**简介**：

PDF是一种广泛使用的文档格式，在本课中我们将学习从PDF文档中提取数据的方法，包括当Selector不稳定时如何使用锚点功能。

**学习大纲**：

如何使用**Read PDF活动**

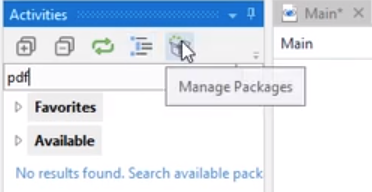
如何使用**Read PDF with OCR活动**

如何使用**锚点**(anchor)从PDF的某个字段里获取数据

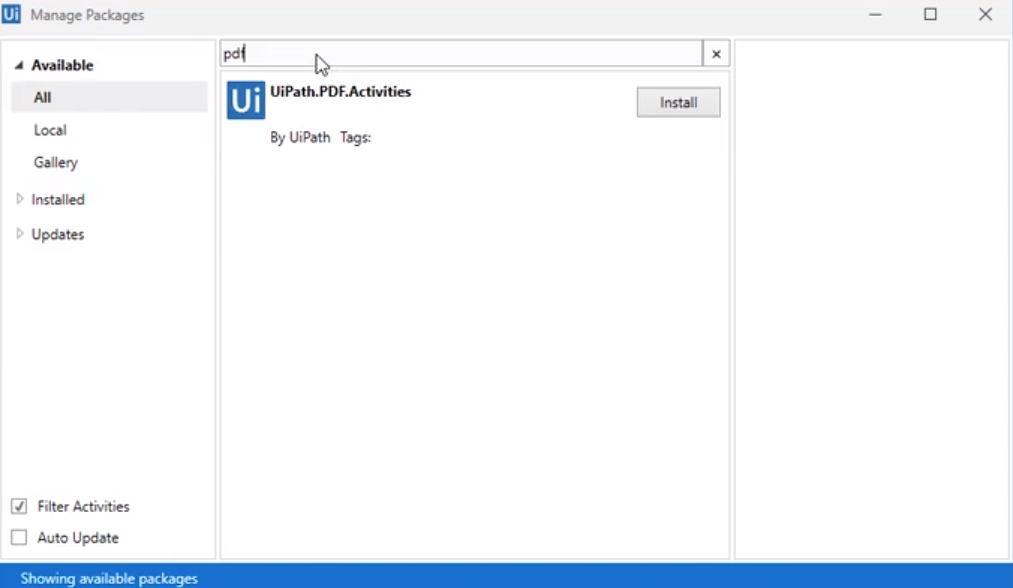
**1. 关于PDF文档**

1) 安装PDF扩展活动包

在Activities面板的搜索栏中搜索"pdf"，如搜索无结果，说明没有安装PDF的活动包。



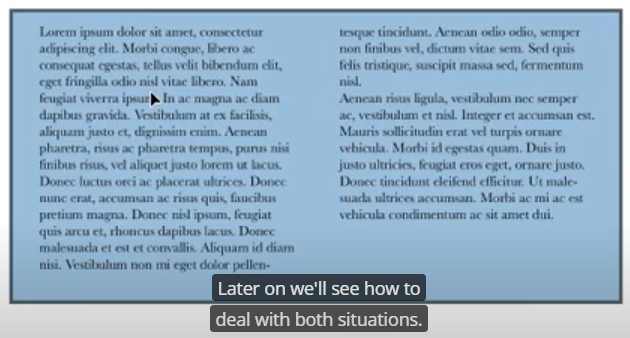
**安装方法**：打开Package Manager，搜索pdf，在搜索结果(UiPath.PDF.Activities)中点击Install。



2) PDF的文本类型

PDF中可以包含文本和图片。有时文本也可能以图片的形式出现。

判断PDF中的某段文字是文本还是图片，可以直接在文件里选中这段文字。如果是可以选中的，就是文本型；如果选择文字时，显示的是一块整的区域（如下图），则是图片型。



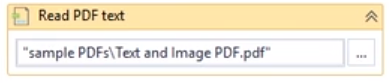
**2. 提取大段文字（三种方法）**

1) **Read PDF Text活动**

FileName参数：选择PDF文件的路径。

Range参数：选择要读取的页码。如"All"，"3"，"3-5"。

Text参数：输出变量。



读取到文本后可以使用**Write Text File活动**把它写入一个txt文档里，或者进行其他的文本操作。

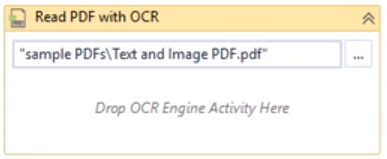
Read PDF Text活动只能读取**可选取的文本**，要读取PDF上的图片，可以使用**Read PDF With OCR活动**。

