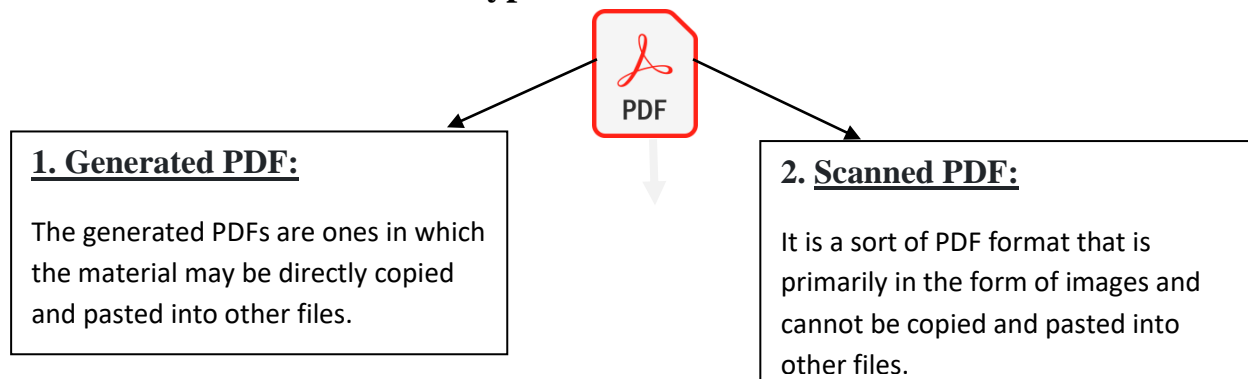


PDF AUTOMATION USING RPA UIPATH

A Portable Document Format (PDF) is a file format for collecting and transmitting electronic documents in their original format. Whereas PDF files play a key role in everyday tasks and own essential points of procedures in every sector.

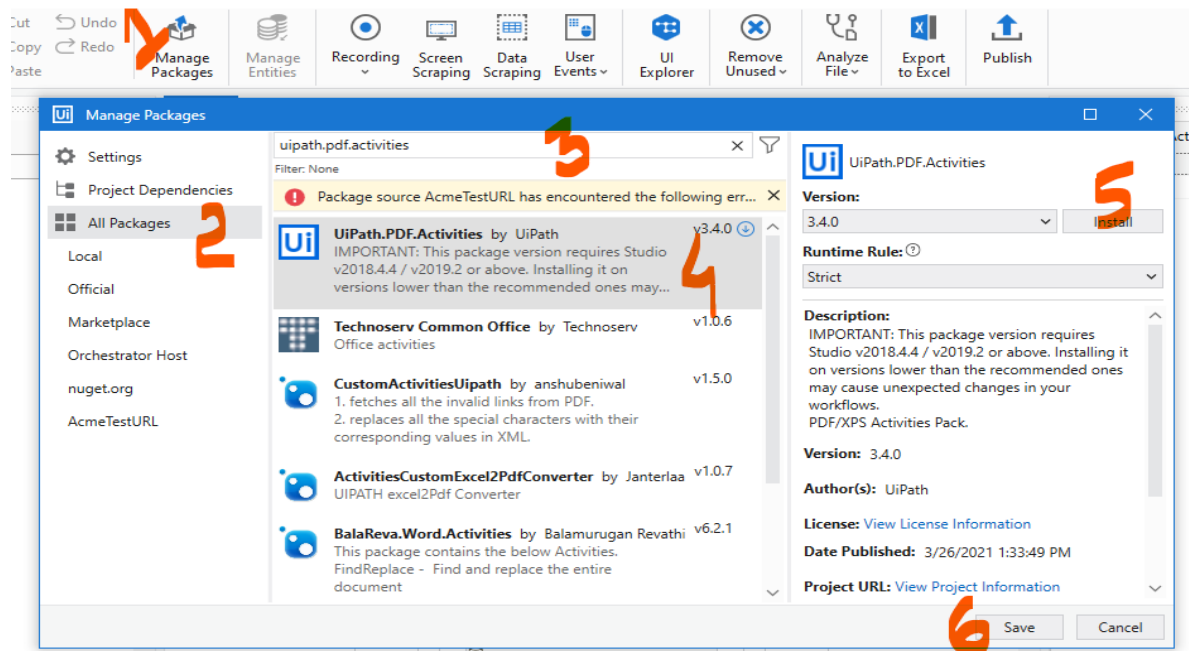
Creating or reading such files is an important component of PDF operations in a variety of market sectors, and it is simple to automate with UiPath.

Two types of PDF Automation:

