xdoc2txt - PDF, WORD, EXCEL, the text from a variety of binary documents such as Ichitaro extraction

xdoc2txt xdoc2txt free

Overview

- xdoc2txt is PDF, WORD, EXCEL, from a variety of binary documents such as Ichitaro, is a general-purpose text converter to extract the text element, will work with the Windows command line.
- xdoc2txt because it has to analyze the structure of the various documents directly, you can convert alone. * Such as WORD or Acrobat, you do not need to install the original application.
- Because it operates at a high speed, it is ideal for filter of various full-text search engine.
- Kind of word-processing document, and then determined from the extension. To the next extension of the files are supported.

.rtf	Rich text
.docx	Microsoft WORD 2007/2010/2013 (OOXML)
.xlsx	Microsoft Excel 2007/2010/2013 (OOXML)
.pptx	Microsoft PowerPoint 2007/2010/2013 (OOXML)
.doc	Microsoft WORD ver5.0 / 95/97/2000 / XP / 2003
.xls	Microsoft Excel ver5.0 / 95/97/2000 / XP / 2003
.ppt	Microsoft PowerPoint 97/2000 / XP / 2003

.sxw / .sxc / .sxi / .sxd	OpenOffice.org
.odt / .ods / .odp / .odg	Open Document
.jaw / jtw	Ichitaro ver5
.jbw / juw	Ichitaro ver6
.jfw / jvw	Ichitaro ver7
.jtd / jtt	Ichitaro ver8 / 9/10/11/12
.oas / oa2 / oa3	OASYS / Win
.bun	New pine / pine 5 / pine 6
.wj2 / wj3 / wk3 / wk4 / 123	Lotus 123
.wri	Windows3.1 Write
.pdf	Adobe PDF
.mht	Web Archive
.html	HTML
.eml	Export format of OutlookExpress

- From Ver2.0, it supports iFilter. * Even with the extension that xdoc2txt does not correspond to the native, text extraction can be done if there is a corresponding iFilter. (This function is only exe version)
- exe version, Dll version, there is a COM component version. Function of text extraction is equivalent.

operating environment

xdoc2txt operates in the following environment.

ver	Operating environment
ver1.x	Windows 95/98 / ME / NT4.0 / 2000 / XP / Vista / Windows 7 (32bit / 64bit) / Windows 8 (32bit / 64bit) / Server 2003 / Windows Server

(MBCS)	2008 R2 (64bit) / Windows Server 2012 (64bit) / Windows Server 2012 R2 (64bit)
ver2.x (Unicode)	2000 / XP / Vista / Windows 7 (32bit / 64bit) / Windows 8 (32bit / 64bit) / Windows 10 (32bit / 64bit) / Server 2003 / Windows Server 2008 R2 (64bit) / Windows Server 2012 (64bit) / Windows Server 2012 R2 (64bit)

- To execute Ver 2.0, the following packages need to be installed.

  When executing xdoc2txt 32 bit (x86) version (Windows OS to execute is 32/64 bit):
  Microsoft Visual C ++ 2010 Redistributable Package (x86)

  When executing xdoc2txt 64bit (x64) version:
  Microsoft Visual C ++ 2010 Redistributable Package (x64)

  "The specified program can not be executed." If you receive an error message such as "The application could not be started due to incorrect side-by-side configuration of this application", install the above package from Microsoft's distribution site please.

Copyright and Terms and Conditions

- xdoc2txt the case of a non-profit, it can be used in free. Use of personal and non-profit organization, the

use of the enterprise and corporate internal intranet, if you want to use to build their own Internet publishing server (including the management of the commercial site), does not hit for commercial, available in free I can do it.

- Because if you want to re-distribute incorporated into commercial products is a xdoc2txt you around to a commercial license, please contact the author.

  xdoc2txt commercial license (2007/5/24 edition)

- The xdoc2txt that is included in the Hyper Estraier, if you want to distribute with the Hyper Estraier are excluded from the commercial license.
- If you wish to re-distribution of xdoc2txt, thank you so be sure to take a distribution permit to the author. In the case of free software, it will not allow the principle distribution.
  If you want to re-distribute the xdoc2txt, please distribute without changing the whole file that is included in the package. ***** Also, Please indicate to where the user is found in the manual such as the fact that you are using the xdoc2txt.
  It should be noted that, in the case where the software is a specification that can be built-in the xdoc2txt as an external filter (if you use the xdoc2txt downloaded separately), the contact of the author is absolutely

unnecessary.

