

I want to extract all the text boxes and text box coordinates from a PDF file. I would like to extract text from a portion (using coordinates) of PDF page, can anyone help me out?

Given a PDF file, output should look something like:

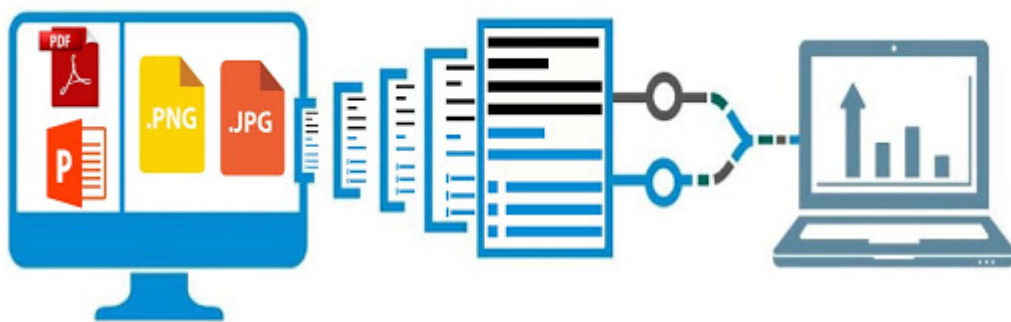
```
489, 41, "Signature"  
500, 52, "b"  
630, 202, "a_g_i_r"
```

Customer #1

Hi,

I was wondering if anyone could recommend a program which can extract the starting (top left) coordinates (x,y) of each word in a PDF file (and the end if possible). Ideally output would be in a format that could be easily inserted into a database.

Customer #2



Data Extraction from Paper

© VeryUtils.com

Sometimes, we have some customers who want to extract text contents and their positions

from PDF pages, the text positions are used to parse the values, such as read invoice numbers from PDF files or looking for some other information.

PDF Extractor SDK (PDF Parser SDK and Command Line) is a good product to extract various information from PDF files, of course, it can extract text contents and text coordinates also.

1. You may download the trial version of PDF Extractor SDK (PDF Parser SDK and Command Line) from this web page first,

<https://veryutils.com/pdf-extractor-sdk-pdf-parser-sdk-and-command-line>

2. After you download it, you may unzip it to a folder.

3. Please run a CMD window first, if you don't know how to run a CMD window, please look at following web page,

<https://veryutils.com/blog/top-10-methods-to-run-a-command-line-window-in-windows-10/>

4. pdfextract.exe is a command line application, it supports following command line options,

```
D:\VeryPDF_PDFExtractTool>pdfextract.exe
```

```
pdfextract.exe version 3.0
```

```
Copyright 1996-2017 VeryPDF.com Inc.
```

```
Product Name: VeryPDF PDF Extract Tool Command Line
```

```
http://www.verypdf.com
```

```
http://www.verydoc.com
```

```
http://support.verypdf.com
```

```
Email: support@verypdf.com
```

```
Usage: pdfextract.exe [options] <PDF-file>
```

```
-f <int>           : first page to print
-l <int>           : last page to print
-opw <string>      : owner password (for encrypted files)
-upw <string>      : user password (for encrypted files)
-outfolder <string>: Set a folder to store extracted files
```

```
-layout           : maintain original physical layout
-textfile         : Extract text contents from PDF file
-textpos         : Extract text and coordinates from PDF file
-nopgbrk         : don't insert page breaks between pages
-h               : print usage information
-help            : print usage information
--help           : print usage information
-?              : print usage information
-$ <string>      : input your license key
```