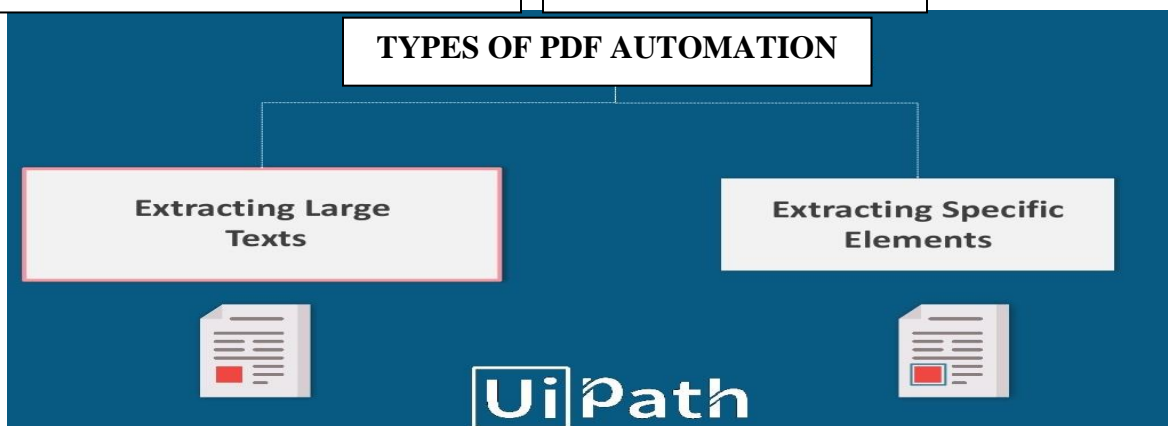
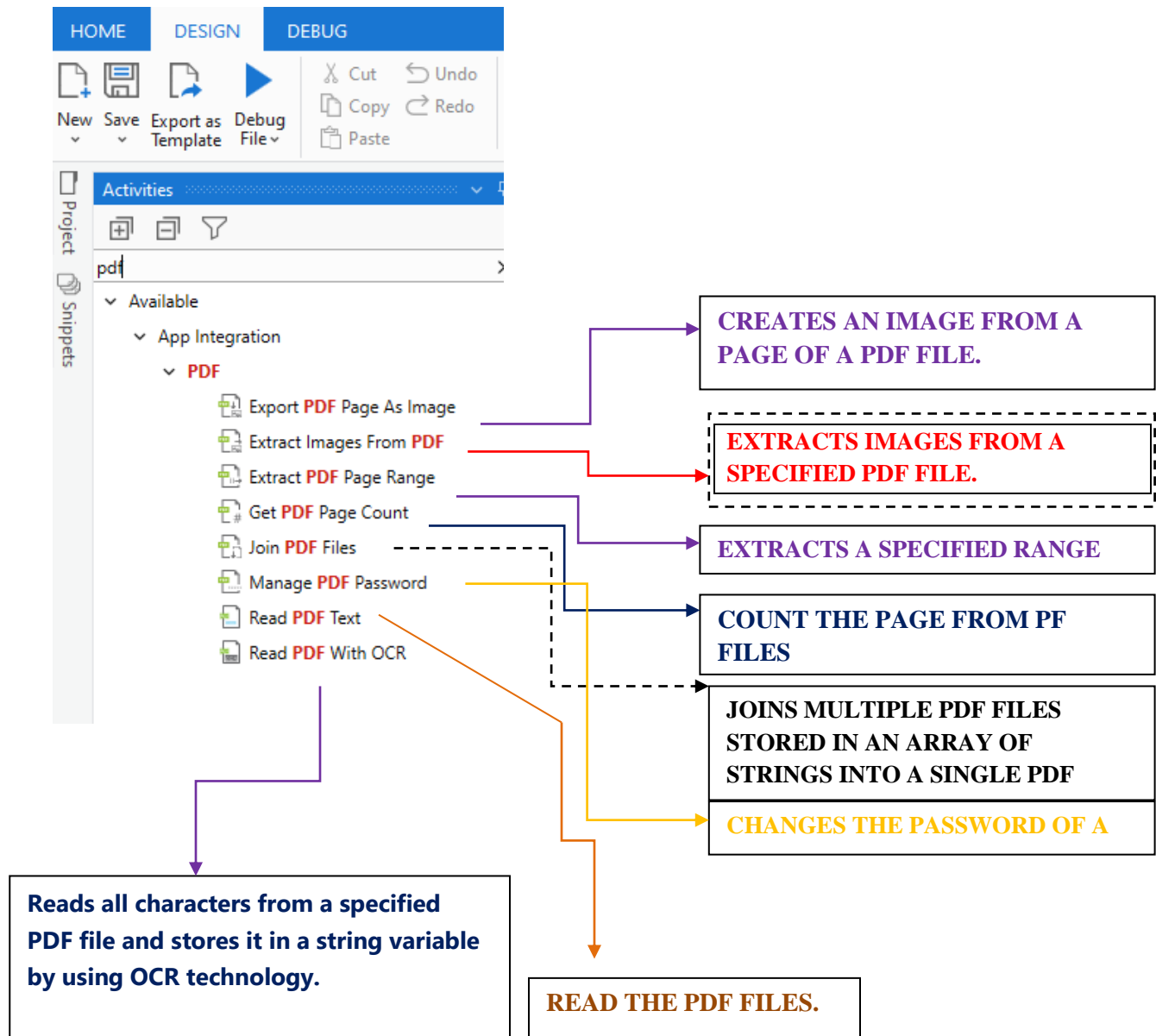


INSTALL THE UIPATH PDF ACTIVITIES PACKAGE:

After you've launched the Process in UiPath Studio, you'll need to install the PDF packages. Go to **Manage packages**, then choose **All packages**, then **UiPath.PDF.Activities**, and install it then save the Packages.



WHAT ARE THE ACTIVITIES AVAILABLE IN UIPATH STUDIO?



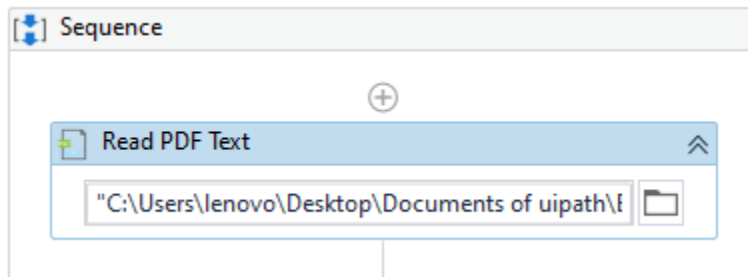
1. EXTRACTING LARGE TEXT:

There may be times when we have a document that is entirely composed of text or a combination of text and graphics. Extraction of big texts, on the other hand, is applicable to documents that include simply text or a combination of text and graphics.

