= TRUE)
 - AL_ExistingActionListExcelPath
- “ActionSettings” シート
 - AC_DocumentValidationAction_Use (= False)

Name	Value
AL_UseExistingActionListExcel	TRUE
AL_ExistingActionListExcelPath	Output/20210601/ActionList.xlsx
AC_DocumentValidationAction_Use	False



Config.xlsx

(BasicSettings, ActionSettings)

1 01_ExtractDocumentsData.xaml を実行する



01_ExtractDocumentsData.xaml

2 Action Center のドキュメント検証アクションタスクを完了する



SKIP

UiPath Action Center

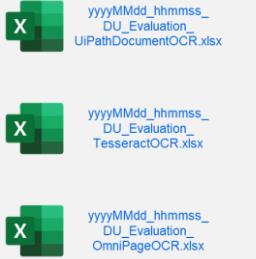
3 02_CopyActualValuesToReport.xaml を実行する



02_CopyActualValuesToReport.xaml

結果

レポートの作成



yyyyMMdd_hhmmss_DU_Evaluation_UiPathDocumentOCR.xlsx

yyyyMMdd_hhmmss_DU_Evaluation_TesseractOCR.xlsx

yyyyMMdd_hhmmss_DU_Evaluation_OmniPageOCR.xlsx

Document Understanding Evaluation Template

4

その他の便利な機能

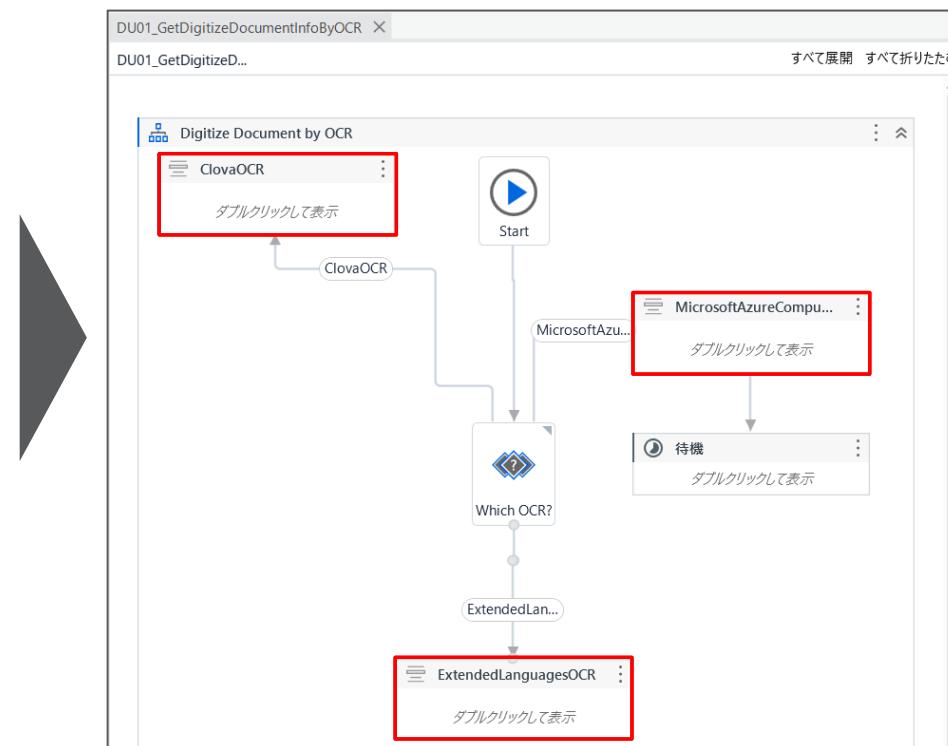
複数OCRエンジンの実行

Config.xlsxで特定のOCRを有効にすることで、DUに複数のOCRを適用し、各OCRの評価レポートを同時に作成することができます。(UiPath Extended Languages OCRは常に利用されます)

Name	Value
OCR_ClovaOCR_Use	TRUE
OCR_ClovaOCR_EndPoint	https://...
OCR_MicrosoftAzureComputerVisionOCR_Use	TRUE
OCR_MicrosoftAzureComputerVisionOCR_EndPoint	https://...