- Copyright of xdoc2txt and accompanying documentation are the property of their hishida.
- xdoc2txt is provided as-is with no warranty. Any damages arising from the use or non-use of xdoc2txt respect (lost profits, interruption of business, including the other monetary damage loss of business information), the author does not take any responsibility.
- Post and inclusion to the software of the magazine, to allow the reprint of on the Internet * If you posted as is, please contact us posted about the magazine until the author because it is quite the ex-post reporting.

command options

```
xdoc2txt.exe [options ..] <filename ...> -h help enc
    It does not perform the file size check.

    <Filename> convert the original file name.

    * Wildcard characters (*?) Is usable.

    * In the case of file names that contain spaces,

    * The following options have been deprecated fro
```

How to Use

- The following example, you write the text that is

included in the sample.doc of MS-Word document to the standard output.

```
xdoc2txt sample.doc
```

By redirecting the output destination in the following manner, it can also be saved to a file.

```
xdoc2txt sample.doc> sample.txt
```

Giving the -f option, you can change the output to a file. Extension will automatically .txt.

```
xdoc2txt -f sample.doc sample.xls
```

Because the wildcard *? Can be used, you can text of collectively multiple files.

```
xdoc2txt -f * .xls
```

In the case of Office documents and Ichitaro document (later Ver8), you can view the document properties in the -p option. * Property will be displayed only the items set.

```
xdoc2txt -p manual .doc
                          * Execution result
                          * <Title> KWIC Finder manual </ Title
                          * <Author> hishida </ Author>
```

```
* <Template> Normal.dot </ Template>
* <LastAuthor> hishida </ LastAuthor>
* <RevisionNumber> 1 </ RevisionNumbe
* <AppName> Microsoft Word 9.0 </ App
* <Lastprinted> 2004/03/23 19:39:00 <
* <Created> 2004/03/23 19:35:00 </ Cr
* <LastSaved> 2004/03/23 19:44:00 </
* <PageCount> 1 </ PageCount>
* <WordCount> 21 </ WordCount>
* <CharCount> 121 </ CharCount>
```

- Is protected by a password WORD / EXCEL / PowerPoint / Ichitaro can not be displayed.
- Since then output as a general rule in the order of the stored text in a file, it may be different from the order of the display of the original application.

about the use of the mouse operation

Create a shortcut on the desktop, you can text of the mouse operation.

1. Explorer right button menu → [send (N)] on xdoc2txt.exe in → [Desktop (create shortcut)
2. Right button menu on the icon that has been created on the desktop → [Properties (R)]

3. At the end of the Target (T)], add the -f.
   Example) "C: \ Program Files \ kwic \ xdoc2txt.exe" -f
4. When you drag and drop the file you want to text into the icon, the extension in the same directory you can file a .txt.

Reference article:
Http://Www.Forest.Impress.Co.Jp/article/2003/11/19/xdoc
([Du NEWS of the window])

About iFilter

- In Ver2.0 or later, it supports iFilter. * The -i option, if the iFilter for the extension is available, will give priority to iFilter.
- We are in the process of validating the following iFilter.
  - Ichitaro IFilter 32 bit for OS
  - DocuWorks Content Filter
  - Microsoft Office Filter Pack
  - Adobe Reader 9.5 accessory iFilter
    ※ Adobe Reader 10 that comes after iFilter and, alone has been distributed in the "Adobe PDF IFilter v6.0", "Adobe PDF iFilter 9 for 64-bit platforms," is not available.
- iFilter support is a feature of only exe version. * It can

not be used iFilter in Dll version.

Download

Ver2.x (Unicode version)

New! 2017/07/06

- xdoc2txt 2.16.1 (xd2tx2161_x64.zip) - x64 (64 bit) version
- xdoc2txt 2.16.1 (xd2tx2161.zip) x86 (32 bit) version

Ver1.x (MBCS version)

- xdoc2txt 1.52 (d2txt152.zip)
- xdoc2txt 1.51 (d2txt151.zip)
- xdoc2txt 1.50 (d2txt150.zip)

- Ver1.00 Cryptlib.Dll (Crypt100.Lzh / 37KB) - encrypted (not required xdoc2txt 2.0 or later) additional DLL in order to search and display the PDF without a password

Filter Case Study

name	Kind	Genre	URL
GoogleXdoc (Incorporating the xdoc2txt to GoogleDeskTop	free	Full-text search	http://softfarm.net/ Soft farm