Example:

```
pdfextract.exe D:\in.pdf
pdfextract.exe -outfolder D:\out\ D:\in.pdf
pdfextract.exe -outfolder D:\out\ D:\in.pdf
pdfextract.exe -opw 123 -upw 456 -outfolder D:\out\ D:\in.pdf
pdfextract.exe -outfolder D:\out\ D:\in.pdf > out.log
pdfextract.exe -outfolder D:\out\ D:\in.pdf out.log
pdfextract.exe D:\in.pdf out.log
pdfextract.exe -textpos D:\in.pdf D:\out.txt
pdfextract.exe -textpos -nopgbrk D:\in.pdf D:\out.txt
pdfextract.exe -textfile D:\in.pdf D:\out.txt
pdfextract.exe -layout -textfile D:\in.pdf D:\out.txt
```

5. You can simple run following command line to extract all information from your PDF file,

```
pdfextract.exe -outfolder D:\VeryUtils\test\ D:\downloads\Test_in.pdf
```

6. You will find a "TextFileWithPosition.txt" file in the "D:\VeryUtils\test" folder, this text file contains all text contents and coordinates for each word, such as,

How to extract text and text coordinates from a PDF file? PDF Parsing with Text and Coordinates. PDF Text Extraction with Coordinates. | 4

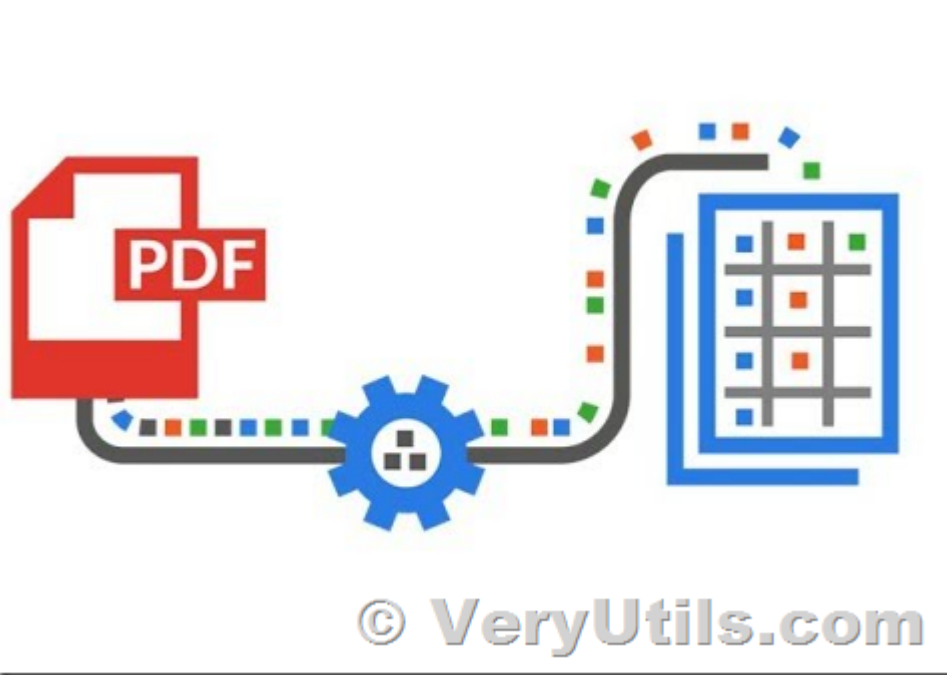
```
1 [Page #1] *** initial words ***
2 word: x=56.80..118.16 y=57.33..72.82 base=69.80 fontSize=14.00 rot=0 link=00000000 'Residence:'
3 word: x=127.70..173.52 y=57.33..72.82 base=69.80 fontSize=14.00 rot=0 link=00000000 'Address'
4 word: x=177.06..202.71 y=57.33..72.82 base=69.80 fontSize=14.00 rot=0 link=00000000 'goes'
5 word: x=206.17..230.26 y=57.33..72.82 base=69.80 fontSize=14.00 rot=0 link=00000000 'here'
6 word: x=56.80..97.15 y=73.43..88.92 base=85.90 fontSize=14.00 rot=0 link=00000000 'Memo:'
7 word: x=127.70..187.55 y=73.43..88.92 base=85.90 fontSize=14.00 rot=0 link=00000000 'Additional'
8 word: x=191.06..257.03 y=73.43..88.92 base=85.90 fontSize=14.00 rot=0 link=00000000 'information'
9 word: x=260.62..286.26 y=73.43..88.92 base=85.90 fontSize=14.00 rot=0 link=00000000 'goes'
10 word: x=289.72..313.82 y=73.43..88.92 base=85.90 fontSize=14.00 rot=0 link=00000000 'here'
11 word: x=56.80..118.16 y=105.63..121.12 base=118.10 fontSize=14.00 rot=0 link=00000000 'Residence:'
12 word: x=127.70..173.52 y=105.63..121.12 base=118.10 fontSize=14.00 rot=0 link=00000000 'Address'
13 word: x=177.06..202.71 y=105.63..121.12 base=118.10 fontSize=14.00 rot=0 link=00000000 'goes'
14 word: x=206.17..230.26 y=105.63..121.12 base=118.10 fontSize=14.00 rot=0 link=00000000 'here'
