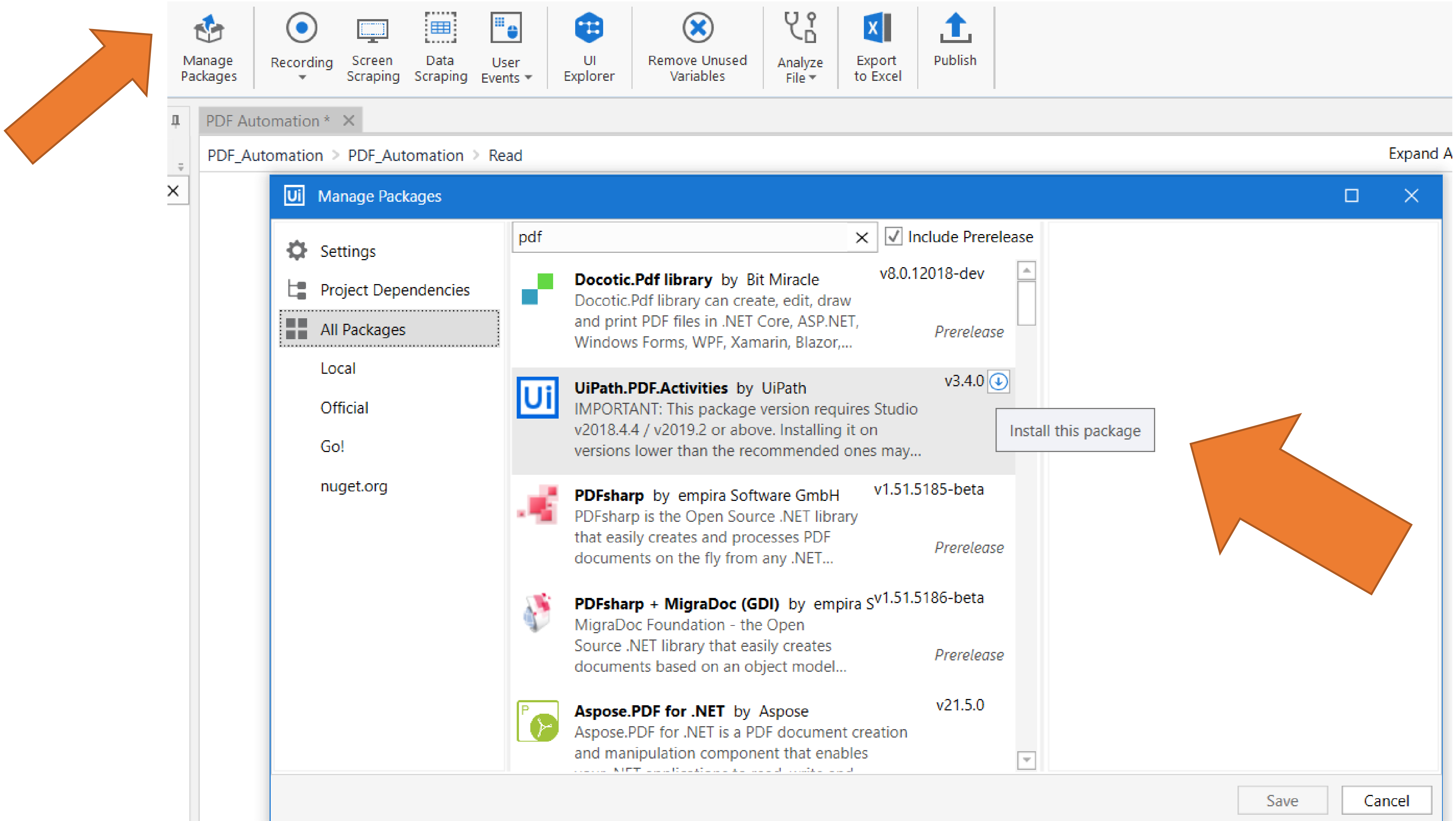
Practice Examples

PDF Automation

- Types of PDF Activities
    - Extracting Large Texts
        - Read PDF text – For files with TEXT only
        - Read PDF with OCR – For files with TEXT and IMAGES
    - Extracting Specific Elements

Extracting Large Text – Read PDF text



Docotic.Pdf library by Bit Miracle v8.0.12018-dev
Docotic.Pdf library can create, edit, draw and print PDF files in .NET Core, ASP.NET, Windows Forms, WPF, Xamarin, Blazor,...
Prerelease

UiPath.PDF.Activities by UiPath v3.4.0
IMPORTANT: This package version requires Studio v2018.4.4 / v2019.2 or above. Installing it on versions lower than the recommended ones may...