PlugIn)			
Namazu for Win32	free	Full-text search	Sample of the document filters using xdo http://www.geocities.co.jp/SiliconValley- Full-text search system Namazu for Win3 http://www.namazu.org/windows/
Hyper Estraier	free	Full-text search	http://hyperestraier.sourceforge.net/
Meadow2	free	editor	http://www.bookshelf.jp/pukiwiki/pukiwi Meadow%20memo%20Wiki Meadow memo Wiki
MiGrep	free	Search	http://homepage3.nifty.com/m-and-i/freet M & I page of
VxEditor	free	editor	http://homepage3.nifty.com/x-labo/ X-Labo WebPage
smoopy	free	Text vertical writing viewer	http://www.vector.co.jp/soft/win95/util/se
Transwise	free	Translation support	http://www6.ocn.ne.jp/~vmel/software/T
EBView	free	Dictionary-text search	http://ebview.sourceforge.net/
Search Cross	Product	Full-text search	http://www.villagecenter.co.jp/soft/search Village Center Co., Ltd.
KOA Direct Server	free (some fee required)	Content sharing system	http://koaproject.sakura.ne.jp/pages/koad KOA Project
HNXgrep	free	Grep Search	http://www.vector.co.jp/soft/winnt/util/se

* Of the software that can be used to xdoc2txt as a filter,

which the author knows.

History

Ver2.x (Unicode version)

2.16.1	2017/07/06	• 64 bit (x64) version added. * To run, Microsoft Visual C ++ 2010 Redistributable Package (x64) is required.
2.16.1	2016/06/28	• Delete dependency of VC90.CRT from manifest
2.16	2016/04/26	• Fixed an issue where some will be displayed in the When you use the control format control information in xlsx
2.15	2016/04/07	• Fixed an issue where the control characters that do not appear in the text is displayed in docx
2.14	2015/11/19	• Fixed an issue where the abnormal termination in part of the PDF
2.13	2015/8/25	• Fixes issue where there is a character with a transparent text of Scan Snap is missing • Fixed an issue you can not read encrypted PDF in the part of the 128bit-AES
2.12	2015/7/18	• Adjustment of character between the parameters of the English • Change the compiler in C ++ 2010 from Visual C ++

		2008
2.11	2015/5/29	• EXCEL2007 performance improvement of text extraction from the format (.xlsx)
2.10	2015/4/15	• bug fixes after ExtractText () file has not been released in the com version PDF • Add a function ExtractTextEx () to com version and Dll version. Command-line options use allowed. (Valid only -r -o -g -x)
2.09	2015/4/09	• Case fixes that a ligature can not be displayed in PDF • Fixed an issue where the abnormal termination in part of the .mht • Add the sample program FileFind using Dll
2.08	2015/3/11	• If skip extension is not the contents of the PDF in .pdf • Improve memory leak at the time of PDF extraction • Explicit cdecl to sample program that calls the Dll
2.07	2014/10/28	• Fixed an issue where the abnormal termination in a particular xlsx
2.06	2014/10/09	• bug fixes that can not be extracted character and change the text color only part of the string in a cell in the xlsx
2.05	2014/08/31	• As much as possible avoid the abnormal termination in the corrupt PDF
		• Fixed an issue that part of the Japanese e-mail are

2.04	2014/07/29	garbled in .eml • Command version is 256MB the input file size limit (-z = can be set in)
2.03	2014/07/16	• Fixed an issue where some .xlsx to fail
2.02	2014/06/14	• [Recommended if you want to continuously use sample added to dynamically load and release the Dll version of LoadLibrary and FreeLibrary
2.02	2014/05/04	• Fixed an issue where the abnormal termination in part of the PDF
2.01	2014/02/16	• Half-width Kana correspondence of EUC code
2.00	2013/01/23	• Official version
2.00β4	2012/12/28	• bug fixes
2.00β3	2012/12/24	• Fixed an issue where the new line of standard output had become to \ r \ r \ n • Fix a problem that when you use the -f option in the non-writable directory to abnormal termination • Fixed an issue that part of the ruby of the Word document does not appear • Fixed an issue that part of the character of the output the PDF can not be extracted with Word2007
2.00β2	2012/12/19	• bug fixes that abnormal termination in the part of the .odt • bug fixes that abnormal termination in the doc of

