blueprism
The Operational Agility Software Company

Interfacing with PDF Documents

Introduction

Blue Prism 을 사용하여 PDF 문서에서 텍스트를 추출하는데 사용할 수 있는 기술이 많이 있습니다. 사용 가능한 기술은 다음과 같습니다.

1. Windows 클립 보드를 사용하여 pdf 문서의 모든 텍스트 복사
2. Blue Prism 'Read Text with OCR' Read 스테이지 액션을 사용하여 PDF 문서 내의 영역에서 텍스트 읽기
3. Adobe Acrobat API 를 사용하여 pdf 를 다른 형식(XML 또는 Microsoft Word)으로 내보내 데이터가 텍스트를 더 쉽게 추출할 수 있습니다.

PDF 문서와의 인터페이스를 고려하고 있다면 다음과 같은 조언을 고려해야 합니다.

* Blue Prism 은 문서 구조가 표준이고 변형이 없거나 매우 제한된 수의 예측이 가능한 경우에만 PDF 에서 데이터를 추출하는 데 사용해야 합니다.
* Blue Prism 은 다양한 형식의 문서에서 데이터를 추출하도록 설계된 전용 OCR 솔루션을 대체하지 않습니다.
* Blue Prism 은 문서에서 손으로 쓴 텍스트를 추출하는 기능이 없습니다.
* OCR 대신 일부 Blue Prism 클라이언트에서 사용하는 100% 신뢰할 수 있는 대안으로 PDF 문서에서 데이터를 Blue Prism 솔루션에 대한 입력으로 사용할 수 있는 구조화된 형식으로 수동으로 추출하는 소수의 직원을 운용하는 것입니다.

Types of PDF Documents

PDF 문서에는 두 가지 주요 유형이 있습니다.

PDF Documents

이러한 PDF 문서는 일반적으로 Microsoft Word 또는 Adobe Acrobat 을 사용하여 작성되며 read only.pdf 형식으로 저장됩니다. Windows 클립 보드를 사용하여 문서에서 텍스트를 복사하여 문서가 진정한 PDF 문서인지 테스트할 수 있습니다.

이러한 '진정한' PDF 문서의 경우 이 가이드에 설명된 기술을 사용하여 데이터를 추출할 수 있습니다.

PDF Images

종종 스캔된 문서는 .pdf 또는 .tiff 형식 이미지로 저장됩니다. 이러한 이미지에서 텍스트를 복사하는 것은 불가능합니다.

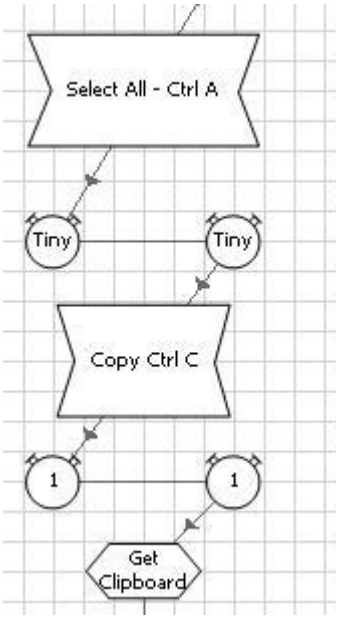
이러한 이미지의 경우 'Reading Text with OCR' 기술만 데이터를 추출하는 데 사용할 수 있습니다. OCR 은 이미지 품질이 충분히 높은 경우에만 작동하며 최소 300dpi 를 권장합니다. Blue Prism 에서 사용하는 Tesseract OCR 엔진은 손으로 쓴 텍스트를 읽는 데 사용할 수 없습니다.

Using Windows Clipboard

이 기술을 사용하기 전에 다음 교육을 받는 것이 좋습니다.

- Surface Automation Training

이 기술을 사용하려면 Object Studio 에서 다음을 수행하는 간단한 Windows 인터페이스 객체를 만들어야 합니다.