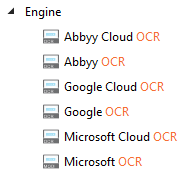
2) **Read PDF With OCR活动**

OCR: Optical Character Recognition

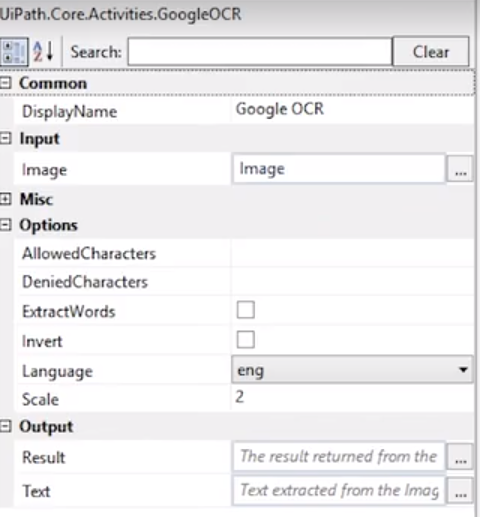
首先要在Read PDF With OCR活动中添加一个OCR Engine。目前UiPath有三种OCR Engine：Google，Microsoft和Abbyy FineReader。

还需要在活动的Output参数里设置一个变量，用来存储读取到的文本。





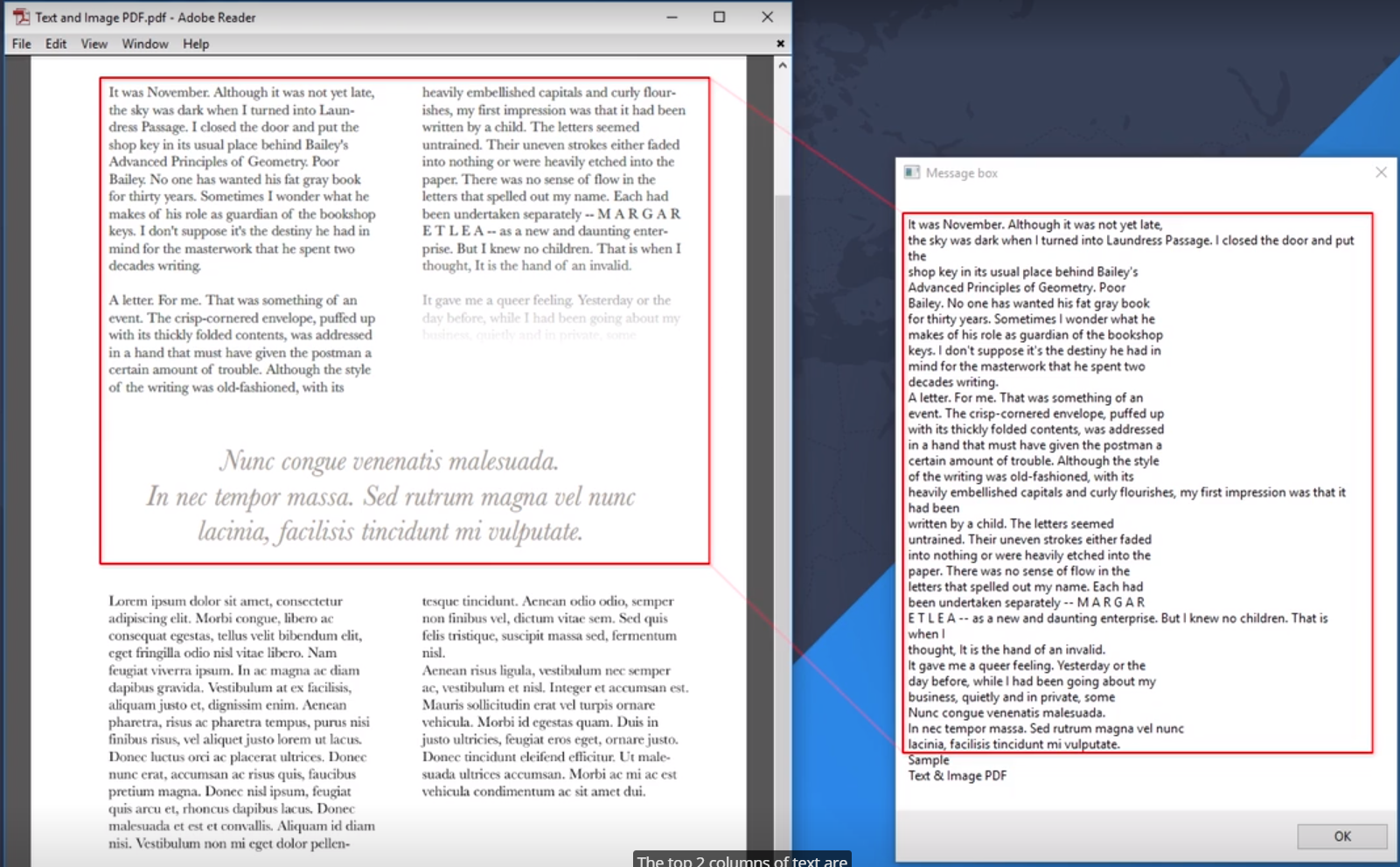
我们可以按需设置OCR Engine的参数（注意：不是Read PDF With OCR活动的参数）。不同的OCR Engine的参数项目可能略有不同。