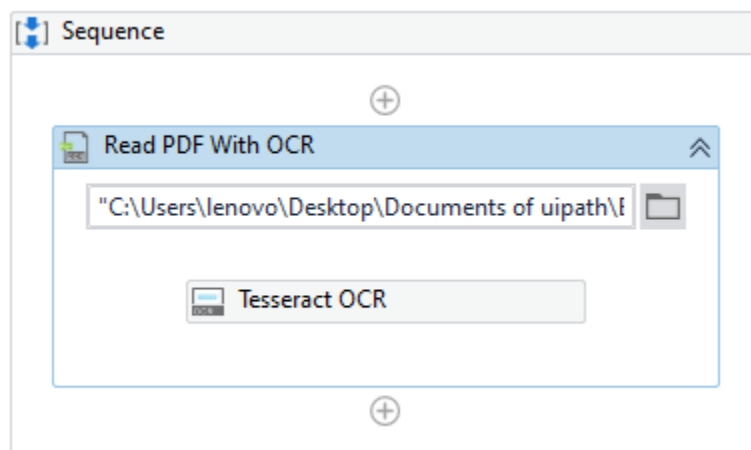
UiPath provides primarily two ways for extracting huge amounts of text. These are the activities:

Activity:

Read PDF Text: The Read PDF activity is used to extract data from the PDF files which have *Text only*.



OCR Activity for Reading PDF: The Read PDF with OCR Activity is used to extract data from the PDF documents which have both *Text and Images*.



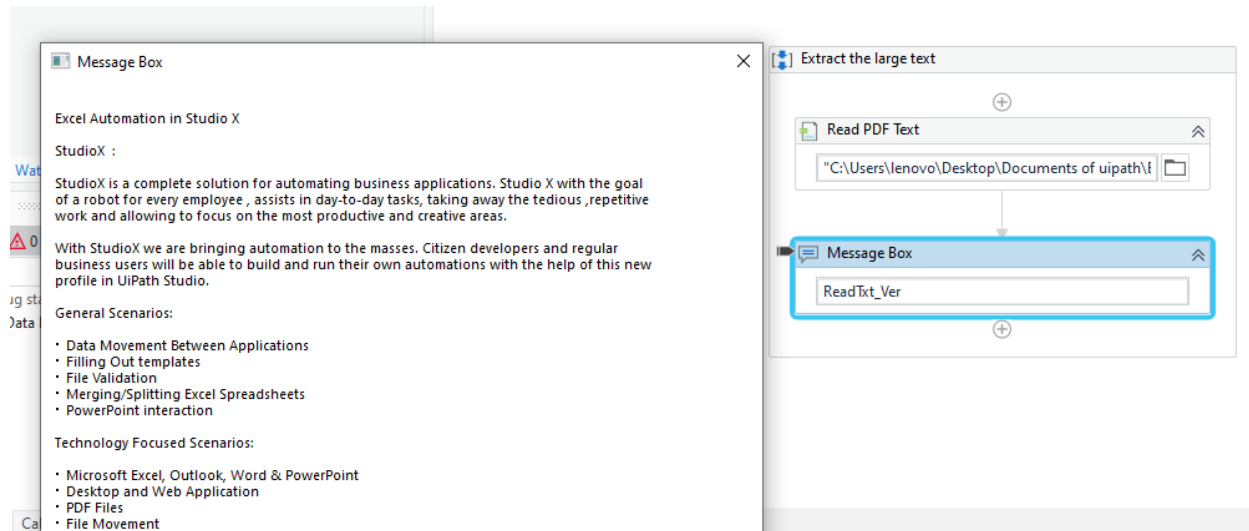
PROCESS:

Step 1: Make a Sequence and rename it if necessary. I've renamed it Extract Large Text in this case.

Step 2: Drag and drop the Read PDF Text Activity into the canvas. Mention the path of the PDF document from which data must be extracted in the activity.

Step 3: To view the output, mention an output variable in the Properties Pane of the Read PDF Text Activity. To create an output variable, press CTRL + K and name it. As output, I've noted it here.

Step 4: Then, in the sequence, drag and drop a message box and mention the output variable in it.



Extracting Dynamic Value:

- **Anchor Base.**
- **Find element.**
- **Find Image.**

Anchor Base Activity: Text and pictures are extracted using the Anchor Base Activity.