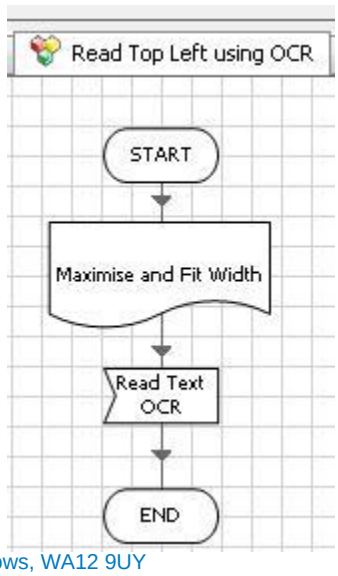
- Adobe Reader 에 표시될 문서를 기동하거나 접속합니다.
- 문서 내부를 클릭하고 Ctrl 과 키 입력을 사용하여 문서 내의 모든 텍스트를 선택합니다.
- Ctrl 및 c 키 입력을 사용하여 선택한 텍스트를 Windows 클립 보드에 복사합니다.
- 계산 스테이지에서 GetClipboard() 함수를 사용하여 PDF 텍스트를 Blue Prism 으로 가져옵니다.

Using Read Text with OCR

이 기술을 시도하기 전에 다음 교육을 받는 것이 좋습니다.

- Surface Automation Training
- Guide to Reading Text with OCR

이 기술을 사용하려면 애플리케이션 모델러의 영역 편집기를 사용하여 텍스트로 읽고자 하는 문서의 영역을 Blue Prism 에게 알려주어야 합니다.



인터페이스에는 다음 스테이지가 포함되어야 합니다.

- Adobe Reader 에 표시될 문서를 기동하거나 접속합니다.
- 인터페이스를 더 표준화하기 위해 문서를 최대화하거나 크기를 조정합니다.
- OCR 기능이 있는 텍스트 읽기를 사용하여 해당 지역의 텍스트를 읽습니다.

> * 참고: Blue Prism 내의 OCR 기능은 큰 문서 영역이 아닌 작은 영역에서 사용할 때 가장 잘 작동합니다.

Using the Adobe Acrobat API

이 기술을 시도하기 전에 다음 데이터 시트를 읽는 것이 좋습니다.

- Blue Prism Data Sheet - Extending Automations using the .NET Framework

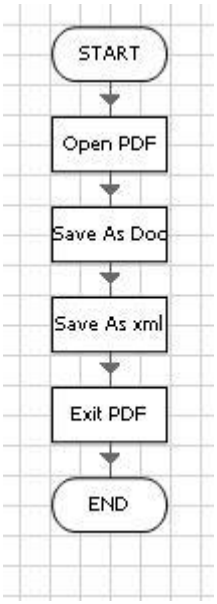
Adobe Acrobat Standard 또는 Professional 이 리소스 PC 에 설치된 경우 Acrobat API 를 사용하여 PDF 문서를 XML 또는 Microsoft Word 와 같은 다른 형식으로 내보내거나 변환할 수 있습니다.

문서를 XML 또는 Word 문서로 저장하는 것은 표와 같은 더 복잡한 문서 구조에서 더 많은 데이터를 추출해야 하는 더 복잡한 문서의 경우 선호될 수 있습니다.

XML 또는 Word .doc 파일로 저장되면 대체 인터페이스를 사용하여 문서와 상호 작용할 수 있습니다(예: Blue Prism MS Word VBO 비즈니스 객체).

Adobe Acrobat API 를 사용하려면 다음 단계를 수행해야 합니다.

- PC 에 Adobe Acrobat Standard 또는 Professional 을 설치해야 합니다. 이에 대한 라이센스 비용이 있으므로 필요한 경우 최소 수의 로봇이 PDF 문서와 상호 작용할 수 있도록 솔루션을 설계하는 것이 좋습니다.
- Acrobat API 를 사용하는 코드 스테이지가 필요합니다.
- API 내의 저장 및 다른 이름으로 저장 기능을 사용하면 PDF 를 다른 형식으로 저장할 수 있습니다.