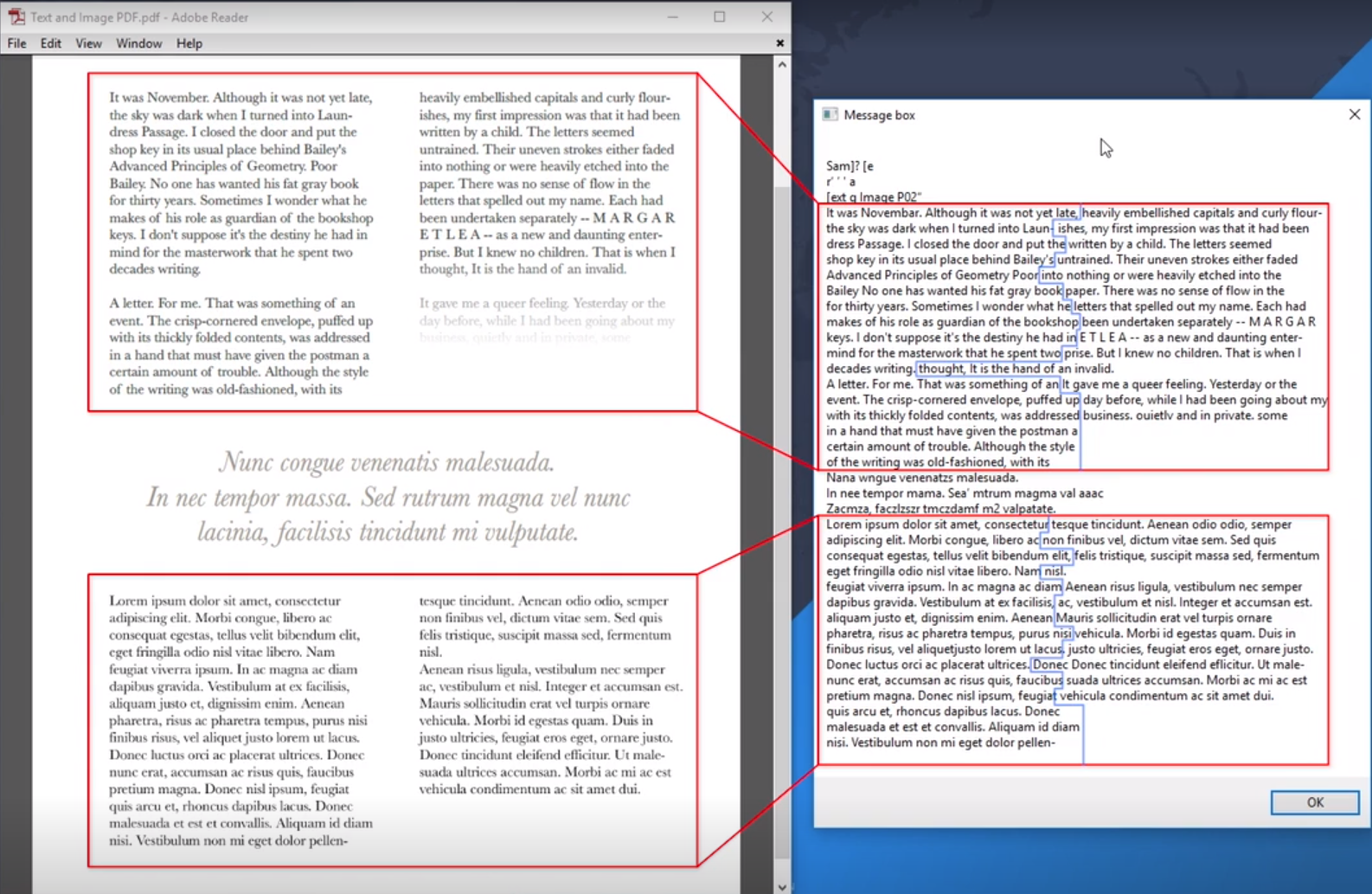
3) **比较**Read PDF Text活动和Read PDF With OCR活动