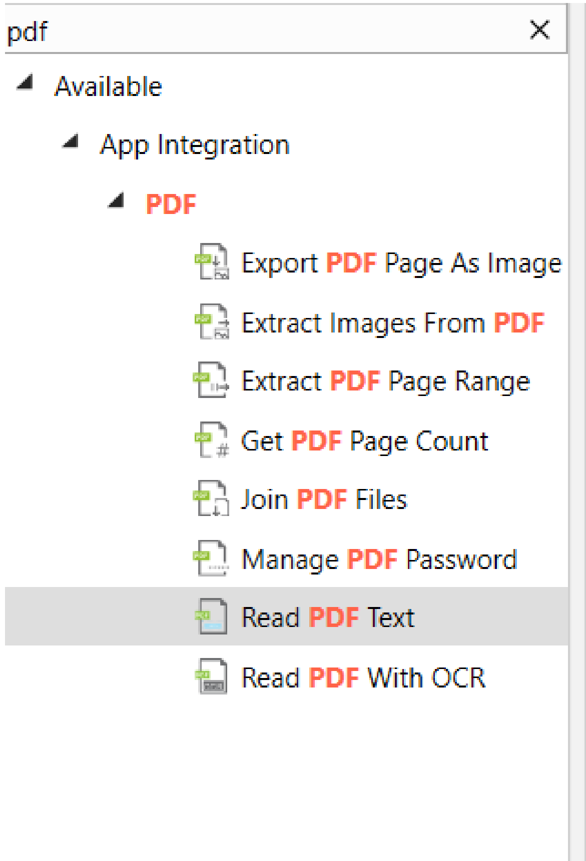
PDFsharp by empira Software GmbH v1.51.5185-beta
PDFsharp is the Open Source .NET library that easily creates and processes PDF documents on the fly from any .NET...
Prerelease

PDFsharp + MigraDoc (GDI) by empira S v1.51.5186-beta
MigraDoc Foundation - the Open Source .NET library that easily creates documents based on an object model...
Prerelease

Aspose.PDF for .NET by Aspose v21.5.0
Aspose.PDF for .NET is a PDF document creation and manipulation component that enables

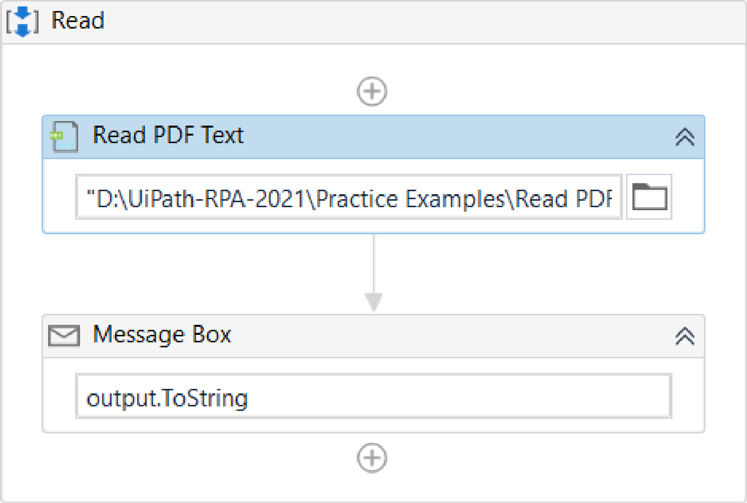
Read PDF Text

- Drag and drop Read PDF Text Activity
- Specify the path to access the pdf file
- Specify the output variable name
- Drag & drop Message box and display output variable



pdf

- Available
  - App Integration
    - PDF
      - Export PDF Page As Image
      - Extract Images From PDF
      - Extract PDF Page Range
      - Get PDF Page Count
      - Join PDF Files
      - Manage PDF Password
      - Read PDF Text
      - Read PDF With OCR



