scanned PDF to XML in VB.NET using ByteScout PDF Extractor SDK

scanned PDF to XML in VB.NET

Today we will explain the steps and algorithm of implementing scanned PDF to XML and how to make it work in your application. ByteScout PDF Extractor SDK was made to help with scanned PDF to XML in VB.NET. ByteScout PDF Extractor SDK is the SDK that helps developers to extract data from unstructured documents, pdf, images, scanned and electronic forms. Includes AI functions like automatic table detection, automatic table extraction and restructuring, text recognition and text restoration from pdf and scanned documents. Includes PDF to CSV, PDF to XML, PDF to JSON, PDF to searchable PDF functions as well as methods for low level data extraction.

VB.NET code snippet like this for ByteScout PDF Extractor SDK works best when you need to quickly implement scanned PDF to XML in your VB.NET application. To do scanned PDF to XML in your VB.NET project or application you may simply copy & paste the code and then run your app! Code testing will allow the function to be tested and work properly with your data.

Trial version can be downloaded from our website. Source code samples for VB.NET and documentation are included.

VB.NET - Program.vb

```
Imports Bytescout.PDFExtractor

' This example demonstrates the use of Optical Character Recognition (OCR) to extract
text into xml
' from scanned PDF documents and raster images.

' To make OCR work you should add the following references to your project:
' "Bytescout.PDFExtractor.dll", "Bytescout.PDFExtractor.OCRExtension.dll".

Class Program

    Friend Shared Sub Main(args As String())

        ' Create Bytescout.PDFExtractor.XMLExtractor instance
        Dim extractor As New XMLExtractor()
        extractor.RegistrationName = "demo"
        extractor.RegistrationKey = "demo"

        ' Load sample PDF document
        extractor.LoadDocumentFromFile("sample_ocr.pdf")

        ' Enable Optical Character Recognition (OCR)
        ' in .Auto mode (SDK automatically checks if needs to use OCR or not)
```

```
        extractor.OCRMode = OCRMode.Auto

        ' Set the location of OCR language data files
        extractor.OCRLanguageDataFolder = "c:\Program Files\Bytescout PDF Extractor
SDK\ocrdata"

        ' Set OCR language
        extractor.OCRLanguage = "eng"  ' "eng" for english, "deu" for German, "fra"
for French, "spa" for Spanish etc - according to files in "ocrdata" folder
        ' Find more language files at https://github.com/bytescout/ocrdata

        ' Set PDF document rendering resolution
        extractor.OCRResolution = 300


        ' You can also apply various preprocessing filters
        ' to improve the recognition on low-quality scans.

        ' Automatically deskew skewed scans
        'extractor.OCRImagePreprocessingFilters.AddDeskew()

        ' Remove vertical or horizontal lines (sometimes helps to avoid OCR engine's
page segmentation errors)
        'extractor.OCRImagePreprocessingFilters.AddVerticalLinesRemover()
        'extractor.OCRImagePreprocessingFilters.AddHorizontalLinesRemover()

        ' Repair broken letters
        'extractor.OCRImagePreprocessingFilters.AddDilate()

        ' Remove noise
        'extractor.OCRImagePreprocessingFilters.AddMedian()

        ' Apply Gamma Correction
        'extractor.OCRImagePreprocessingFilters.AddGammaCorrection()

        ' Add Contrast
        'extractor.OCRImagePreprocessingFilters.AddContrast(20)


        ' (!) You can use new OCRAnalyzer class to find an optimal set of image
preprocessing
        ' filters for your specific document.
        ' See "OCR Analyser" example.


        ' Save extracted text to file
        extractor.SaveXMLToFile("output.xml")

        ' Cleanup
        extractor.Dispose()

        ' Open output file in default associated application
        System.Diagnostics.Process.Start("output.xml")

    End Sub

End Class
```

FOR MORE INFORMATION AND FREE TRIAL:

Download Free Trial SDK (on-premise version)

Read more about ByteScout PDF Extractor SDK

Explore documentation

Visit www.ByteScout.com

or

Get Your Free API Key for www.PDF.co Web API