使用Message Box活动显示Read PDF Text活动和Read PDF With OCR活动读取的PDF数据。

**Read PDF Text活动的运行结果**：PDF上半部分的可选取文本正常显示，且按照分栏顺序显示。下半部分的图片文本显示为“Sample Text & Image PDF”。



**Read PDF With OCR活动的运行结果**：能够显示PDF下半部分的图片文本。上半部分的可选取文本虽然也能显示，但是不能按照分栏顺序显示。



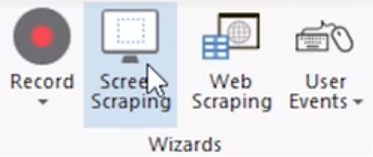
\*目前大部分OCR Engine还不够智能，不能识别出文本中的分栏布局。但Abby是个例外，它可以保留文本的结构。因此在上面的PDF里，使用Abby可以按照正确的顺序显示文本。

\*OCR读取数据的准确率和PDF原图的质量有很大关系。当图片质量较低时，OCR的准确率也会在很大程度上下降。

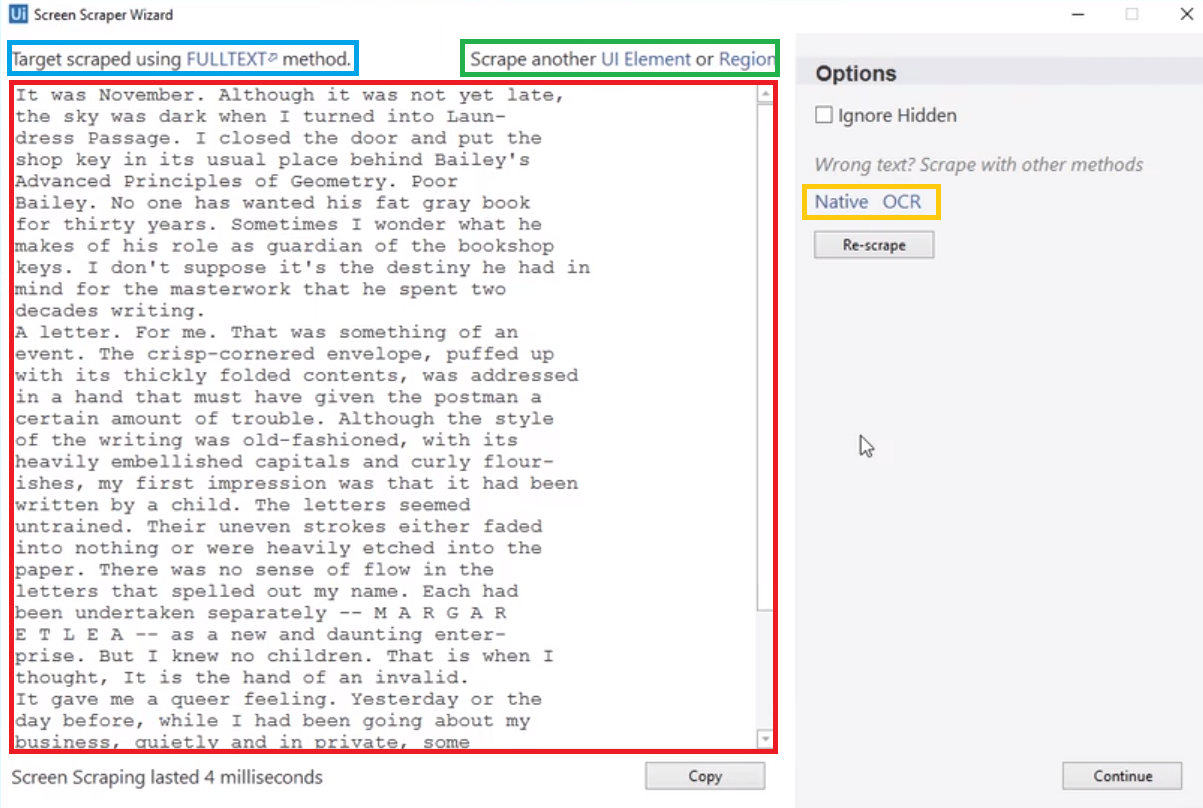
**注意**：上述两种PDF活动都是内部活动，可以在后台运行，且**不需要**打开其他的应用程序（比如阅读PDF的应用Adobe）。