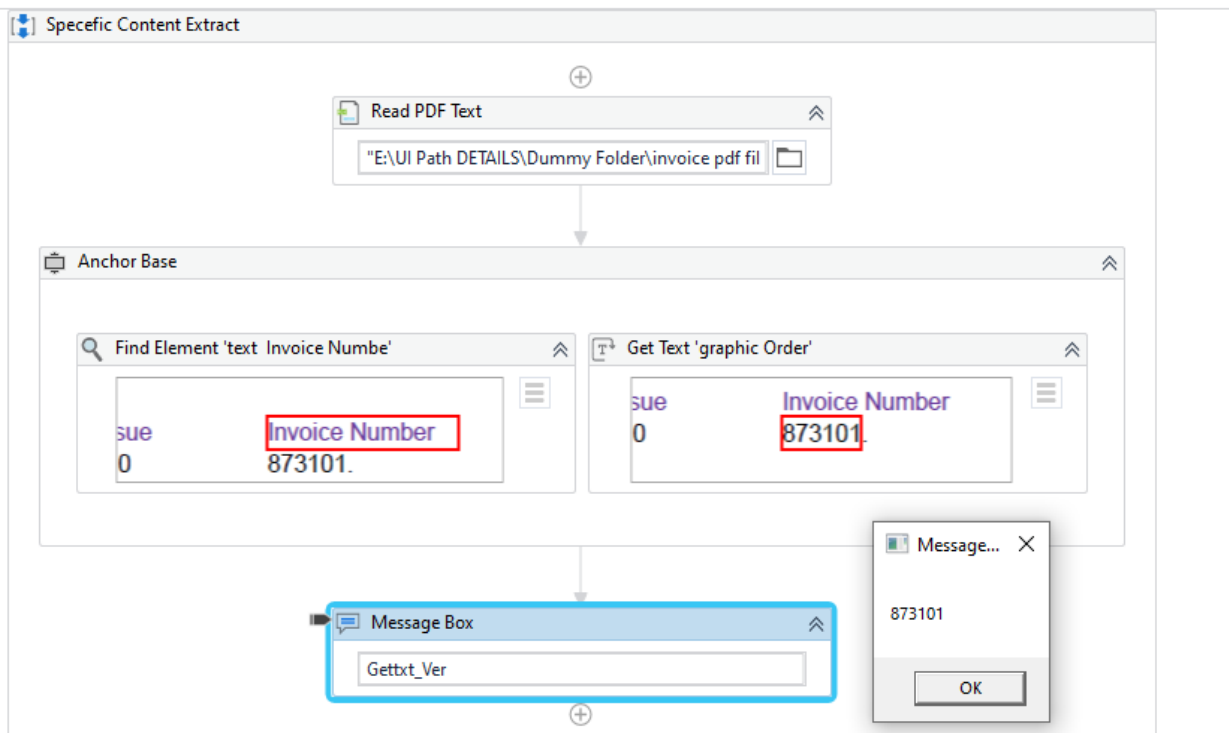
This activity consists of two actions since it executes one in connection to another fixed element or anchor.

As a result, a typical anchor-base activity often includes two activities that are employed beneath it:

- **Find Element/Find Image Activity:** The Find Element / Find Image Activity is used to locate an element, such as text or an image. You can deploy the activities as you see fit. Because the Anchor Base activity is a relative activity, you may perform the Get Text Activity, as previously indicated.
- **Get Text Activity:** This action simply indicates the element you want to remove. Text may be extracted with this activity, and an output variable can be utilized. Following that, you can use a Message Box or a Write Text File Activity to specify the output variable.

➤ Extract a single piece of information from a pdf document

Task: We need to extract the data i.e., Invoice Number Only from the Invoice Process PDF files.

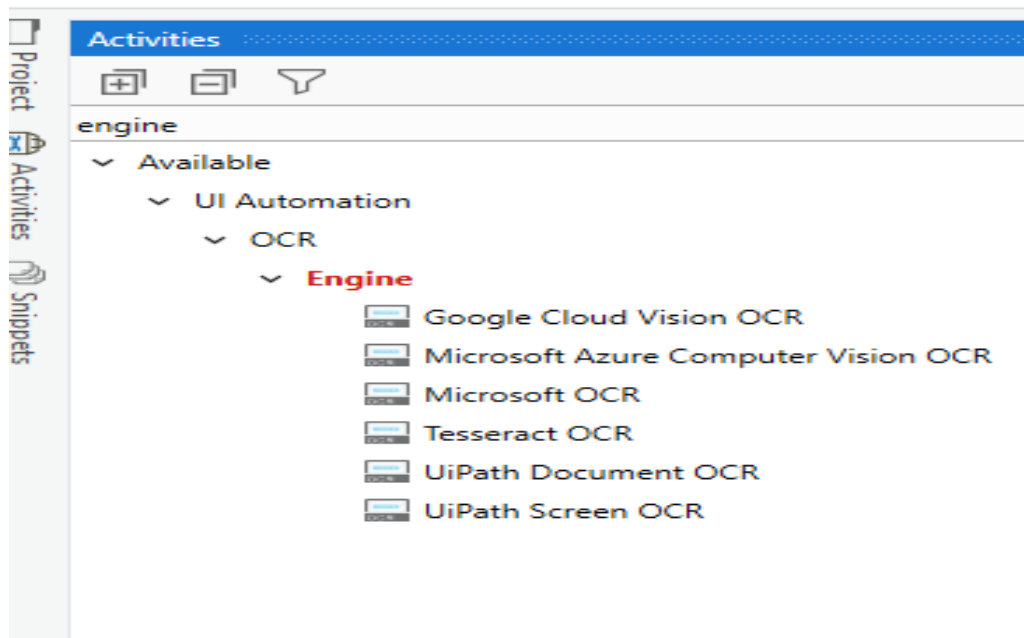


➤ **Extract large text segments from PDF using the Read PDF activity, the *Read PDF with OCR activity Method:***

Under the OCR activity: Optical Character Recognition

To Read PDF with OCR Activity is used to extract data from the PDF documents which have both **Text and Images**. So, if you have any images apart from the text in the document, this activity would extract data from those images and give a Text output.

