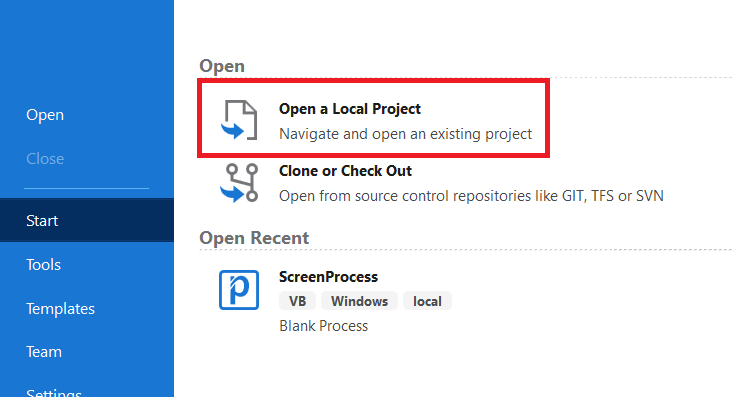
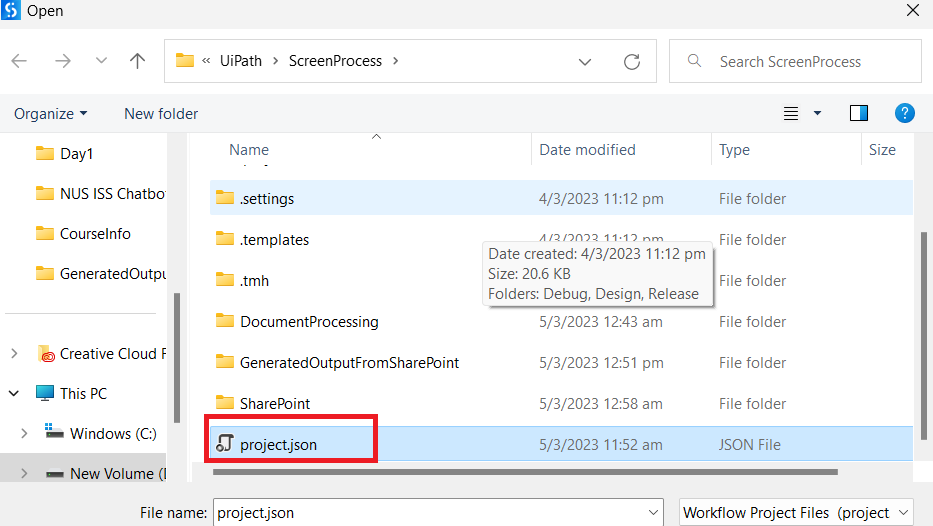
QuickStart Guide

Can follow through the following youtube videos:

1. [UiPath Document Understanding #1 | ExpoHub | By Rakesh - YouTube](https://www.youtube.com/watch?v=PisdmWlanvE&list=PLhTE7-JU1rhYX6RmzEsCRSGqjyJIaKlG6)
2. Install UiPath Studios (Community)
3. After launching UiPath Studio, you need to open the project.