4) **Screen Scraping工具**

在工具栏里选择Screen Scraping工具，在PDF中指定想要抓取的文本元素，UiPath会显示一个预览界面。



 在下图中，红色部分是预览区域，从元素里识别出的文本信息会在这个区域中显示；蓝色部分表示当前使用的Scraping方法；黄色部分表示其他可使用的Scraping方法；点击绿色部分中的UI Element可以重新指定元素。



\*使用Screen Scraping工具的Full Text方法，最终生成的活动是一个**Attach Window容器**和一个**Get Full Text活动**。

\*使用Screen Scraping工具必须先打开PDF文件，且不能在后台运行。

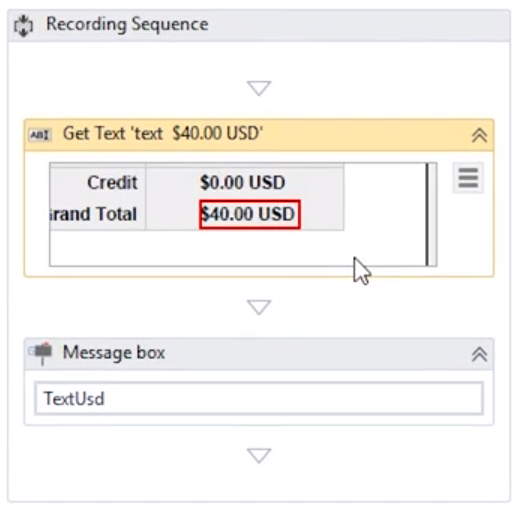
**3. 提取某个特定文本**

PDF数据类型：**Native Text** (本地文本)。这种文本元素可以直接被UiPath识别和选中。与之对应的是扫描版的PDF，文本是图片型的文本。



1) 提取这份PDF上的Grand Total的值。

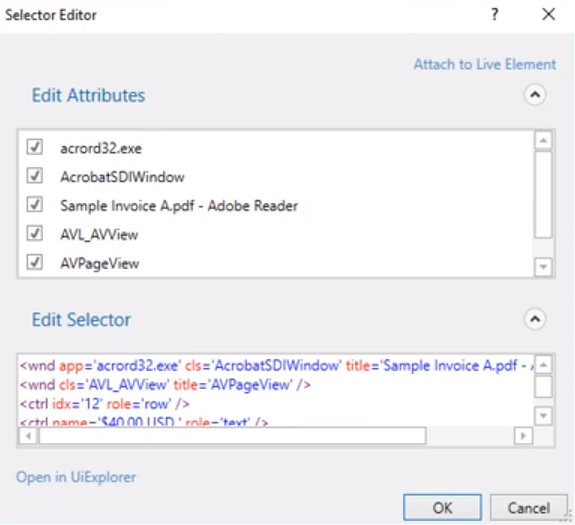
使用**Get Text活动**（或者直接使用Recording工具录制），选择PDF文件里的Grand Total的值。



