Automation techniques in RPA

giovedì 23 maggio 2024 10:44

EXTRACTION AND ITS TECHNIQUES

Extraction is the process of retrieving data from a data source for further processing or storage.

Data sources can be: Screen, PDF, Image, Excel, Email, Citrix

In Studio there are 3 techniques for extraction:

SCREEN SCRAPING

It's the process of extracting data from a specified UI element or document, such as a PDF file. Screen Scraping methods are the core of all the activities that enable extracting data from a specified UI element or document.

It's used to take data out from various screens or documents

It enables data extraction from different UI elements that the automation workflows interact with.

In Studio, to extract text from a UI element, there are 3 methods

To use these methods, a Screen Scraping Wizard is used.

↳ It enables to point at a UI element and use one of the 3 methods and extract text from it.

(1) Full Text Method

It's the default output method in Studio.

It's FAST, ACCURATE and it can work in the BACKGROUND

(-) It captures all the text from the terminal screen, including the hidden text (=> ex: options in a dropdown list)

(-) It doesn't work on Citrix and different virtual environments

(-) It doesn't retain formatting and text position

(2) Native Method

It can be used only with applications that are built to render text with "Graphics Device Interface"

It [illegible] if each word are nested

(-) Useful when the [illegible]

(-) Useful when the font and color of the text are required to be retrieved accurately

(-) Only extracts the visible text

(-) Cannot work w the background and doesn't support Cmtix

This method offers 2 options: < No formatting option / Get Words Info option

(3) OCR method

Both FullText and Native have excellent results in term of accuracy and speed, there are specific cases in which both are unusable.

(-) It should be used if you need to extract information from virtual environments or "read" text from images

(-) It is based on the OCR technology used in recognition of scanned documents

(-) It attempts to recognize each letter of text given on an image in the target document

(-) It is slow when compared to other methods and has lower accuracy

(-) It cannot extract hidden text and cannot work in the background

The OCR method has 3 default engines

Tesseract OCR | Microsoft OCR | UiPath Screen OCR

The use of these engines depend on the type of information being extracted, in general it's better to switch between the methods to see which engines bring better results for each situation.

Tesseract OCR

It gets better results for character recognition on smaller size areas and supports color inversion. It offers multiple customization options through filters that can be used to select only specific categories of characters.

It offers 5 options:

1) Language : English by default

2) Characters : enables you to select which types of characters are to be extracted

3) Scale : the scaling factor of the selected UI element or image, the higher the number is, the more enlarged the images.

This can provide a better OCR rate and is recommended for small images

4) Invert : when this checkbox is selected, the colors of the UI elements are inverted before scraping. It's useful when the background is darker than the text color.

5) Get Words Info : it helps in getting the onscreen position of each scraped word.

Microsoft OCR

It's used to work with Microsoft fonts and large-size images.

It supports multiple languages.

It offers 3 options :

1) Languages 2) Scope 3) Get Words Info

UiPath Screen OCR

It can be used in any UI Automation scenario in which an OCR engine is needed.

It offers 3 options :

1) End Point : here the endpoint of the UiPath Screen OCR is entered.

2) API Key : it's used to provide you access to the UiPath Screen OCR. It's not required for the preview period.

3) Get Words Info

Data Scraping

It's the process of extracting structured data from :

Browser, Application, Document

It's used to create a Data Table at run time.

(•) It extracts all the pattern-based data and store it into the form of the Data Table automatically

(•) Generates a Container such as Attach Browser or Attach Window with a Selector for the top-level

(b) Detects if a table cell is [illegible] table is to be extracted

(c) Scrapped information is stored in a Data Table Variable that can be used later to populate a database.

PDF Extraction

It's the process of extracting the raw data from PDF documents. PDF files can contain text, images, and sometimes text that are actually images. PDFs can be 2 types:

1) Scanned PDF:
The information stored in these PDFs is image format only and generally cannot be searched or copied.

2) Native PDF:
The information stored in these PDFs is in text and image format and can be searched and copied.