15 word: x=56.80..97.15 y=121.73..137.22 base=134.20 fontSize=14.00 rot=0 link=00000000 'Memo:'
16 word: x=127.70..187.55 y=121.73..137.22 base=134.20 fontSize=14.00 rot=0 link=00000000 'Additional'
17 word: x=191.06..257.03 y=121.73..137.22 base=134.20 fontSize=14.00 rot=0 link=00000000 'information'
18 word: x=260.62..286.26 y=121.73..137.22 base=134.20 fontSize=14.00 rot=0 link=00000000 'goes'
19 word: x=289.72..313.82 y=121.73..137.22 base=134.20 fontSize=14.00 rot=0 link=00000000 'here'
20 word: x=56.80..118.16 y=153.93..169.42 base=166.40 fontSize=14.00 rot=0 link=00000000 'Residence:'
21 word: x=127.70..173.52 y=153.93..169.42 base=166.40 fontSize=14.00 rot=0 link=00000000 'Address'
22 word: x=177.06..202.71 y=153.93..169.42 base=166.40 fontSize=14.00 rot=0 link=00000000 'goes'
23 word: x=206.17..230.26 y=153.93..169.42 base=166.40 fontSize=14.00 rot=0 link=00000000 'here'
24 word: x=56.80..97.15 y=170.03..185.52 base=182.50 fontSize=14.00 rot=0 link=00000000 'Memo:'
25 word: x=127.70..187.55 y=170.03..185.52 base=182.50 fontSize=14.00 rot=0 link=00000000 'Additional'
26 word: x=191.06..257.03 y=170.03..185.52 base=182.50 fontSize=14.00 rot=0 link=00000000 'information'
27 word: x=260.62..286.26 y=170.03..185.52 base=182.50 fontSize=14.00 rot=0 link=00000000 'goes'
28 word: x=289.72..313.82 y=170.03..185.52 base=182.50 fontSize=14.00 rot=0 link=00000000 'here'
29 word: x=56.80..118.16 y=202.23..217.72 base=214.70 fontSize=14.00 rot=0 link=00000000 'Residence:'
30 word: x=127.70..173.52 y=202.23..217.72 base=214.70 fontSize=14.00 rot=0 link=00000000 'Address'
31 word: x=177.06..202.71 y=202.23..217.72 base=214.70 fontSize=14.00 rot=0 link=00000000 'goes'
32 word: x=206.17..230.26 y=202.23..217.72 base=214.70 fontSize=14.00 rot=0 link=00000000 'here'
33 word: x=56.80..97.15 y=218.33..233.82 base=230.80 fontSize=14.00 rot=0 link=00000000 'Memo:'
34 word: x=127.70..187.55 y=218.33..233.82 base=230.80 fontSize=14.00 rot=0 link=00000000 'Additional'
35 word: x=191.06..257.03 y=218.33..233.82 base=230.80 fontSize=14.00 rot=0 link=00000000 'information'
36 word: x=260.62..286.26 y=218.33..233.82 base=230.80 fontSize=14.00 rot=0 link=00000000 'goes'
37 word: x=289.72..313.82 y=218.33..233.82 base=230.80 fontSize=14.00 rot=0 link=00000000 'here'
```