2) 假设有很多份类似的PDF，提取每一份PDF上的Grand Total。

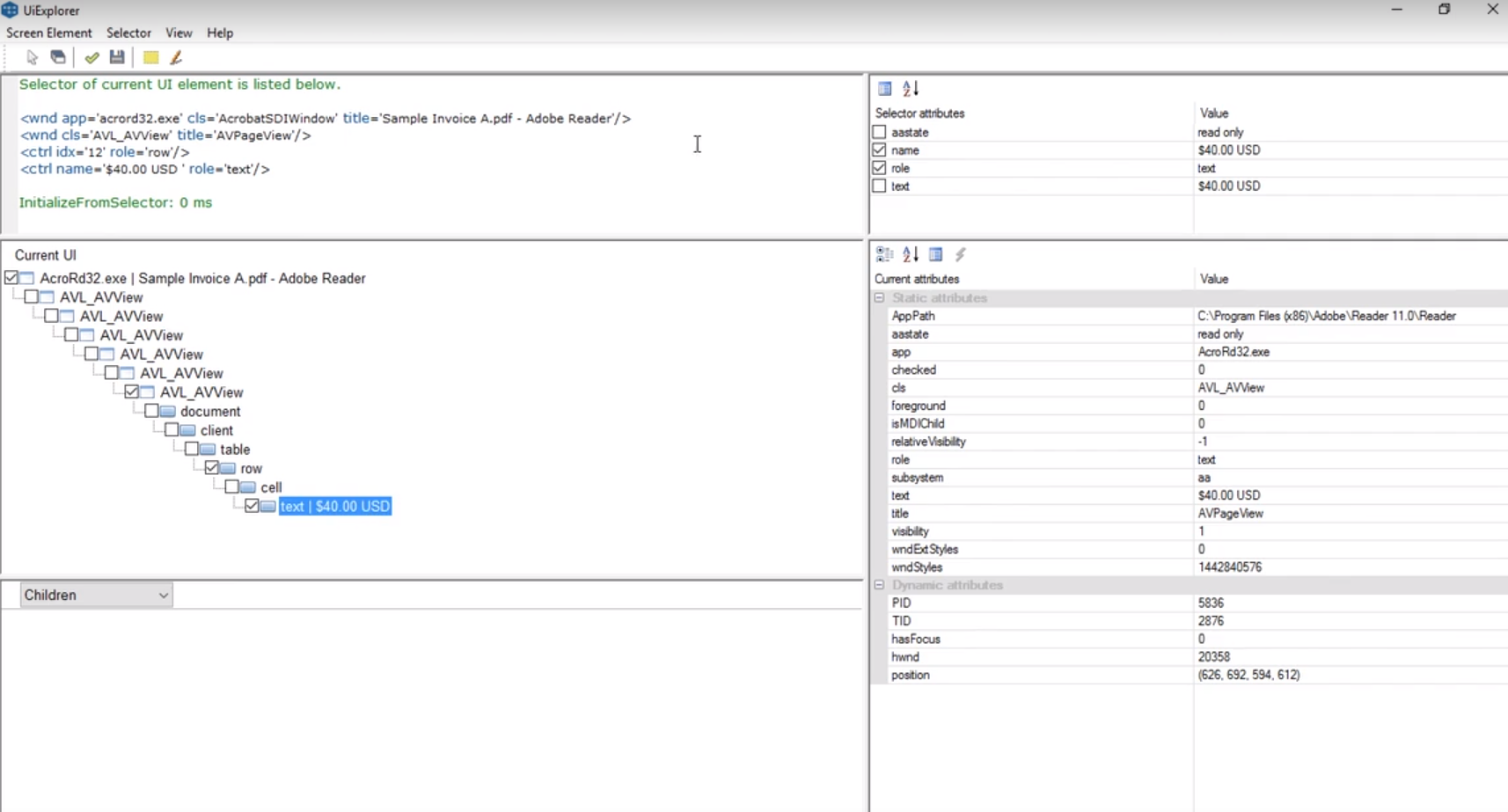
**思路**：修改Get Text活动的Selector，以扩大它的识别范围。

**方法1**：UiPath自动修改。在Get Text活动的参数面板里编辑Selector，在Selector Editor里点击右上角的**Attach to Live Element**，再选择另一个PDF文件里的Grand Total。

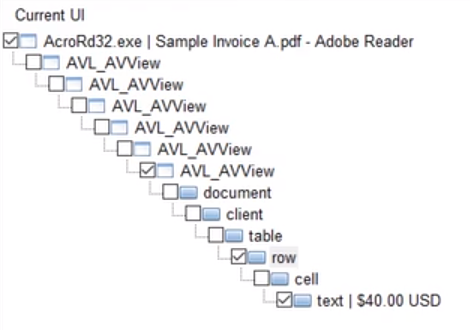


**方法2**：手动修改（修改方法仅限此例，更多详情参见Lesson 6 - Selectors）。

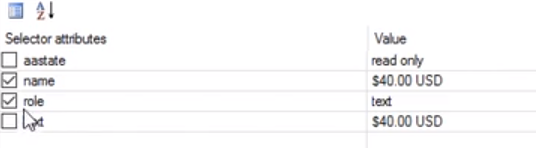
在Get Text活动的参数面板里编辑Selector，在Selector Editor里点击左下角的**Open in UiExplorer**。



在Current UI面板中勾选上的元素即该Selector的组成部分。查看每一个勾选的元素，并在Selector Attributes面板中查看/修改属性。



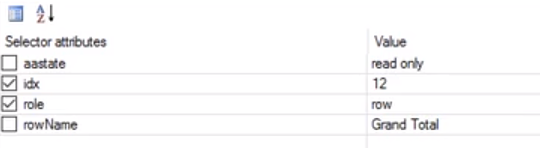
* text | $40.00 USD：去掉name属性。



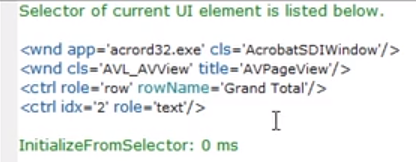
* AcroRd32.exe | Samplee Invoice A.pdf - Adobe Reader：去掉title属性。