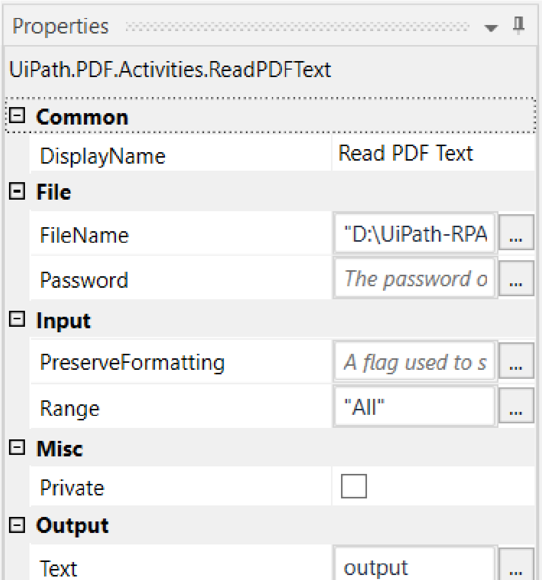
Read

Read PDF Text

"D:\UiPath-RPA-2021\Practice Examples\Read PDF

Message Box

output.ToString



Properties

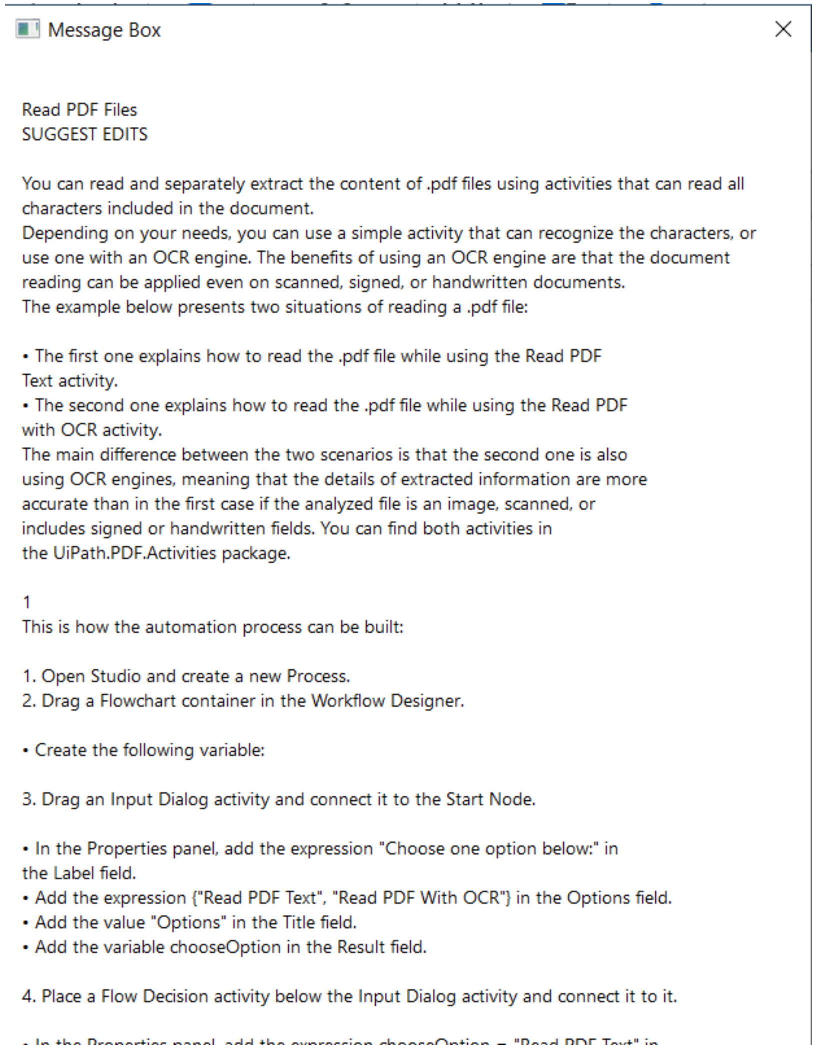
UiPath.PDF.Activities.ReadPDFText

Common	
DisplayName	Read PDF Text
File	
FileName	"D:\UiPath-RPA
Password	*The password o*
Input	
PreserveFormatting	*A flag used to s*
Range	"All"
Misc	
Private	□
Output	
Text	output

Hit Run to check the result

- Output will read all the contents (page1,2)
- Observer that First page has only TEXT part
- The last line in the first page is missing.
- The reason is that last line is of type image.
- To access image content we should use
- Read PDF with OCR Activity
- Note: Read PDF Text activity has property

Called **Range** default its “All”, change it to Page numbers. Example: 1, 2-5, 3-6 etc.



Read PDF Files
SUGGEST EDITS

You can read and separately extract the content of .pdf files using activities that can read all characters included in the document.
Depending on your needs, you can use a simple activity that can recognize the characters, or use one with an OCR engine. The benefits of using an OCR engine are that the document reading can be applied even on scanned, signed, or handwritten documents.
The example below presents two situations of reading a .pdf file:

• The first one explains how to read the .pdf file while using the Read PDF Text activity.
• The second one explains how to read the .pdf file while using the Read PDF with OCR activity.
The main difference between the two scenarios is that the second one is also using OCR engines, meaning that the details of extracted information are more accurate than in the first case if the analyzed file is an image, scanned, or includes signed or handwritten fields. You can find both activities in the UiPath.PDF.Activities package.

1
This is how the automation process can be built:

1. Open Studio and create a new Process.
2. Drag a Flowchart container in the Workflow Designer.

• Create the following variable:

3. Drag an Input Dialog activity and connect it to the Start Node.

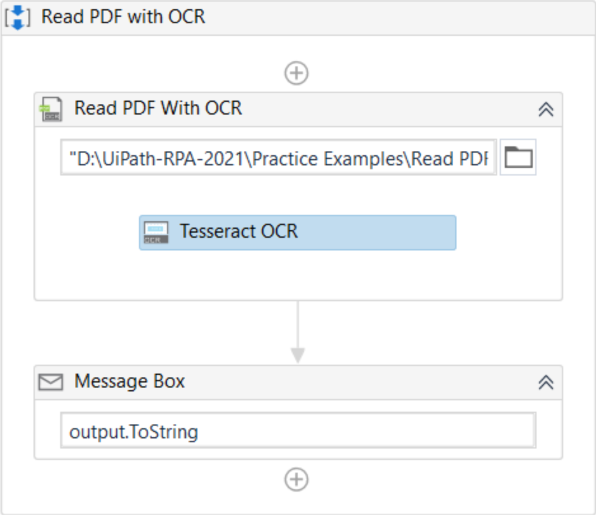
• In the Properties panel, add the expression "Choose one option below:" in the Label field.
• Add the expression {"Read PDF Text", "Read PDF With OCR"} in the Options field.
• Add the value "Options" in the Title field.
• Add the variable chooseOption in the Result field.

4. Place a Flow Decision activity below the Input Dialog activity and connect it to it.

• In the Properties panel, add the expression chooseOption = "Read PDF Text" in

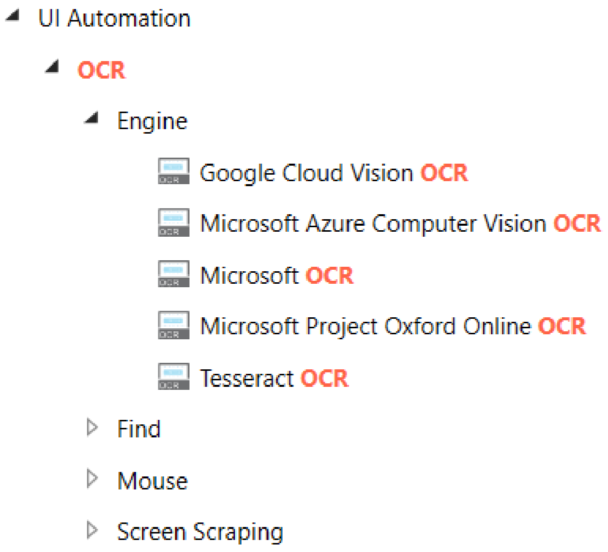
Extracting Large Text – Read PDF with OCR