Config.xlsx
(OcrSettings)



yyyyMMdd_hhmmss_
DU_Evaluation_
ExtendedLanguagesOCR.xlsx



yyyyMMdd_hhmmss_
DU_Evaluation_
ClovaOCR.xlsx



yyyyMMdd_hhmmss_
DU_Evaluation_
MicrosoftAzureComputerVisionOCR.xlsx

Clova OCRとMS Read OCRが
3rd Party OCRエンジンとして利用可能です

信頼度閾値による自動検証

自動検証の閾値として、ConfidenceとOcrConfidenceを使用することができます。
これらの両方が閾値以上であれば、標準フィールドが自動的に検証されます。

Confidence = 抽出器の信頼度
OcrConfidence = OCRの信頼度

Name	Value
DU_AutoVerifyMinimumThreshold_Confidence	99.98%
DU_AutoVerifyMinimumThreshold_OcrConfidence	95.99%



Config.xlsx
(DuSettings)



UiPath™ Action Center

検証アクションにOcrConfidenceを表示する

(現在のACでは信頼度とOCR信頼度の切替が可能なため、この機能は不要となりました)

Action Centerに表示される値として、Confidence または OcrConfidence が選択できます。
扱う文書によって異なりますので適切な方を選択してください。

Confidence = 抽出器の信頼度
OcrConfidence = OCRの信頼度

Name	Value
DU_ValidationConfidenceType	OcrConfidence



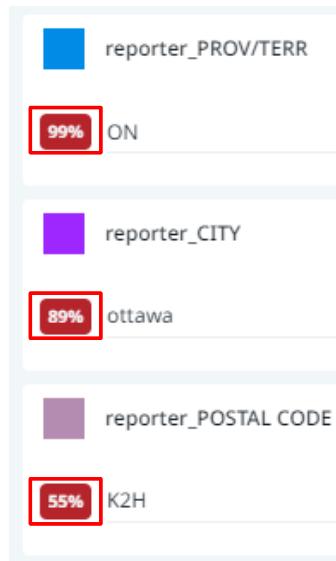
Config.xlsx
(DuSettings)



Confidence



OcrConfidence



UiPath Action Center

既存ActionList.xlsxの内容を実際の値として利用する

(29ページと同様の問題が発生した場合は、ActionList.xlsxへ空行を追加することで対応が可能です)

過去に「実行ステップ」を実行し、同じタクソノミーを使用した同じドキュメントのリストに対して ActionList.xlsx を生成した場合、次回からアクションセンターでのステップ検証と 02_CopyActualValuesToReport.xaml の実行を省略することができます。

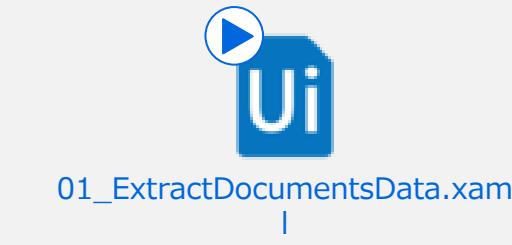
Name	Value
AL_UseExistingActionListExcel	TRUE
AL_ExistingActionListExcelPath	Output/20210601/ActionList.xlsx



Config.xlsx
(BasicSettings)