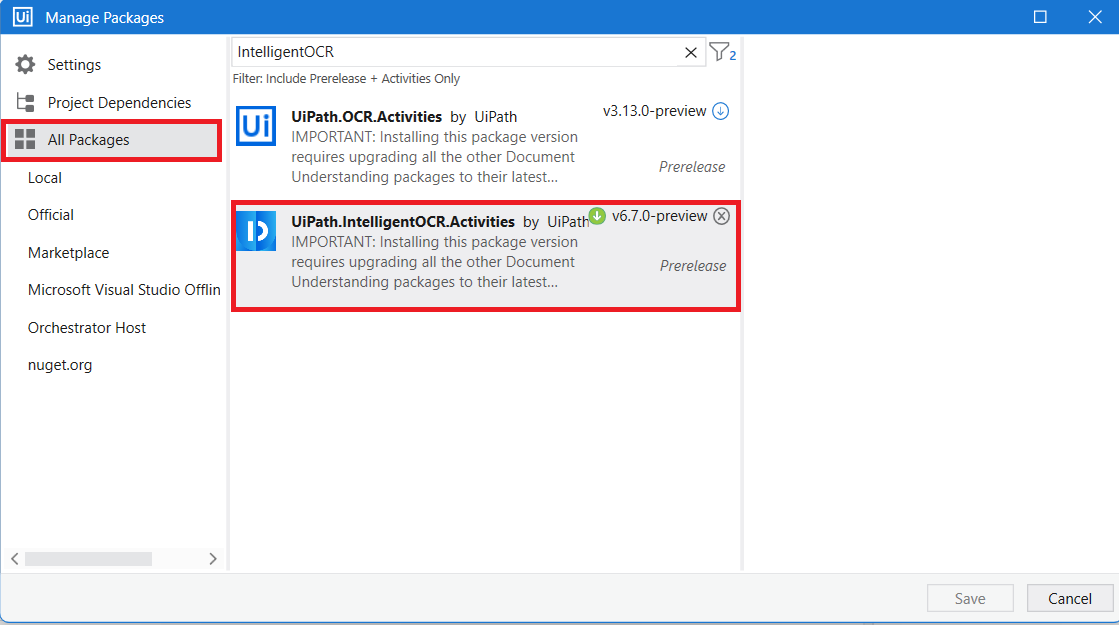
1. Navigate to the following project:



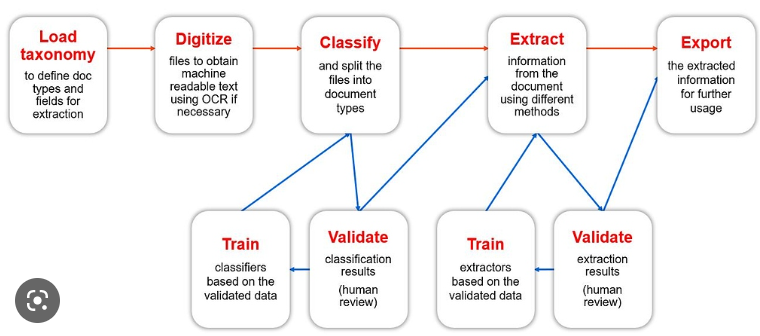
1. you need to pre-install some packages
2. Click on this ribbon:



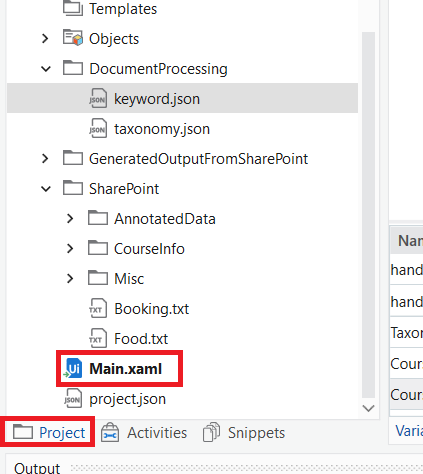
1. Install this:



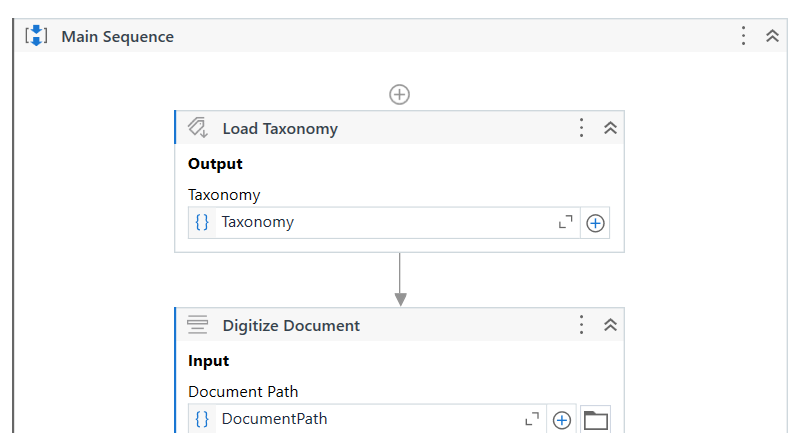
1. UIPath has the following flow, and it is mapped into your project as activities.