- Drag and drop Read PDF with OCR Activity
- Specify the path to access the pdf file
- Specify the output variable name
- Drag & drop OCR Engine
- Drag & drop Message box and display output variable
- Hit Run and observer the output
- It will read all the content including IMAGE text also.



"D:\UiPath-RPA-2021\Practice Examples\Read PDF

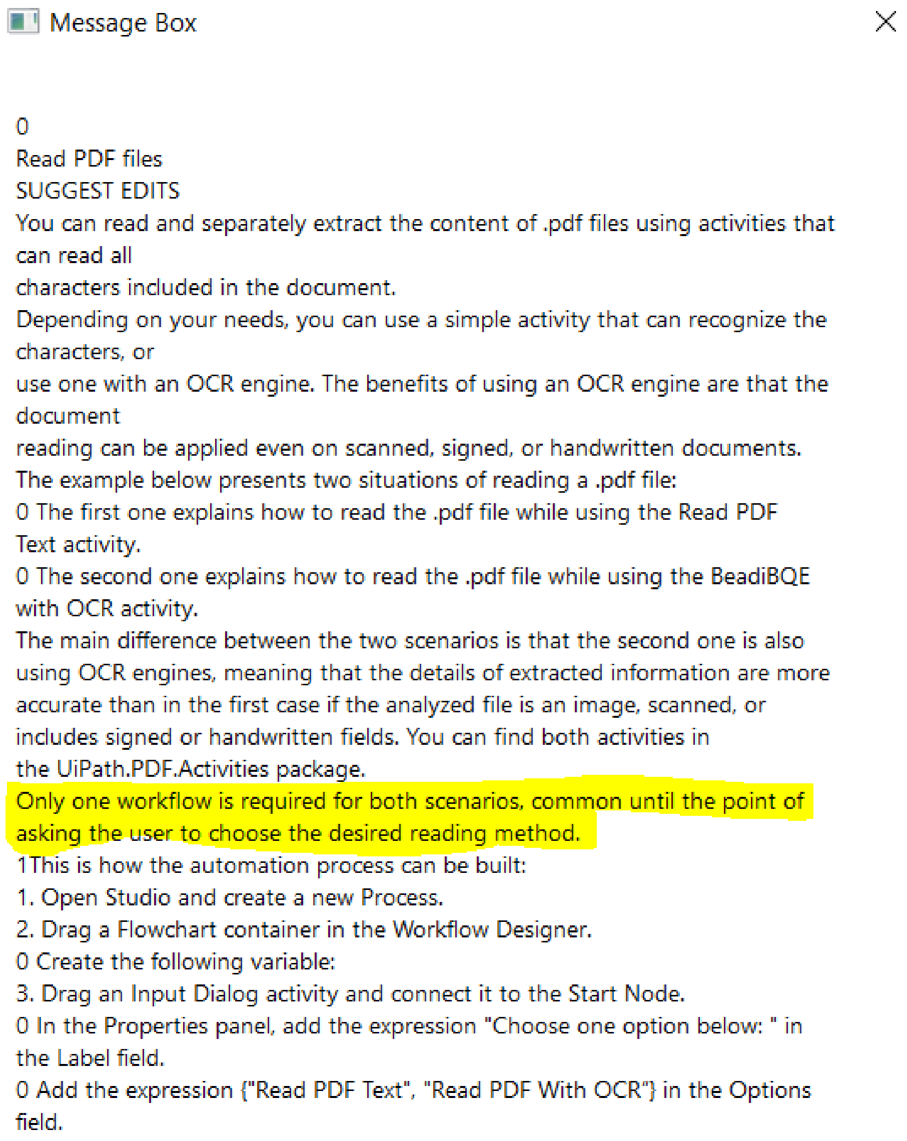
output.ToString



Output

- All the text including IMAGE content

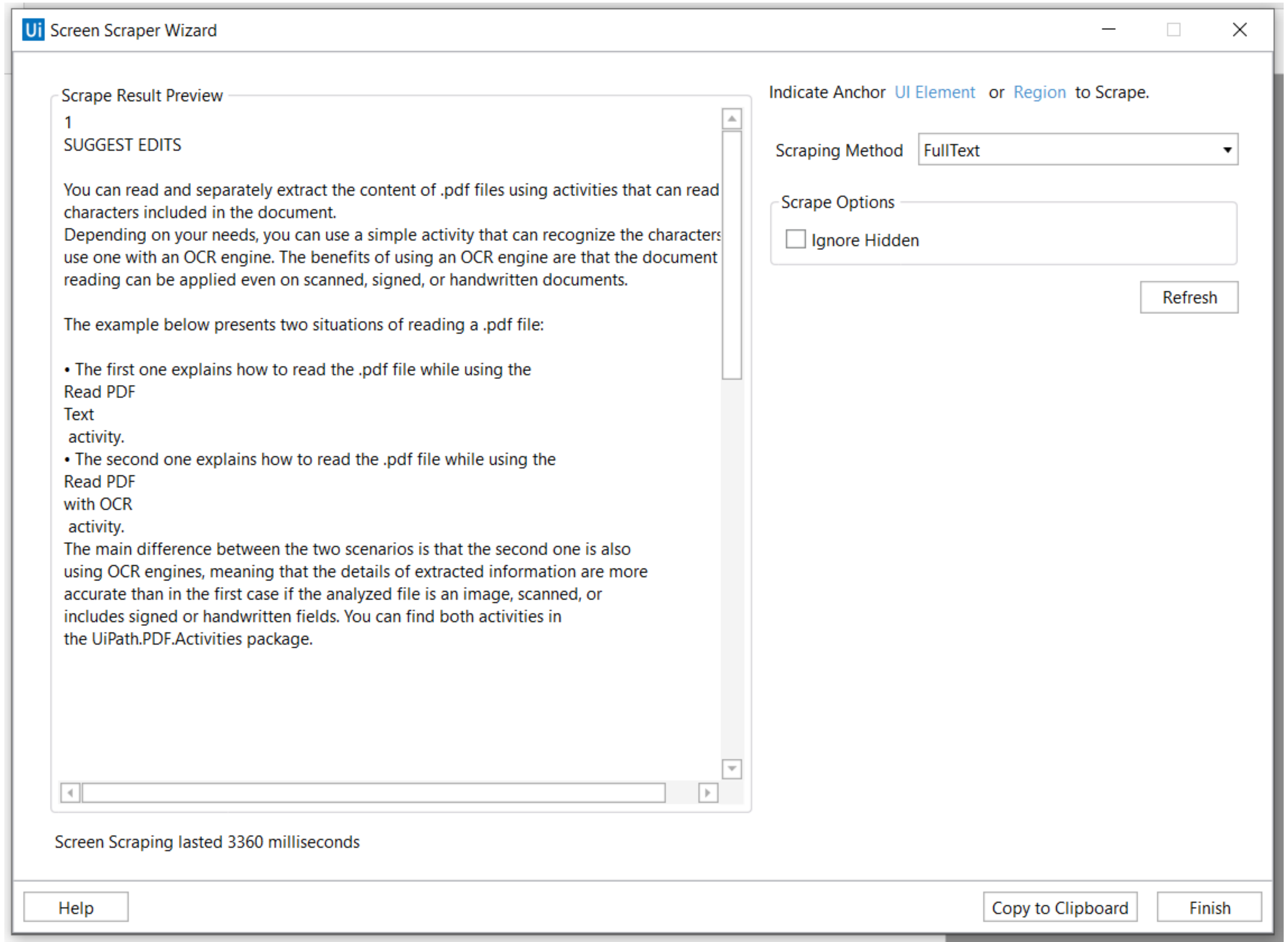
Also got extracted.



```
0
Read PDF files
SUGGEST EDITS
You can read and separately extract the content of .pdf files using activities that
can read all
characters included in the document.
Depending on your needs, you can use a simple activity that can recognize the
characters, or
use one with an OCR engine. The benefits of using an OCR engine are that the
document
reading can be applied even on scanned, signed, or handwritten documents.
The example below presents two situations of reading a .pdf file:
0 The first one explains how to read the .pdf file while using the Read PDF
Text activity.
0 The second one explains how to read the .pdf file while using the BeadiBQE
with OCR activity.
The main difference between the two scenarios is that the second one is also
using OCR engines, meaning that the details of extracted information are more
accurate than in the first case if the analyzed file is an image, scanned, or
includes signed or handwritten fields. You can find both activities in
the UiPath.PDF.Activities package.
Only one workflow is required for both scenarios, common until the point of
asking the user to choose the desired reading method.
1This is how the automation process can be built:
1. Open Studio and create a new Process.
2. Drag a Flowchart container in the Workflow Designer.
0 Create the following variable:
3. Drag an Input Dialog activity and connect it to the Start Node.
0 In the Properties panel, add the expression "Choose one option below: " in
the Label field.
0 Add the expression {"Read PDF Text", "Read PDF With OCR"} in the Options
field.
```