- 1 01_ExtractDocumentsData.xaml を実行する。



- 2 アクションセンターのドキュメント検証アクションタスクを完了する

- 2 アクションセンターのドキュメント検証アクションタスクを完了する



- 3 02_CopyActualValuesToReport.xaml を実行する。



02_CopyActualValuesToReport.xaml

結果

レポートの作成



Document Understanding Evaluation Template

5

制限事項

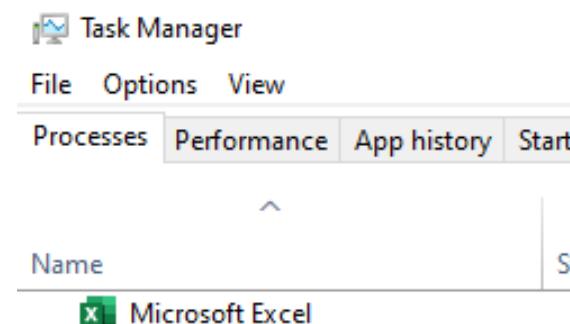
既知の問題

制限事項

- 1つのフィールドに複数の抽出結果候補がある場合、ConfidenceとOcrConfidenceが最も高い結果がレポートに記載されます。

既知の問題

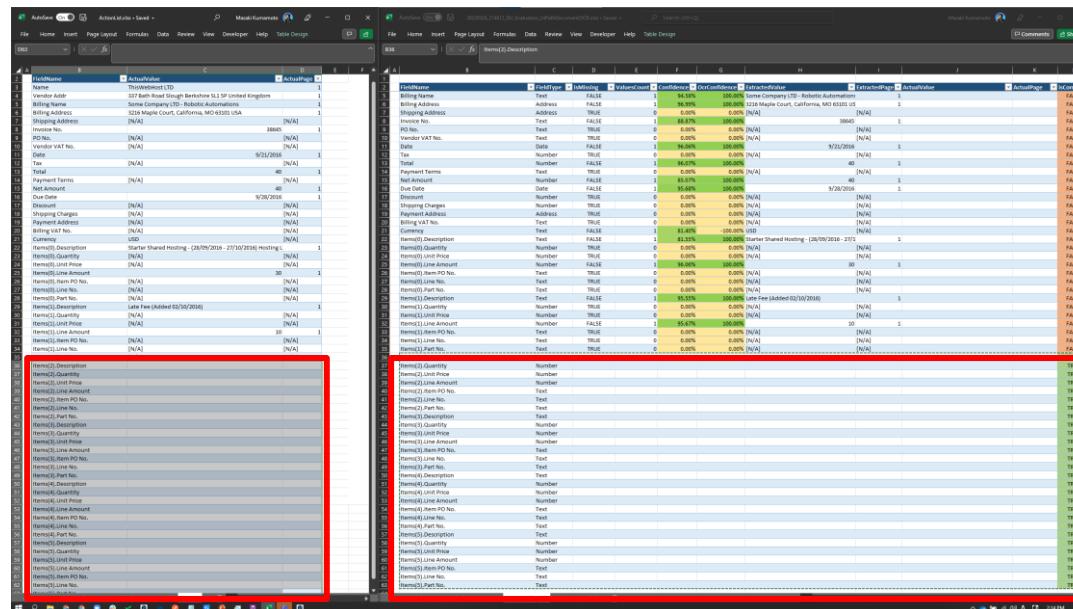
- ワークフロー実行時にバックグラウンドでも「Microsoft Excel」が起動していると何らかの原因でワークフローでエラーが発生してしまう。



制限事項と既知の問題

既知の問題

- ML 抽出器を利用する環境でタクソノミーにテーブルフィールドがある場合、Action Centerから検証値を取得する際にエラーが発生する場合があります。これは抽出された行数がドキュメント内の実際の行数よりも少ない場合、ワークフローがフィールドの動的な数を処理できないためです。
ワークフローを実行する前に、適切なフィールド名と空の値を持つ行を手動で追加することで、この問題を回避することができます。



Document Understanding Evaluation Template

6

リリースノート

リリースノート

パッケージはこちら⇒[パッケージ](#)、リリースはこちら⇒[リリース](#)

日付	バージョン	説明	リポジトリリンク
2021年6月3日	0.0.1	<ul style="list-style-type: none">初期プロジェクト	リンク
2021年6月8日	0.0.3	<ul style="list-style-type: none">UiPathパブリックエンドポイントのMLモデル（請求書、発注書、領収書、公共料金請求書）にデフォルトのタクソノミーを追加。	リンク
2021年6月8日	0.0.4	<ul style="list-style-type: none">バグ修正：抽出された値が0の場合、Excelレポートの "isCorrect" カラムが "Correct" と表示され、"Actual" カラムに空の値が表示される不具合を修正しました。	リンク
2021年6月9日	0.0.7	<ul style="list-style-type: none">障害者の活動を閉じるDU_Evaluation.xlsxのSummaryシートから "Table" と "TableColumn" を削除する。エクセルシートの2行目を凍結し、ヘッダーが常に上に来るようになる。IDカードとパスポートの分類法を追加	リンク
2021年7月16日	0.2.1	<ul style="list-style-type: none">自動塗りつぶしのバグ修正(0.0.7から)バグを修正：保留中のアクションがある場合、02_Processでエラーが発生する。サマリーシートの全体的な結果がずれる問題を修正。(0.0.7からの問題)障害者の活動を閉じるDU_Evaluation.xlsxのSummaryシートから "Table" と "TableColumn" を削除する。エクセルシートの2行目を凍結し、ヘッダーが常に上に来るようになる。IDカードとパスポートの分類法を追加AL_UseExistingActionListExcelがTrueの場合の動作を修正しました。ExtractedPageとActualPageをレポートに含める。バケツフォルダ名にタイムスタンプを含めるAL_UseExistingActionListExcel が True の場合、新しい Excel ファイルを作成しない。バグフィックス：エクセルレポートの文書画像の並び順が、10ページ以上の場合、正しくない。	リンク
2022年3月24日	0.2.3	<ul style="list-style-type: none">安定性のために「ディレイ」アクティビティを追加パッケージの更新タスクデータ取得の "Orchestrator Folder" プロパティに "OC_FolderPath" を追加しました。	リンク
2024年5月16日	0.3.0	<ul style="list-style-type: none">日本のお客様向けに情報を整理	リンク