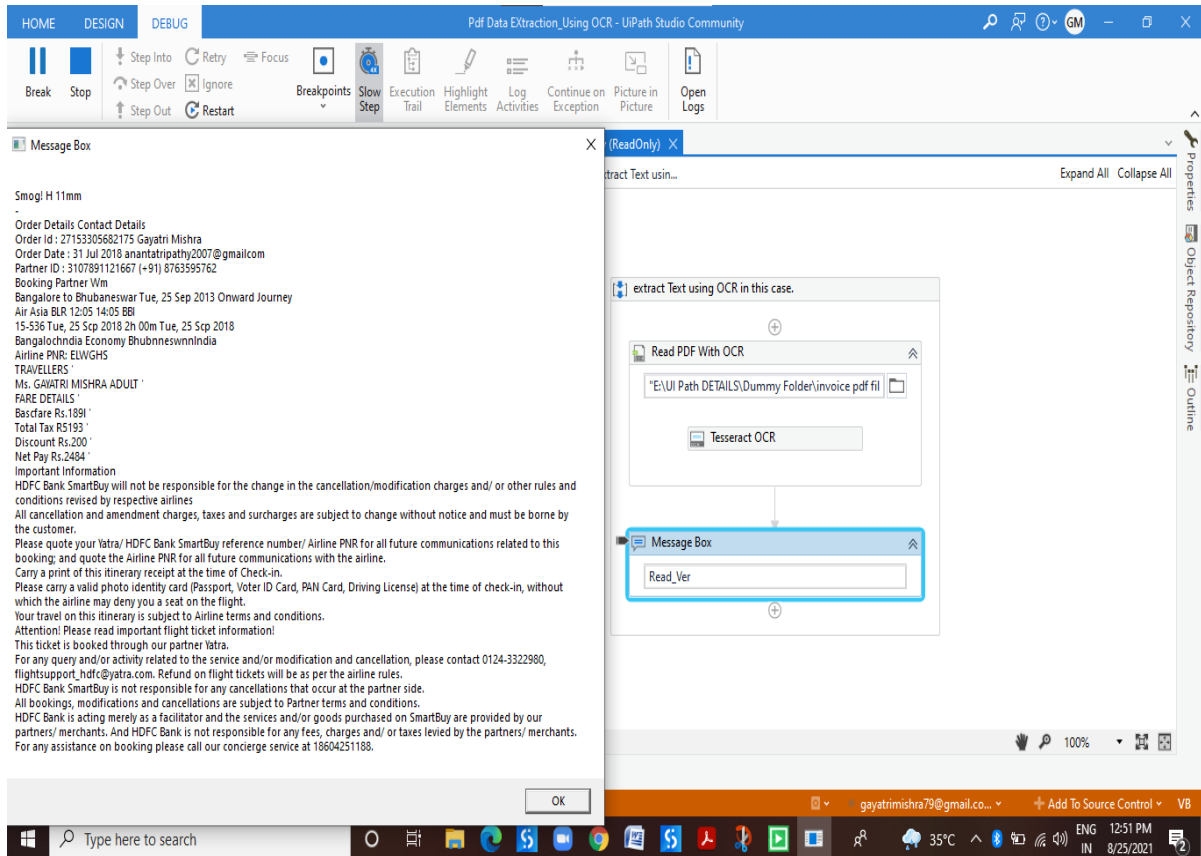
These activities used for extract text with image.



Follow the below steps, to create automation for extracting text present inside images.

The steps are as follows

1. Make a Sequence and rename it if necessary. I renamed it to extract Text using OCR in this case.
2. Drag and drop the Read PDF with OCR Activity into the screen. Mention the path of the PDF document from which data must be extracted in the activity.
3. Now, look for an OCR Engine and drag and drop one based on which is installed. I utilized Teesseract OCR Engine in this case.
4. To see the output, mention an output variable in the Properties Pane of the Read PDF with OCR Activity. To configure an output variable Give it a name by pressing CTRL + K. As output, I've noted it here.
5. Then, in the sequence, drag and drop a message box and mention the output variable in it.



Extract large text segments from PDF using the Read PDF activity and the Screen Scraping wizard:

Task: We need to extract the data from Flight ticket

Step Followed:

1. Make a Sequence and rename it if necessary. I renamed it to extract Text using Screen scraping in this case.
2. Click on the Screen scraping
3. Now You look for the Attach window Activity and Get OCR text Activity.
4. To see the output, mention an output variable in the Properties Pane of the Get OCR text Activity. To configure an output variable Give it a name by pressing CTRL + K. As output, I've noted it here.
5. Drag and drop a message box and mention the output variable in it.

Screen Scraping

Expand All Collapse All

UiPath.Core.Activities.GetOCRText

Common

ContinueOnError Spec

DisplayName Get OCR Text

Input

Target Target

Misc

Private

Output

Text Graphic

WordsInfo TA

GraphicOrder

Attach Window 'Ticketpdf Acrobatsd'

Do

Get OCR Text 'graphic Order'

Microsoft OCR

Message Box

GraphicOrder

SmartAad
deal
Order Details
Olde Id
Olde Date
Bo o king Pa_ltnr
Bangalore to Bhubaneswar
Air Asia
Airline PNR: ELWGH5
Ms. GAYATRI Nil.SHRA
FARE DE -ram S HDFC BANK
: 27153305682175
: 31 Jul 2018
: 3107891121667
yatra
BLR 12:05
Tue. 2-5 sep 2018 14:05 BBI
Tue, 25 sep 2018 Contact Details
Gayatri Misl_na
mantabipathy2007@gmailcom
Onward Journev
ADULT