Extracting Large Text – Screen Scraping

- Click on Screen scrape
- Select the item on pdf



1
SUGGEST EDITS

You can read and separately extract the content of .pdf files using activities that can read
characters included in the document.
Depending on your needs, you can use a simple activity that can recognize the characters
use one with an OCR engine. The benefits of using an OCR engine are that the document
reading can be applied even on scanned, signed, or handwritten documents.

The example below presents two situations of reading a .pdf file:

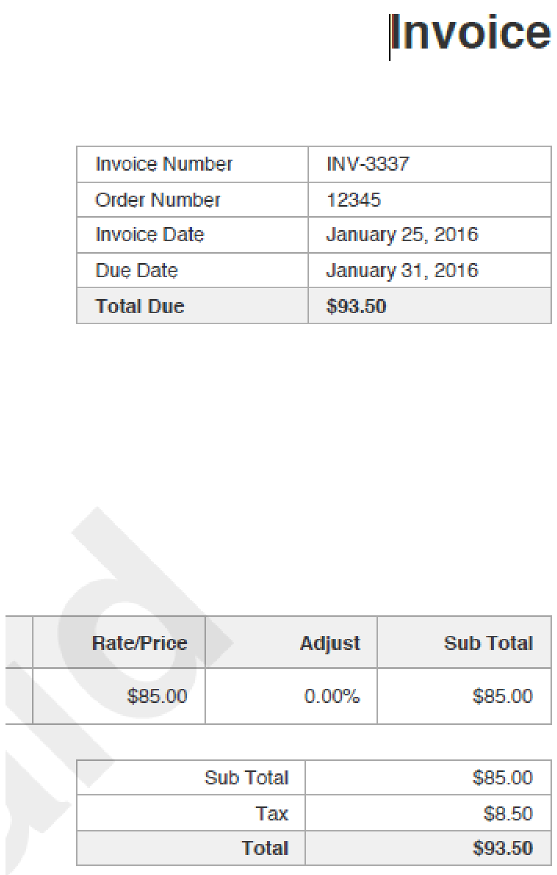
• The first one explains how to read the .pdf file while using the
Read PDF
Text
activity.
• The second one explains how to read the .pdf file while using the
Read PDF
with OCR
activity.
The main difference between the two scenarios is that the second one is also
using OCR engines, meaning that the details of extracted information are more
accurate than in the first case if the analyzed file is an image, scanned, or
includes signed or handwritten fields. You can find both activities in
the UiPath.PDF.Activities package.

