HOWTO: How to OCR a PDF

NOTE From Support:

This article has been flagged for review. It contains possibly outdated information.

You may wish to review the Searchable PDF demo as it contains correct/tested code for this use case

Original Article Content:

The OCR process is most efficient when you use a class derived from ImageSource that lazily loads each image one at a time, so that all of the pages of the document are not kept in memory. For PDF documents, we have created PdfImageSource, which you will find in the PDF Reader add-on, in the Atalasoft.Imaging.ImageSources namespace. It has the following features:

1. Lazy loads each page on request
2. Extracts the exact image from the page if the page is a single image (like from a scanned document)
3. Rasterizes pages that are not a single image An instance of this class can be passed to Translate() and Recognize() on any OcrEngine. This assumes that the OcrEngine has been initialized and that it supports searchable PDF Translation.

C# Sample Code:

```
   publicvoid TranslatePdftoSearchablePdf(OcrEngine ocrEng, String pdfIn, String
searchablePdfOut)
  {
      using (Stream pdfStream = File.OpenRead(pdfIn))
    {
         using (PdfImageSource pdfSource = newPdfImageSource(pdfStream))
       {
        ocrEng.Translate(pdfSource, "application/pdf", searchablePdfOut);
       }
    }
  }
```

Original Article:

Q10301 - HOWTO: How to OCR a PDF

Atalasoft Knowledge Base
https://www.atalasoft.com/kb2/KB/50174/HOWTO-How-to-OCR-a-PDF