To view the flow in UiPath Studios, click on here:

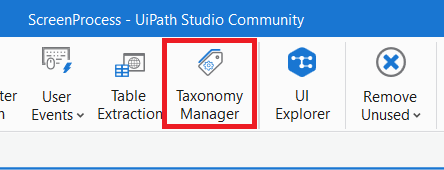


This will bring up the following page on the right:



It should always have Taxonomy first. Which determines the general types of document to read. In current example, we only have 2 types. One is handbook PDF, the other is the course information PDF.

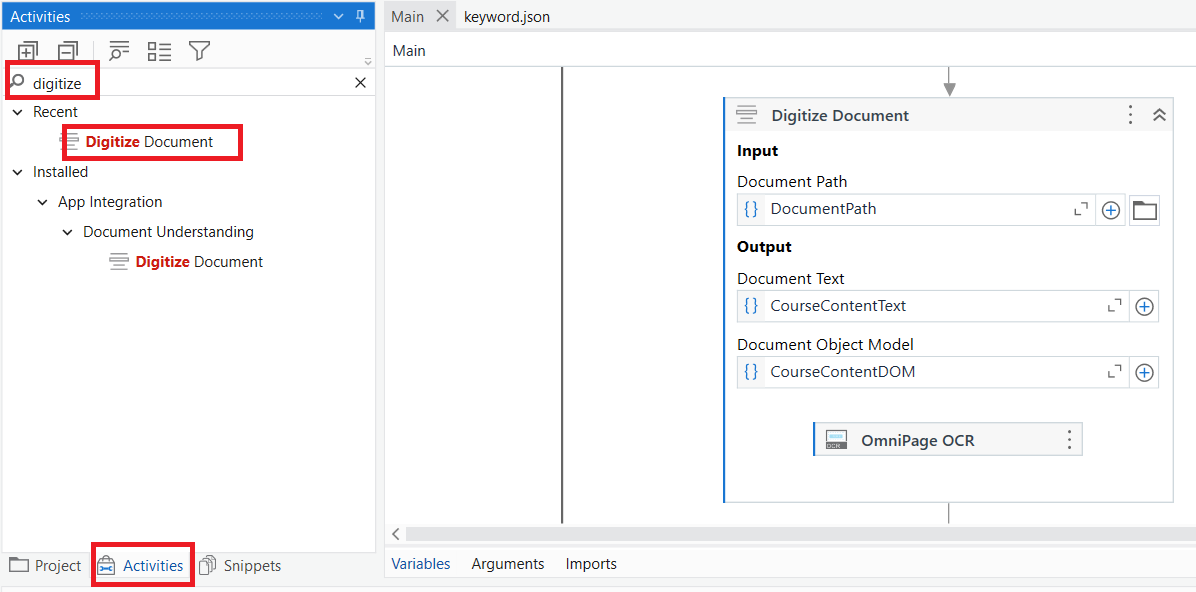
To edit the Taxonomy, click the following Ribbon:



The Taxonomy details should be quite straightforward.

The next part “Digitize Document” is an activity we use to translate PDF into text.

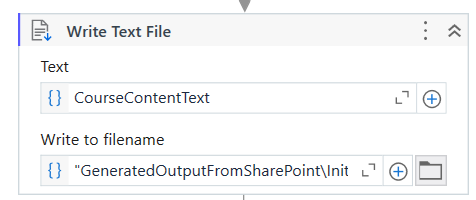
We can insert different activities instead of this one via the activity tab here.