Studio offers several activities to extract data from PDFs

↳ You must install the UiPath.PDF.Activities package on the system with the help of the Managed Package section

Two activities to read PDF data:

(1) Read PDF Text Activity:
It reads all characters from a specified PDF file and stores them in a string variable. It chooses the file to be read and outputs a text variable with the contents of the file. The result can be saved as a text file and displayed in a message box.

The activity extracts or converts only the text part of the document, and any images in the document are ignored.

(2) Read PDF with OCR Activity
It reads all characters from a specified PDF file and stores it in a string variable by using OCR technology.
It scans the images with OCR inside the PDF document and output raw the text as a variable.
It requires an OCR engine for the scanning procedure.

An important property of these activities is Range
↳ It specifies the range of pages that you want to read.

Both PDF activities are self-contained, that is they don't require other applications to be open, so they can run in the background.

a) Get PDF Page Count activity: provides total number of pages in a PDF file

c) Export PDF Page As Image activity: it creates an image from a page in a specified PDF file

e) Extract Images From PDF Activity: it extracts images from a specified PDF file and saves them in a folder.

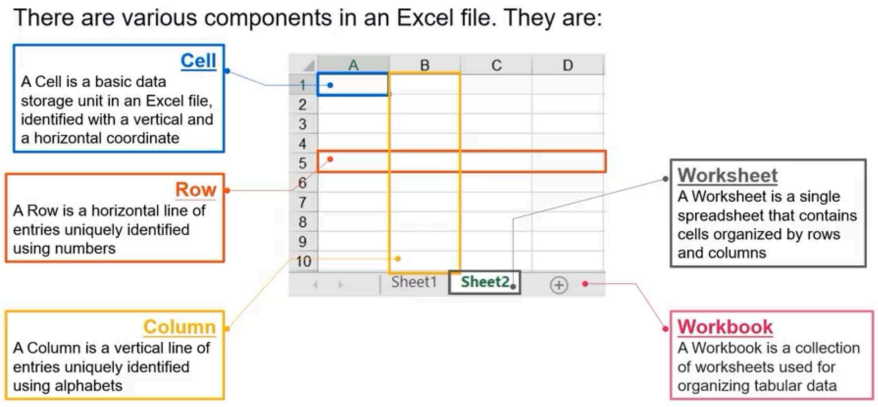
e) Extract PDF Page Range Activity: it extracts text from a specified range of pages from a PDF document.

c) Join PDF Files Activity: joins multiple PDF files in an array of strings into a single PDF file.

c) Manage PDF Password Activity: it manages the password of a specified PDF file if current password details are known.

Automation Techniques

Excel Workbook and Automation



In many business scenarios, databases are stored in workbooks. From there, they can be inserted into Data Tables and processed further. Studio offers two sets of activities to access and manipulate workbooks:

1) Workbook activities

are executed in the background; doesn't require Excel to be installed on the computer; faster and more reliable for some

operations when the user doesn't open the file;
works only for .xlsx files.

2) Excel Activities:

Should open Excel; it requires Excel to be installed on the computer;
if the file isn't open, it will be opened, saved and closed
for each activity; All activities can be set to either be visible
to the user or run in the background;
works with .xls and .xlsx, and it has some specific
activities for working with .csv.

Common Activities to Workbook and Excel:

(•) Append Range: it adds the information from a Data Table to the end of a specified Excel spreadsheet. If the sheet does not exist, it creates it.

(•) Get Table Range: it locates and extracts the range of an Excel table from a specified spreadsheet using the table name as input.

(•) Read Cell: it reads the content of a given cell and stores it as a string

(•) Read Cell Formula: it reads the formula from a given cell and stores it as a String.

(•) Read Column: it reads a column starting with a cell inputted by the user and stores it as an IEnumerable < Object > variable.

(•) Read Row: it reads a row starting with a cell inputted by the user and stores it as an IEnumerable < Object > variable.

(•) Read Range: it reads a specified range and stores it in a Data Table. If "Use Filter" is checked, it will read only the filtered data. This option does not exist for the Read Range activity under "Workbook".

(•) Write Cell: it writes a value into a specified cell.
If the cell contains data, it will overwrite the content.
If the sheet specified doesn't exist, it will be created.

(•) Write Range: it writes the data from a Data Table variable in a spreadsheet starting with the cell indicated

Excel Application Scope

The integration with Excel is enabled by using an Excel Application Scope container.

