Week 8 : Extraction and its techniques

- ตั้งชื่อ Project ด้วย week8_รหัสนักศึกษา เช่น week8_62070001
- ในแต่ละข้อ สร้าง sequence flow ด้วย เลขที่ข้อ_รหัสนักศึกษา เช่น 8.1_62070001
- ต้อง capture รูป flow แปะใน word ให้เห็นชัดเจนและตั้งชื่อไฟล์ เป็น เลขที่ข้อ_รหัสนักศึกษา.docx เช่น 8.1_62070001.docx
- ส่งงานทุกข้อเป็น 1 Project ในที่รวม folder ทั้งหมดในโปรเจคนั้นๆ ตั้งชื่อไฟล์ week8_รหัสนักศึกษา.zip เช่น week8_62070001.zip

Exercise 8.1 (see clip 8.1 ScreenScraping)

Demonstrate how to scrape text from a UiPath blog post using **Screen Scraping** wizard and store it in a Notepad File.

Prerequisite:

1. Open a blog post on UiPath Website (https://www.uipath.com/blog).
2. Choose any blog you like.

Algorithm

1. START
2. Go to UiPath Studio, and click **Screen Scraping** button from the Design ribbon. Indicate the text you want to copy from the UiPath blog page.
3. In the Screen Scraper Wizard, switch between different scraping methods. Choose each one by one to see their results in the preview area.
    - Click the Scraping Method drop down and select **Native**.
    - Click the **Refresh** button and see the preview.
    - Again click the Scraping Method drop down and select **OCR** and click **Refresh** button. The output of OCR is in the paragraph style as seen on the website.
    - Click the Scraping Method drop down and select **Full Text** and click **Refresh** button and see the preview.
4. Drag and drop **Write Text File** activity below Screen Scrapping container.
5. In the Text box, enter the variable in which the scraped content is stored. In the Write to filename: box, enter the filename “question_studentID.txt” i.e. **8.1_62070001.txt**.
6. STOP
7. Save and run the workflow.
8. Open the saved Notepad to see if the scraped content is stored

Outcome:

- The content is stored in the file.

Exercise 8.2

- Build a workflow using **Screen Scraper Wizard** that scrapes text using **Full Text** scraping method and stores in a Notepad file.
- Search for “UiPath” in Google Search.
- Scrape information about UiPath shown on top right of the result page using **Full Text** scraping method.
- Store text in a Notepad file.

Algorithm

1. START
2. Use **Open Browser** activity to open URL – “www.google.com”
3. Choose **BrowserType** property
4. Use **Type Into** activity in Open Browser activity to indicate search bar of Google and search for “UiPath” and Plus **Enter** key
5. Use **Screen Scraping** button in Design ribbon and select information about UiPath shown on the right of the Google search results page.
6. Choose **Full Text** as Screen Scraping method.
7. Use **Write Text File** activity to store scraped content in a Notepad. In the Write to filename: box, enter the filename “question_studentID.txt” i.e. 8.2_62070001.txt.
8. STOP

Outcome

- All the texts from the PDF file is stored here.

Exercise 8.3

Build a workflow using **Screen Scraping** wizard that scrapes text using **Tesseract OCR** scraping method from an image and stores in a Notepad file.

- Search for “text images” in Google Images
- Pick one image containing text from the search results
- Scrape the text from the image using Tesseract OCR
- Store text in a Notepad file

Algorithm

1. START
2. Use **Open Browser** activity and enter URL www.google.com/images
3. Use **Type Into** activity in Open Browser activity and search “text images” in the URL and Plus **Enter** key
4. Use **Screen Scraping** from Design ribbon to select an image containing text.
5. Choose **Tesseract OCR** as Screen Scraping method. Change **Scale to 5**.
6. Use **Write Text File** activity to store content in a Notepad file. In the Write to filename: box, enter the filename “question_studentID.txt” i.e. 8.3_62070001.txt.
7. STOP

Outcome

- All the texts from the PDF file is stored here.

Exercise 8.4 (see example in clip PDF extraction)

Demonstrate how to extract text (page 2) from a pdf file using **Read PDF Text** activity and store in a Notepad file.

Prerequisite:

1. Go to UiPath Studio and click on Manage Packages button in the Design ribbon. In the pop up window, navigate to All Packages
2. Search for UiPath.PDF.Activities and click the first result. In the right panel, click **Install** and **Save**. In the terms and conditions pop up window, select "**I Accept**". Wait for few seconds till the dependencies are installed.
3. Add folder named "PDF" in this project and put "samplePDF.pdf" in this folder
4. Open samplePDF.pdf to see how it looks like and close it

Algorithm

1. START
2. Drag and drop **Read PDF Text** activity in the designer panel.
3. Click the "PDF" Folder icon and select the "samplePDF.pdf" file.
4. Go to the Properties panel of the Read PDF Text activity, and in the Output property, press Ctrl+K and insert a new variable called **content**.
5. Change from "All" to **"2"** in the **Range** property
6. Insert **Write Text File** activity and insert below Read PDF Text activity.
7. Enter the variable **content** in the **Text** input box, and "question_studentID.txt" in the **Write to Filename** input box.
8. STOP
9. Save and run the workflow.
10. Go to the folder where the Notepad is saved and open it

Outcome

- All the texts from the PDF file is stored here.

Exercise 8.5

Build a workflow using Read PDF Text activity and extract only Email IDs and Phone Number from a PDF file and store in a MS Word file. Finally, show the number of emails and phone numbers.

Prerequisite:

1. Put “challenge.pdf” in the “PDF” folder
2. Open challenge.pdf to see how it looks like and close it
3. Find the Word icon and Pin it to Task bar

Algorithm

1. START
2. In desktop recording, choose **Click** and click at MS Word at the taskbar
3. Use **Attach windows** to perform attaching to an already opened window
4. Use **Read PDF Text** activity to read content of the PDF file and store in a string variable.
5. Use **Matches** activity below Read PDF Text activity
    a. In RegEx column, select Email.
6. Use **For Each** activity to iterate through each email item and store it line by line in a MS Word file using **Type Into** activity.
7. Use another **Matches** activity below previous Matches activity.
    a. In RegEx column, select Advanced.
    b. In Value column, enter the expression: (407)([0-9])
8. Use **For Each** activity to iterate through each phone number item and store it line by line in MS Word file using Type Into activity.
9. Display **message** with "Done ! There are แสดงจำนวนอีเมล emails and แสดงจำนวนเบอร์โทรศัพท์ phone numbers.
10. STOP

Outcome

- All the texts from the PDF file is stored in Word saved with “question_studentID.docx”.