Scraping Method: FullText

□ Ignore Hidden

Screen Scraping lasted 3360 milliseconds

Extracting specific Elements



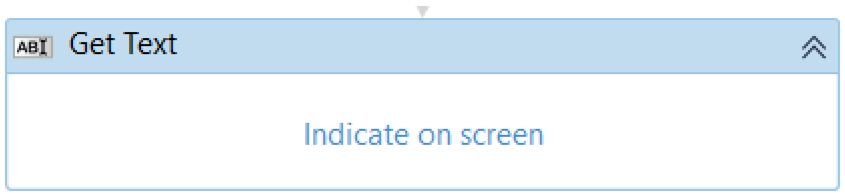
Invoice

Invoice Number	INV-3337
Order Number	12345
Invoice Date	January 25, 2016
Due Date	January 31, 2016
Total Due	**$93.50**

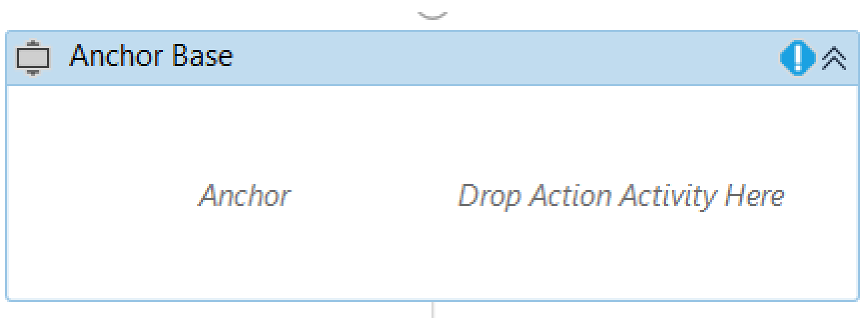
Rate/Price	Adjust	Sub Total
$85.00	0.00%	$85.00

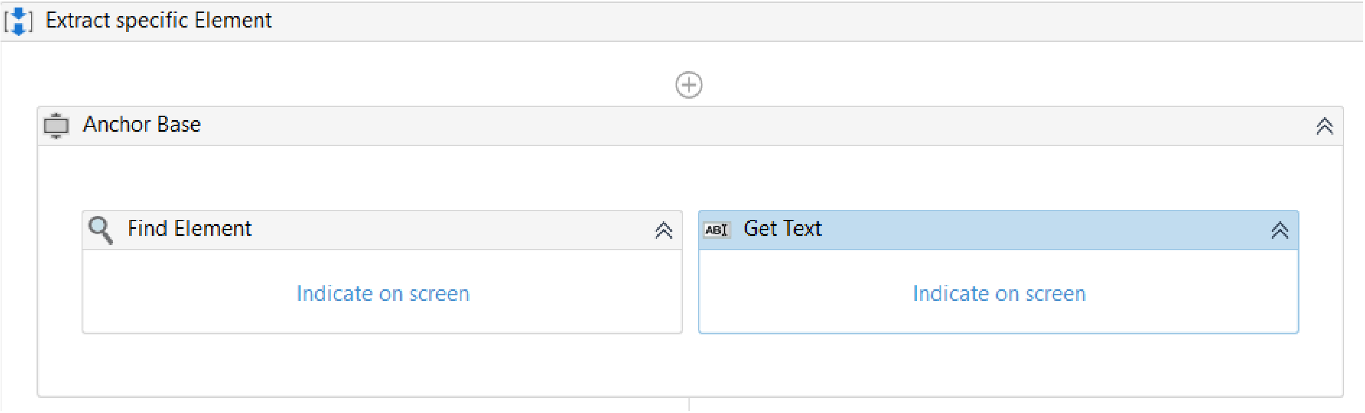
Sub Total	$85.00
Tax	$8.50
Total	**$93.50**

- Get Text – To extract TEXT in a PDF file



- Anchor Base – To extract TEXT and IMAGE in a PDF file





Advantage of Anchor Base

- Get Text activity will fail if Total is getting changed in Invoice
- Anchor Base is the best option to automate or read the Total which is getting changed in each Invoice document.
- Using Find Element Identify the static element (label : Total)
- Using Get Text extract the required text.
- Anchor base will extract the text which is top/left/right side of the Anchor.