The purpose of the container is to open the Excel workbook and provide a scope for Excel activities.

There are some important properties:

- Workbook path: used for the full path to the Excel workbook.
- Read-only: helps prevent editing or writing to the file in scope.
- Visible: if checked helps open and read the file using Microsoft Excel

There are specific activities to Excel App Integration

- Range: can read data, insert and delete rows and columns, and even copy/paste entire ranges. They are similar to the corresponding activities under Data Table, but they work directly in the Excel file. The activities under this category are:
  - Delete Column
  - Insert/Delete Column
  - Insert Column
  - Insert/Delete Row
  - Select Range
  - Get Selected Range
  - Delete Range
  - Auto Fill Range
  - Copy Paste Range
  - Lookup Range
  - Remove Duplicate Range

- Table = this activities create, filter and sort tables directly in Excel files. The activities under this category are:
  - Filter Table → only rows that meet the filter will be displayed
  - Sort Table → based on values of a given column
  - Create Table

- File: these activities work directly with the Excel files, either by saving or closing them. The activities are:
  - Close Workbook
  - Save Workbook

- Cell Color: these activities can capture and modify the background color of cells in Excel files. The activities are:

(-) Get Cell Color (-) Set Range Color

(-) Sheet : These activities can perform various actions over the sheets in an Excel file. The activities are :

(-) Get Worksheet Sheet (1) (-) Copy Sheet

(-) Pivot Table : These activities facilitate working with pivot tables.

The activities are :

(-) Refresh Pivot Table (-) Cache Pivot Table

{

Useful when the Pivot table's source changes, as the table doesn't refresh automatically

(-) Macro : These activities can execute macros already defined in the Excel file or invoke macros from other files.

These activities work with .xlsm files and they are :

(-) Execute Macro (-) Invoke VBA

Email Automation

From an RPA perspective, there are 2 situations for interacting with emails:

1) Email as input for the Business Process :

The automation uses information from an incoming email.

Data is received in textual form as a part of the subject or body of the email. Data can also come in the form of an attachment.

For example, names or ID numbers from the subject or body, or input files coming as attachments.

2) Email as output for the Business Process :

Automation generates and sends email according to the set rules.

Email is used for sending status updates to users regarding different automation projects or when business or application exceptions occur.

For example, progress reports and exception alerts.

Email automation involves a number of configuration steps and different types of applications, and the process requires the usage of a number of

"If Else" control statements and standard, structured and unstructured input and output types.

Email Protocols

Email Protocol is a method by which a communication channel is established between two computers and email is transferred between them.

Depending on the scope of automation, different sets of activities and protocols can be used.

The different protocols are:
- (•) SMTP
- (•) POP
- (•) IMAP
- (•) Microsoft Exchange
- (•) Microsoft Outlook

1) SMTP ~> can be used with Gmail to send emails

2) POP3 ~> used for reading messages and to retrieve email from Gmail

3) IMAP ~> offers useful features to mark messages as read or move them between folders. Input is email, it can also be used to retrieve email from Gmail.

4) Microsoft Exchange ~> it used to send and receive emails.
It uses a proprietary protocol called MAPI to talk to email clients.
The Exchange Scope is used to connect to exchange and provide scope for other Exchange activities

5) Microsoft Outlook ~> it's used as an email client. It sends and receives emails. It also comes with a calendar, task manager, scheduling meeting requests and other information management facilities.

Categories of Email Automation

1) Generate and Send automated messages ~> SMTP, Outlook, IBM Notes, Exchange
2) Retrieve messages and extract data ~> POP3, Outlook, IMAP, IBM Notes, Exchange
3) Manage messages ~> Outlook, IMAP, IBM Notes, Exchange
4) Save email attachments and essential information in a well-defined structured form ~> Generic email activities
5) Save messages ~> Generic email activities

There are 2 generic email activities used to save:

⊙ Save Attachments => it saves the mail message attachments to the

specified folder; if the folder doesn't exist it's created;
if no folder is specified, the downloads are saved in the
project folder; files in the specified folder with the same name
as the attachments are overwritten.

⊙ Save Mail Message