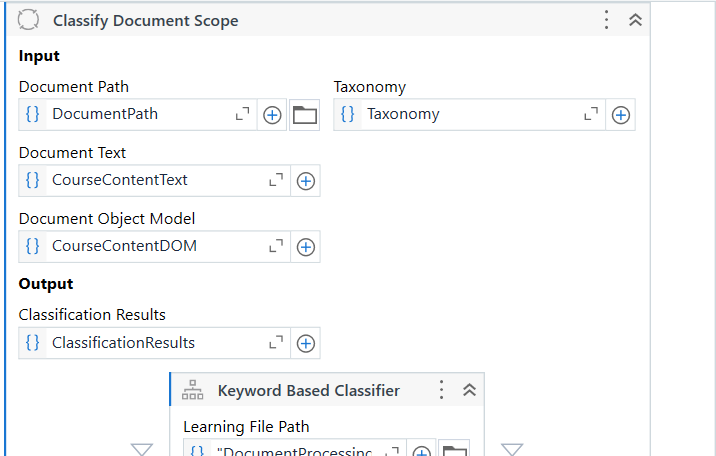
Instead of going to Digitize Document, we can actually do something else.

We do not necessarily use OmniPage OCR. For this example, I’m using this because the youtube video taught me so.

This part writes the text read from pdf into the file. It’s not actually required. Just for fun.



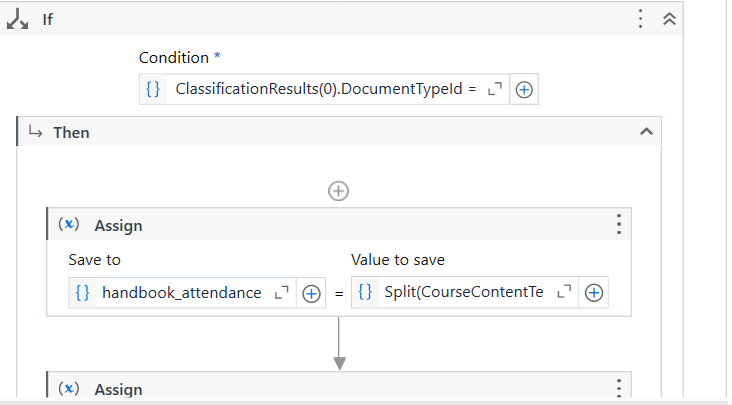
This section classifies whether the text belongs to handbook or course information



This section is supposed to be the extraction part.

There is an IF-ELSE statement, where I currently only handles the “Handbook”

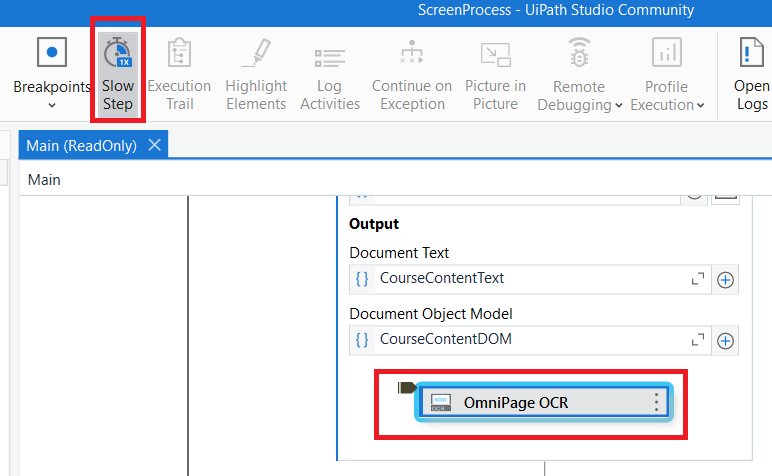
Instead of using the extraction models (regex, flow, ML), I used simple VB scripts to read the required data directly. In this example, I abstracted content from the CourseContent(which is handbook), and I dump the whole section 3 into the variable handbook\_attendance.



1. To run and debug, you should see the message.



While debugging, you can click on “Slow Step” to show the steps executed:



The OCR takes a long time.

1. Generated text file from handbook could be found in directory from :

UiPath\ScreenProcess\GeneratedOutputFromSharePoint\InitialParsedTextContent.txt