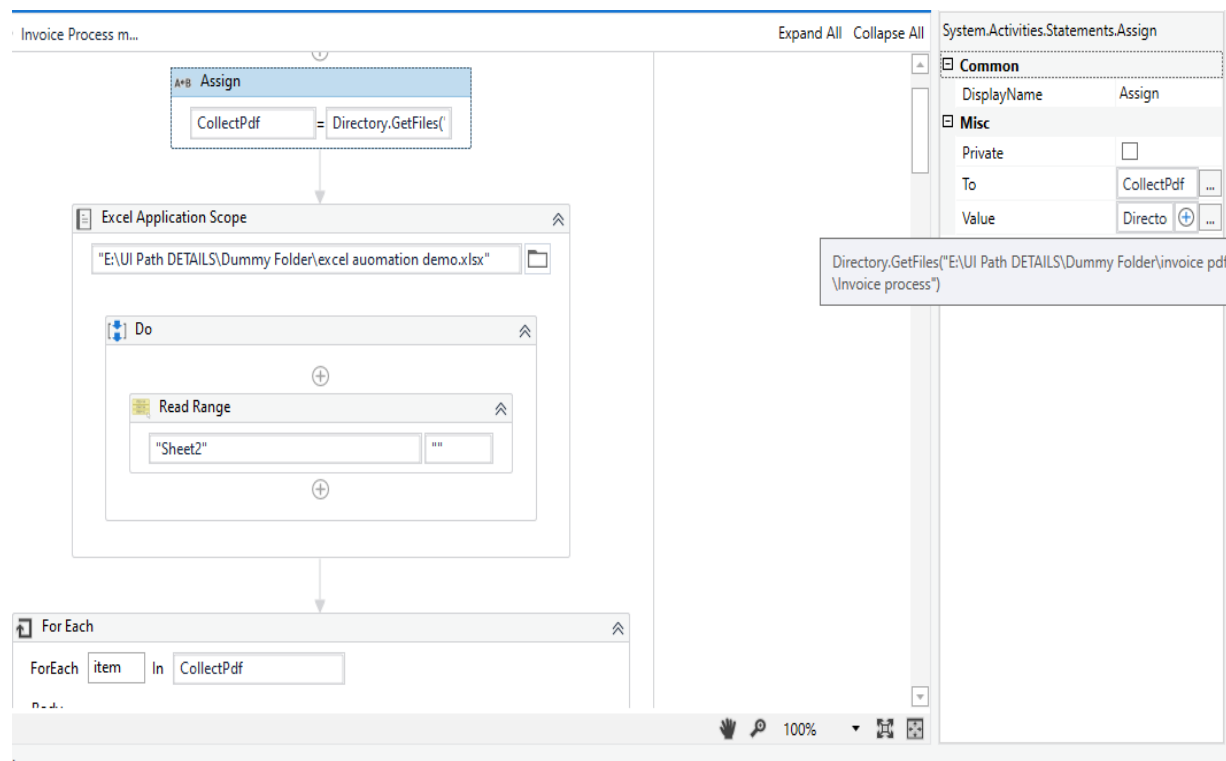
OK

100%

REAL TIME SCENARIO:

Extract Multiple Pdf Files:

- 1.Create array to store all the files using assign activity.
- 2.Excel application scope (For Put the path of the folder)
- I. Read Range activity (Read the excel sheet)



Variable panel:

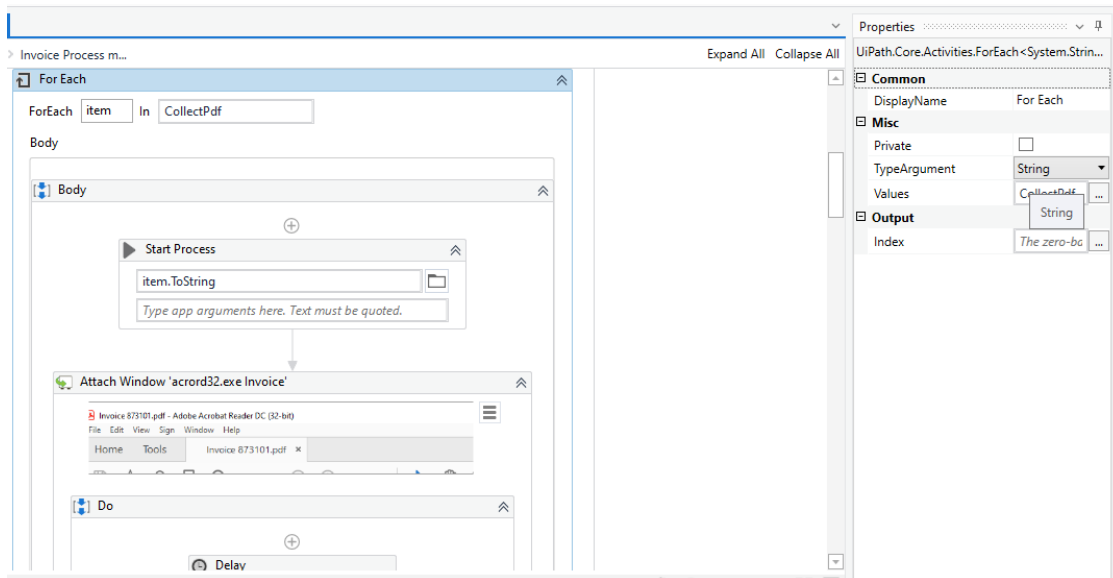
Name	Variable type	Scope	Default
SoldTo_Ver	String	Do	Enter a VB expression
OrderDate_Ver	String	Do	Enter a VB expression
OrderNumberVer	String	Do	Enter a VB expression
InvoiceNum_Ver	String	Do	Enter a VB expression
CollectPdf	String[]	Invoice Process m...	Enter a VB expression
ReadExcel_Ver	DataTable	Invoice Process m...	Enter a VB expression
TOTAL_Ver	String	Invoice Process m...	Enter a VB expression