7. "PageContents.xml" is a XML file which contain coordinates for each character, such as,

Drawing pages 1-1...

```
<tree file="D:\downloads\Test_in.pdf" page="1" mediabox="0 0 612 792" rotate="0">
  <over>
    <over>
      <mask>
        <text font="BAAAAA+TimesNewRomanPSMT" matrix="14 0 0 14" fontsize="14">
          <g c="<0052>" x="56.8" y="722.2" />
          <g c="<0065>" x="66.096" y="722.2" />
          <g c="<0073>" x="72.298" y="722.2" />
          <g c="<0069>" x="77.702" y="722.2" />
          <g c="<0064>" x="81.594" y="722.2" />
          <g c="<0065>" x="88.594" y="722.2" />
          <g c="<006e>" x="94.796" y="722.2" />
          <g c="<0063>" x="101.88" y="722.2" />
          <g c="<0065>" x="108.082" y="722.2" />
          <g c="<003a>" x="114.284" y="722.2" />
        </text>
        <solid colorspace="DeviceRGB" alpha="1" v="0 0 0" />
      </mask>
      <mask>
        <text font="BAAAAA+TimesNewRomanPSMT" matrix="14 0 0 14" fontsize="14">
          <g c="<0041>" x="127.7" y="722.2" />
          <g c="<0064>" x="137.794" y="722.2" />
          <g c="<0064>" x="144.794" y="722.2" />
          <g c="<0072>" x="151.696" y="722.2" />
          <g c="<0065>" x="156.386" y="722.2" />
          <g c="<0073>" x="162.672" y="722.2" />
          <g c="<0073>" x="168.076" y="722.2" />
          <g c="<0020>" x="173.564" y="722.2" />
          <g c="<0067>" x="177.064" y="722.2" />
          <g c="<006f>" x="184.064" y="722.2" />
          <g c="<0065>" x="191.064" y="722.2" />
          <g c="<0073>" x="197.266" y="722.2" />
          <g c="<0020>" x="202.67" y="722.2" />
          <g c="<0068>" x="206.17" y="722.2" />
          <g c="<0065>" x="213.17" y="722.2" />
          <g c="<0072>" x="219.372" y="722.2" />
          <g c="<0065>" x="224.062" y="722.2" />
        </text>
        <solid colorspace="DeviceRGB" alpha="1" v="0 0 0" />
      </mask>
      <mask>
        <text font="BAAAAA+TimesNewRomanPSMT" matrix="14 0 0 14" fontsize="14">
          <g c="<004d>" x="56.8" y="706.1" />
          <g c="<0065>" x="69.204" y="706.1" />
          <g c="<006d>" x="75.49" y="706.1" />
          <g c="<006f>" x="86.186" y="706.1" />
          <g c="<003a>" x="93.27" y="706.1" />
        </text>
```

8. Now, you can write a simple PHP or Python application to read and parse X/Y positions from these PDF files, then you can process these PDF files easily.

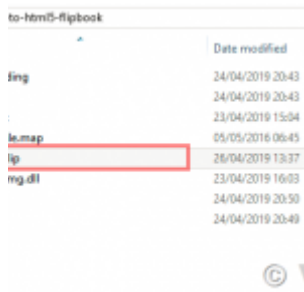


If you wish extract more information from PDF files, such as hyperlinks, colorspace, attachments, bookmarks, pictures, embedded fonts, forms, etc. elements, please feel free to contact us, we are glad to assist you asap,

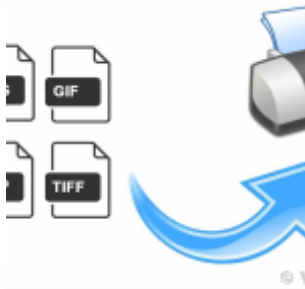
<https://veryutils.com/contact>

Related posts:

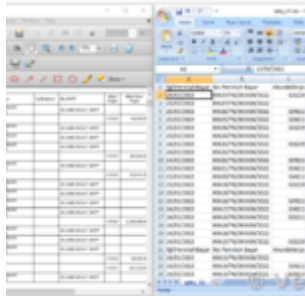
How to extract text and text coordinates from a PDF file? PDF Parsing with Text and Coordinates. PDF Text Extraction with Coordinates. | 7



How to use PDF to HTML5 Flipbook Maker software?



