www.bytescout.com

How to read PDF with scanned image and hindi text in C# using ByteScout PDF Suite

How to write a robust code in C# to read PDF with scanned image and hindi text with this step-by-step tutorial

The documentation is designed for a specific purpose to help you to apply the features on your side. ByteScout PDF Suite is the bundle that provides six different SDK libraries to work with PDF from generating rich PDF reports to extracting data from PDF documents and converting them to HTML. This bundle includes PDF (Generator) SDK, PDF Renderer SDK, PDF Extractor SDK, PDF to HTML SDK, PDF Viewer SDK and PDF Generator SDK for Javascript. It can read PDF with scanned image and hindi text in C#.

The following code snippet for ByteScout PDF Suite works best when you need to quickly read PDF with scanned image and hindi text in your C# application. Simply copy and paste in your C# project or application you and then run your app! Want to see how it works with your data then code testing will allow the function to be tested and work properly.

You can download free trial version of ByteScout PDF Suite from our website with this and other source code samples for C#.

FOR MORE INFORMATION AND FREE TRIAL:

Download Free Trial SDK (on-premise version)

Read more about ByteScout PDF Suite

Explore API Documentation

Get Free Training for ByteScout PDF Suite

Get Free API key for Web API

visit www.ByteScout.com

Source Code Files:

Program.cs

```
using System;
using System.Collections.Generic;
using System.IO;
using System.Text;
using System.Drawing.Imaging;
using Bytescout.PDFExtractor;
using Bytescout.PDF;
using System.Diagnostics;

namespace ReadPDFWithImageHindiText
{
    class Program
    {
        static void Main(string[] args)
        {
            try
            {
                // Files
                string fileName = "hindi_text_with_image.pdf";
                string destFileName = "output_hindi_text_with_image.pdf";
                string destFileName_serachable = "output_hindi_text_with_image_searchab

                // Read all text from pdf file
                string allTextExtracted = "";
                using (TextExtractor extractor = new TextExtractor())
                {
                    // Load PDF document
                    extractor.LoadDocumentFromFile(fileName);

                    // Read all text directly
                    allTextExtracted = extractor.GetText();
                }

                // Get image from pdf file
                MemoryStream memoryStream = new MemoryStream();
                using (ImageExtractor extractor = new ImageExtractor())
                {
                    // Load PDF document
                    extractor.LoadDocumentFromFile(fileName);

                    if (extractor.GetFirstImage())
                    {
                        extractor.SaveCurrentImageToStream(memoryStream, ImageFormat.Pn
                    }
                }

                // Load image from file to System.Drawing.Image object (we need it to g
                using (System.Drawing.Image sysImage = System.Drawing.Image.FromStream(
                {
                    // Compute image size in PDF units (Points)
                    float widthInPoints = sysImage.Width / sysImage.HorizontalResolutio
                    float heightInPoints = sysImage.Height / sysImage.VerticalResolutio
```

```
            // Create new PDF document
            using (Document outPdfDocument = new Document())
            {
                outPdfDocument.RegistrationName = "demo";
                outPdfDocument.RegistrationKey = "demo";

                // Create page of computed size
                Page page = new Page(widthInPoints, heightInPoints);

                // Add page to the document
                outPdfDocument.Pages.Add(page);

                Canvas canvas = page.Canvas;

                // Create Bytescout.PDF.Image object from loaded image
                Image pdfImage = new Image(sysImage);

                // Draw the image
                canvas.DrawImage(pdfImage, 0, 0, widthInPoints, heightInPoints)

                // Dispose the System.Drawing.Image object to free resources
                sysImage.Dispose();

                // Create brush
                SolidBrush transparentBrush = new SolidBrush(new ColorGray(0))

                // ... and make it transparent
                transparentBrush.Opacity = 0;

                // Draw text with transparent brush
                // Need to set Font which supports hindi characters.
                Font font16 = new Font("Arial Unicode MS", 16);
                canvas.DrawString(allTextExtracted, font16, transparentBrush, 4

                // Save document to file
                outPdfDocument.Save(destFileName);
            }
        }


        // Make PDF file with hindi text searchable to OCR.
        using (SearchablePDFMaker searchablePDFMaker = new SearchablePDFMaker()
        {
            //Load PDF document
            searchablePDFMaker.LoadDocumentFromFile(destFileName);

            // Set the location of "tessdata" folder containing language data f
            /*
             * It used following files for hindi language support. Need to put
             * https://github.com/tesseract-ocr/tessdata/tree/3.04.00
             hin.traineddata
             hin.cube.bigrams
             hin.cube.lm
             hin.cube.nn
             hin.cube.params
             hin.cube.word-freq
             hin.tesseract_cube.nn
             */
            searchablePDFMaker.OCRLanguageDataFolder = @"c:\Program Files\Bytes
```

```
                    // Set OCR language
                    searchablePDFMaker.OCRLanguage = "hin";

                    // Need to set Font which supports hindi characters
                    searchablePDFMaker.LabelingFont = "Arial Unicode MS";

                    // Set PDF document rendering resolution
                    searchablePDFMaker.OCRResolution = 300;

                    searchablePDFMaker.MakePDFSearchable(destFileName_serachable);
                }

                // Open document in default PDF viewer app
                Process.Start(destFileName_serachable);
            }
            catch (Exception ex)
            {
                Console.WriteLine("ERROR:" + ex.Message);
            }

            Console.ReadLine();
        }
    }
}
```

ReadPDFWithImageHindiText.sln

```
Microsoft Visual Studio Solution File, Format Version 12.00
# Visual Studio 15
VisualStudioVersion = 15.0.27703.2026
MinimumVisualStudioVersion = 10.0.40219.1
Project("{FAE04EC0-301F-11D3-BF4B-00C04F79EFBC}") = "ReadHindiText", "ReadPDFWithImageH
EndProject
Global
        GlobalSection(SolutionConfigurationPlatforms) = preSolution
                Debug|Any CPU = Debug|Any CPU
                Release|Any CPU = Release|Any CPU
        EndGlobalSection
        GlobalSection(ProjectConfigurationPlatforms) = postSolution
                {99735776-2956-463D-9795-EBCE16928C30}.Debug|Any CPU.ActiveCfg = Debug
                {99735776-2956-463D-9795-EBCE16928C30}.Debug|Any CPU.Build.0 = Debug|An
                {99735776-2956-463D-9795-EBCE16928C30}.Release|Any CPU.ActiveCfg = Rele
                {99735776-2956-463D-9795-EBCE16928C30}.Release|Any CPU.Build.0 = Releas
        EndGlobalSection
        GlobalSection(SolutionProperties) = preSolution
                HideSolutionNode = FALSE
        EndGlobalSection
        GlobalSection(ExtensibilityGlobals) = postSolution
                SolutionGuid = {3ABE3EEF-B212-4E8B-9A74-67A52FD333AC}
```

```
        EndGlobalSection
EndGlobal
```

VIDEO

https://www.youtube.com/watch?v=NEwNs2b9YN8

ON-PREMISE OFFLINE SDK

60 Day Free Trial or Visit ByteScout PDF Suite Home Page
Explore ByteScout PDF Suite Documentation
Explore Samples
Sign Up for ByteScout PDF Suite Online Training

ON-DEMAND REST WEB API

Get Your API Key
Explore Web API Docs
Explore Web API Samples

visit www.ByteScout.com

visit www.PDF.co