PDF created with Accessibility Features

PDF 문서는 접근성을 염두에 두고 Adobe Acrobat 을 사용하여 만들 수 있습니다. 양식 및/또는 태그와 같은 Acrobat 의 기능을 사용하여 만든 문서는 Blue Prism Active Accessibility 인터페이스를 사용하여 인터페이스할 수 있습니다.

인터페이스해야 하는 PDF 문서가 조직 내부에서 생성된 경우 접근성 기능을 사용하여 Robotic Process Automation 을 더 쉽게 만들 수 있는지 문서 소유자와 논의하는 것이 좋습니다.

Extracting data from text

위에 설명된 기술 중 하나를 사용하여 PDF 문서 텍스트를 캡처한 후에도 텍스트 내에서 원하는 데이터를 추출하는 몇 가지 논리를 구현해야 할 수 있습니다.

For example:

'Read Text with OCR' 기능을 사용하여 구매 주문서의 왼쪽 상단 영역에서 아래 텍스트를 캡처했습니다.

The following number must appear on all
related correspondence, shipping papers,
and invoices:
P.O. NUMBER: 00012345678
TO:
Mr J Bloggs
Wigits R Us
202 Factory Street
Glasgow G5 5CD
Phone 0330 200 3000

PDF 에서 가져온 위의 텍스트에서 우리가 자동화하고 있는 비즈니스 프로세스에 사용하기 위해 숫자 '00012345678'을 추출하려고 합니다.

캡처된 PDF 텍스트에서 원하는 데이터를 추출하는 데 사용할 수 있는 두 가지 방법이 있습니다.

- InStr, Left, Mid 등과 같은 텍스트 표현을 사용하는 계산 스테이지;

  PO 번호를 추출하는 이 예에서는 InStr 표현식을 사용하여 "P.O. Number:" 텍스트와 그 뒤의 다음 줄 바꿈 문자를 찾은 다음 Mid 표현식을 사용하여 PO 번호를 데이터 항목으로 잘라낼 수 있습니다.

  예를 들어 다음 표현식은 "P.O. NUMBER" 텍스트의 위치를 반환합니다.

  ```
  InStr([Purchase Order Text],"P.O. NUMBER:")
  ```

  Blue Prism 에서 자신만의 계산식을 작성하는 방법을 배우려면 계산 스테이지 속성 창에서 Expression Function Builder 및 Evaluate Expression 기능을 사용하는 것이 좋습니다.

- Regular Expressions

  PO 번호를 추출하는 이 예에서는 "P.O. Number" 텍스트 뒤의 첫 번째 숫자 필드를 검색할 수 있는 정규식을 사용할 수 있습니다.

  예를 들어 다음 정규식은 "P.O. NUMBER"를 포함하는 텍스트 줄을 반환합니다.

  ```
  (?:P(.?)*O(.?)*NUMBER).*
  ```

  자신만의 정규식을 만드는 방법을 배우려면 http://regexr.com/ 웹 사이트를 추천합니다. 여기에는 정규식의 두 가지 예가 모두 포함되어 있으며 자신의 텍스트를 붙여 넣고 정규식 구문을 실험하여 원하는 결과를 얻을 수 있습니다.

* 참고: 일부 기술 개발자는 코드 스테이지의 텍스트에서 데이터를 추출하기위한 논리를 만들 수 있습니다. 이것이 유효한 방법이지만 조직에서 생성하는 맞춤형 코드 스테이지의 향후 지원 용이성을 고려해야 합니다.

Testing your PDF data extraction Logic

많은 수의 PDF 문서에 대해 솔루션을 테스트해야 합니다.

텍스트를 추출하기 위해 개발된 논리는 PDF 문서 구조의 예측 가능성에 따라 다르므로 데이터 추출을 위해 개발한 논리는 가능한 한 많은 PDF 문서 예제를 사용하여 테스트하는 것이 좋습니다.