How to use PHP Folder Watcher to monitor a folder and print JPG files to a certain tray of the print...



How to convert text report PDF file to Excel CSV file?

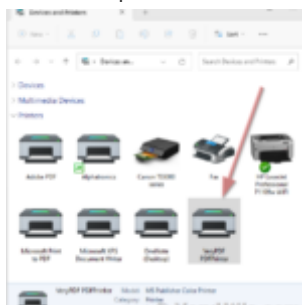
How to extract text and text coordinates from a PDF file? PDF Parsing with Text and Coordinates. PDF Text Extraction with Coordinates. | 8



Best Royalty Free PDF Printer for Developers. Develop your own PDF Printer based on VeryUtils PDF Pr...

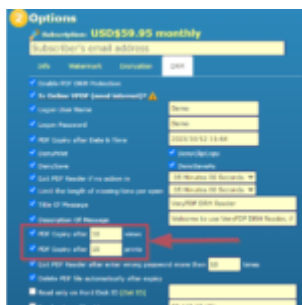


VeryUtils PDF Viewer SDK is a Fully Featured and Customizable PDF Library for Developers



Integration of the Virtual PDF Printer Driver SDK & API into your application Royalty Free

How to extract text and text coordinates from a PDF file? PDF Parsing with Text and Coordinates. PDF Text Extraction with Coordinates. | 9



PDF DRM Protector

Solution for Business:

Safeguard Your PDF

Documents



VeryUtils pdfsearch

Command Line Tool is a

Powerful PDF Search Utility

Related Posts

- pdfsearch is a powerful command line tool to search text in PDF files on Windows system
- Streamlining PDF to Excel Conversion with VeryUtils PDF to Excel Converter Command Line
- Simplify Text Extraction with VeryUtils Text Extraction Command Line Software
- VeryUtils OCR and Data Extraction SDK for C# and VB.NET applications to extract data from PDFs and scanned image files
- VeryUtils Text Extractor Command Line is a Windows console utility that extracts plan text from 200+ file formats
- VeryUtils PDF to Word Converter Command Line for Developers Royalty Free
- How to convert text report PDF file to Excel CSV file?

- [VeryUtils PDF to Text Command Line Extraction](#)
- [OCR TIFF to Text File using VeryUtils ScanOCR software](#)
- [VeryUtils PDF Table Extractor software does extract text columns from selectable or searchable PDF files to tables with CSV and JSON formats easily](#)
- [VeryUtils HTML to PDF Converter Command Line can convert web pages to PDF files from C#, VB.NET, ASP.NET, PHP, etc. program languages](#)
- [jpdfkit is a Java PDF Library to Merge & Split PDF documents](#)
- [How to use PDF to HTML5 Flipbook Maker software?](#)
- [Use PDF Signer Cloud Service to Sign Any Document Online In Seconds](#)
- [Use VeryUtils HotFolder Windows Desktop Application to Automate your workflow](#)
- [VeryUtils AI Writer custom development service to automatically generate unique and best quality blog contents](#)
- [Email Finder is the Fastest Way to Find Email Addresses from websites](#)
- [Boost Your Productivity with VeryUtils Desktop Search: Find Your Local Files Instantly](#)
- [PHP Batch Email Sender – Email Marketing Software – Bulk Email Sender software](#)
- [Use VeryUtils Print Job Tracker software to track all print jobs from Windows system](#)