Variables Arguments Imports

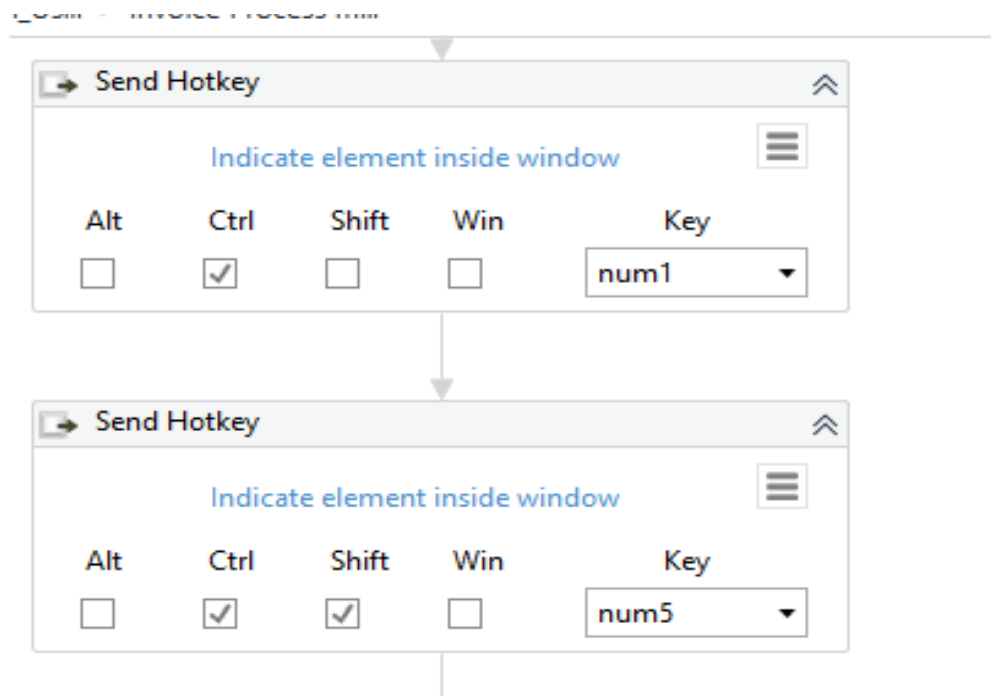
100%

4. For each loop (Open each File)

- Start process
- Attach window

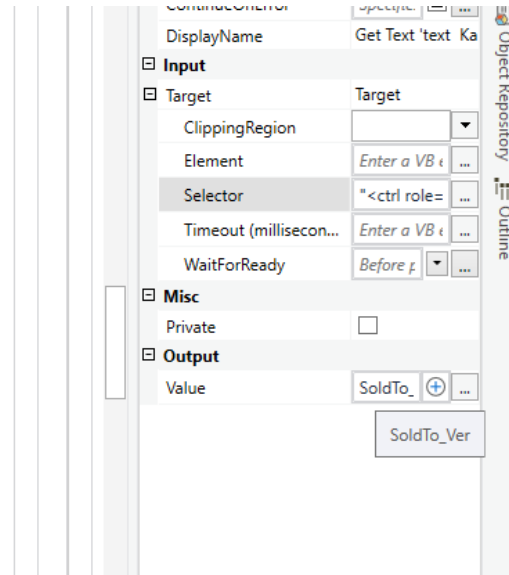
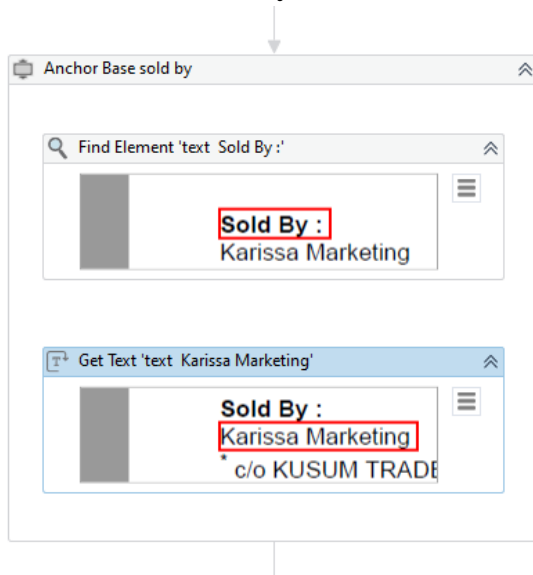