* row：添加rowName属性。

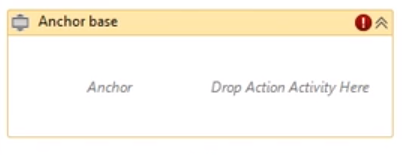


把 最终得到的Selector复制到之前的Selector Editor中。



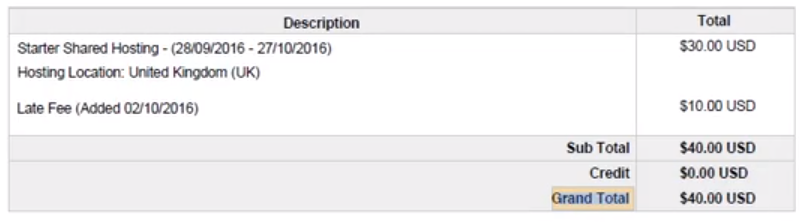
**4. 使用锚点提取特定文本**

1) 添加一个**Anchor Base活动**。

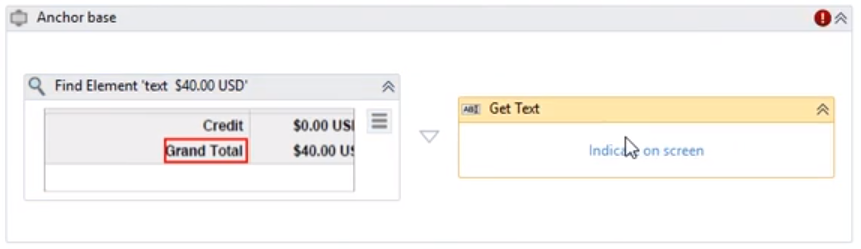


2) 最典型的Anchor (锚点) 就是**Find Element活动**。添加该活动，用它来定位某一个固定的元素。在此例中，指定的锚点是PDF上的"Grand Total"。





3) 使用锚点的目的是提取Grand Total的值，因此在Anchor Base活动的右边添加**Get Text活动**，并指定PDF上Grand Total的值。

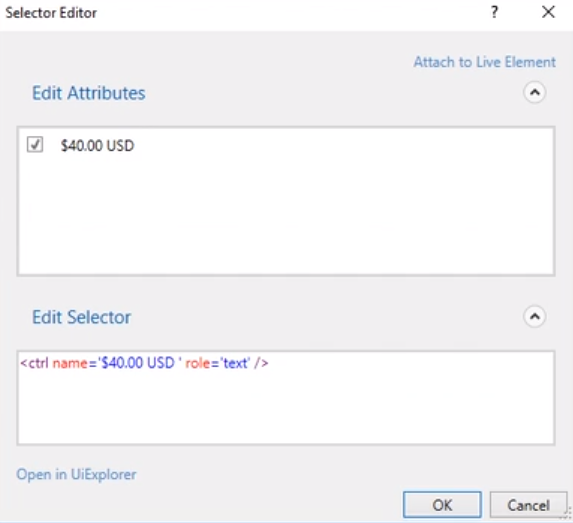




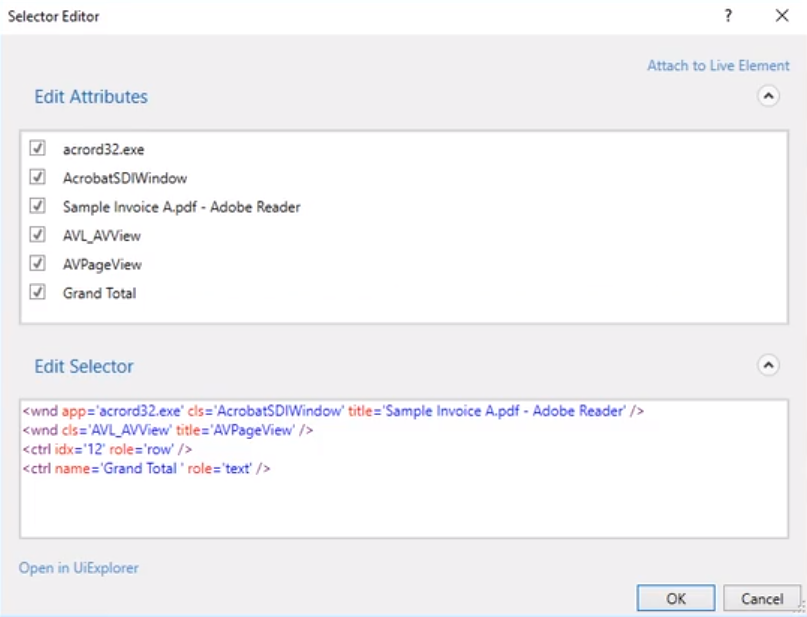
4) 不论是Find Element活动还有Get Text活动，它们都有自己的Selector参数。

**Get Text活动的Selector**：$40.00 USD数值过于具体，可以使用通配符\*进行替换，即<ctrl name='$ \* USD ' role='text'/>

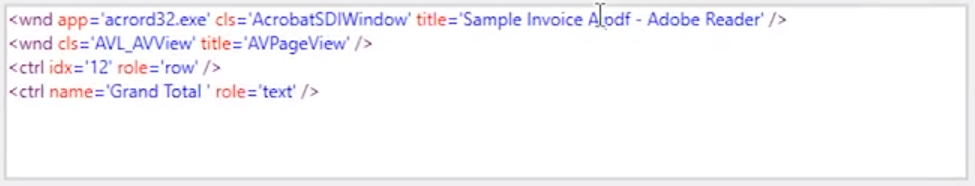
\*由于Get Text活动的Selector根据锚点元素进行定位，这里只显示了完整的Selector的最后一行，即它是一个Partial Selector。



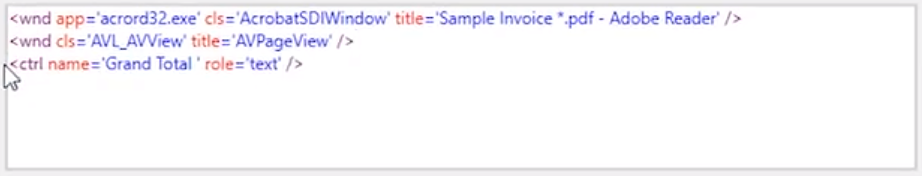
**Find Element活动的Selector**：



修改前：

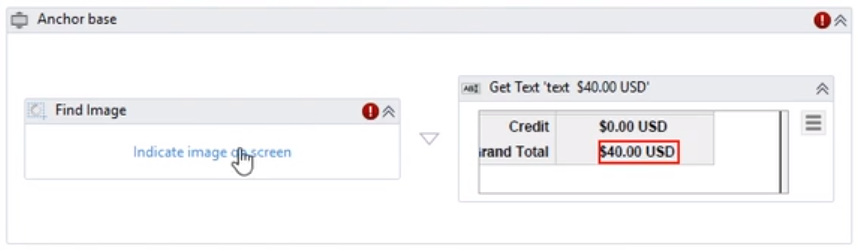


修改后：标题的部分用通配符\*替换；删除第三行，因为它不包含任何唯一的识别标识。



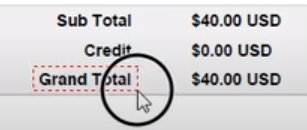
\*Anchor Base活动有一个**Anchor Position参数**，可以设定锚点相对于数据的位置。在此例中，我们可以设定该参数为Left。

5) 使用**Find Image活动**作为锚点，替换Find Element活动。



**优点**：使用Find Element活动作为锚点，只能在相同结构的PDF文档中准确的提取数据。但使用Find Image活动作为锚点，可以不考虑PDF文档的结构，只需要在PDF可见部分的任一位置包含有指定的特定图像即可。另外，使用Find Image活动也不需要花时间去修改和调试Selectors。

* 首先将PDF文档调整为实际大小，以确保获取图片的完整性和准确性：View → Zoom → Actual Size。
* 添加Find Image活动，选取图像。



\*Find Image活动能适应一定范围内的尺寸变化。

\*使用Anchor Base活动，特别是配合Find Image活动一起使用，常比使用其他方法**更加可靠**。因为就算PDF文档的主要结构发生变化，只要文字或图片能显示出来，并且和数据的对应关系不变，UiPath就可以提取出数据。

\*使用Anchor Base活动**必须先打开PDF文档**，并且操作的数据**必须是可见的**（如果数据不在当前页面上，UiPath无法操作该数据）。

**5. 其他方法**

1) **Find Relative Element活动**

2) Scrape Relative工具

3) 对于扫描版的PDF，请参阅Lesson 7  - Image and Text Automation学习如何使用UiPath处理图像。

**6. 练习**

**要求**：假定每个PDF文档中的数据内容都是不同的，使用锚点来获取Date和Bill To的信息。

**思路**：添加**Anchor Base活动**；用**Find Element活动**确定Date和Bill To的文字标签的位置；用**Get Text活动**读取数据。注意**调整Selector**以适应可变的数据。

**\*本课使用过的新活动、方法、函数等：**

**Read PDF Text**

**Read PDF With OCR**

**Find Relative Element**