- Delay activity
- Send hot key



5.Anchor base activity

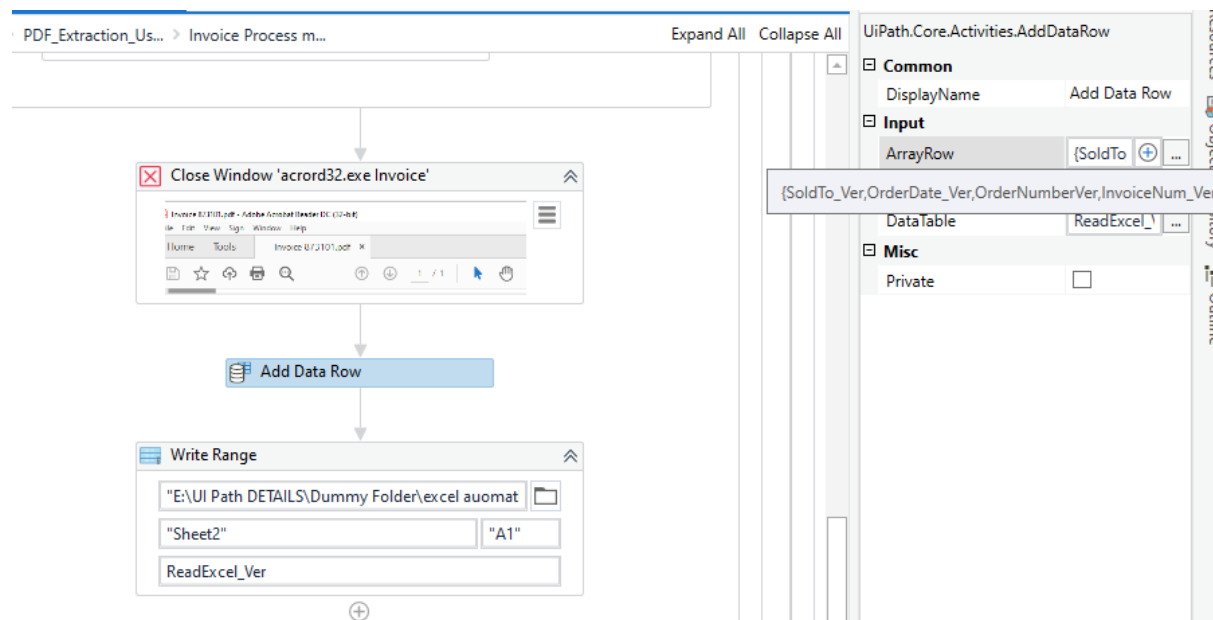
- Find Element
- Get text activity



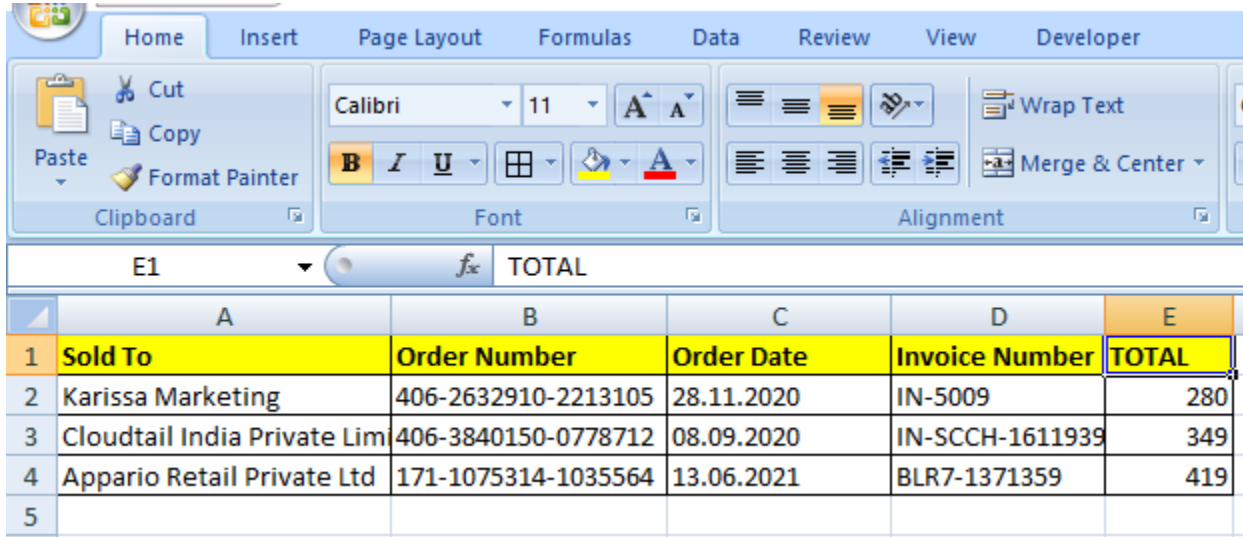
6.Close window

7.Add data row

8.Write Range Activity



OUTPUT:



The screenshot shows the Microsoft Excel interface with the 'Home' tab selected. The ribbon includes 'Clipboard', 'Font', and 'Alignment' groups. The active cell is E1, which contains the formula '=TOTAL'. The table below is displayed in the worksheet.

	A	B	C	D	E
1	Sold To	Order Number	Order Date	Invoice Number	TOTAL
2	Karissa Marketing	406-2632910-2213105	28.11.2020	IN-5009	280
3	Cloudtail India Private Lim	406-3840150-0778712	08.09.2020	IN-SCCH-1611939	349
4	Appario Retail Private Ltd	171-1075314-1035564	13.06.2021	BLR7-1371359	419
5					

Reference:

1. <https://www.youtube.com/watch?v=EEIzfs11hv8&list=PLhTE7-JU1rhapBUgDoJOqxEmBN0XxYrDE&index=5&t=2201s>
2. https://www.youtube.com/watch?v=vuSszbM_oHc&t=762s
3. <https://www.edureka.co/blog/ui-path-pdf-data-extraction/>

For any RPA Implementation/Resources in your Organization please reach out to rpa@gxplabs.com.

THANK YOU