		length 0
2.00β1	2012/12/01	• Add version resource to xdoc2txt.exe (to distinguish it from the 1.x series)
2.00β0	2012/11/26	• Fixed an issue to freeze part of the PDF
2.00α3	2012/11/17	• Office2007 / 2010 Fixed an issue where the entity reference has not been interpreted in a document
2.00α2	2012/11/15	• VC ++ sample added, the argument order of DLL version of the same as the COM version.
2.00α1	2012/11/14	• Add a COM component version (xd2txcom.dll)
2.00α0	2012/11/13	• It performs an internal Unicode reduction. ***** Change the compiler from VC ++ 6.0 to VC ++ 2008. • -u (UTF16) in output options, - Add 8 (UTF8) • Corresponding to iFilter (-i option). ***** Even with the extension that xdoc2txt does not correspond to the native, text extraction can be performed if there is a corresponding iFilter. • / Corresponds to the encoded PDF in LZWDecode (for Unisys patent has expired). • Integrated cryptlib.dll, corresponding standard with encrypted PDF with password-free. • It provides a DLL version (xd2txlib.dll). ***** Attach a sample to call from C # and VB.Net

Ver1.x (MBCS version)

Development of MBCS version (Ver1.x) has ended.
Please use the Ver2.x system in the future.

1.52	2015/11/19	• Fixed an issue where the abnormal termination in part of the PDF
1.51	2015/8/25	• Fixes issue where there is a character with a transparent text of Scan Snap is missing • Fix some of the problems you can not read encrypted PDF with 128bit-AES (required cryptlib.dll)
1.50 1.50	2014/10/28	• Fixed an issue where the abnormal termination in a particular xlsx
1.49 1.49	2014/10/09	• bug fixes that can not be extracted character and change the text color only part of the string in a cell in the xlsx
1.48	2014/05/04	• Fixed an issue where the abnormal termination in part of the PDF
1.47	2013/11/30	• bug fixes and abnormal termination to contain the long link in a Word document
1.46	2012/12/24	• Fixed an issue that part of the ruby of the Word document does not appear • Fixed an issue that part of the character of the output the PDF can not be extracted with Word2007
1.45	2012/11/26	• Fixed an issue to freeze part of the PDF

1.44	2012/11/17	• Corresponding to the garbled some of the characters in the docx
1.43	2012/10/17	• Fixed the problem that if there is to be "Office Open XML File Formats" abnormal termination in .xlsx that do not conform to
1.42	2012/05/16	• Fixed an issue where if there is a crash in the part of the PDF
1.41	2011/07/31	• bug fixes that the last character is lacking in EXCEL of the text box (bug that is mixed with 1.37)
1.40	2011/05/17	• Fixed the problem that if there is the display of the text is missing in some of the PDF.
1.39	2011/04/28	• Fixed an issue where extra characters in the part of the PDF is displayed (/ document the language specified is used by Lang)
1.38	2010/12/21	• Abnormal termination to fix a problem in the (part of the conditions in the case of PDF1.5 or later / XRef is used) part of the PDF
1.37	2010/05/16	• Deal with the problem that PDF output in the part of the PDF writer (Brava! Desktop) is garbled • Fixed a case where dust enters in EXCEL of the text box
1.36	2010/01/09	• Fixed an issue where the EUC encoding PDF is garbled

1.35	2009/08/28	• Corresponding to the air of Office2007 document
1.34	2009/06/22	• Office2007 document dated password is to be displayed with the "encrypted file.". • zlib.dll from this version is not required (changes to the static link)
1.33	2009/06/07	• Fixed a case to be abnormally terminated with corrupt PDF • Fixed a problem that garbled in OpenOffice documents with password (indicated as "encrypted file.") • Fixed a case where the number and the number row seat to fail in the extremely large Excel2007 document. • Add the -x option. ***** Display only the cells that exist in EXCEL2007
1.32	2008/12/01	• Corrupted Fixed some cases result in an infinite loop in PDF • After Acrobat7.0, PDF of reading correspondence that has been encrypted with 128bit AES (necessary to introduce a cryptlib.dll)
1.31	2008/11/05	• Fixed a problem that garbled in corrupted Office2007 file. • Fixed an issue where the number of sheets is completed abnormal when it exceeds around 100 in Excel2007
1.30 R2	2008/08/18	• Add the AtiveX version xdoc2txt.ocx. Distribution conditions are the same as exe version.
1.30	2008/05/22	• -p option: Add the "company name", "classification", "administrator name" in the property view of the Office document

1.29	2008/05/18	• Fixed an issue where if there is a crash in the PDF file that has been created in the PDF creation software other than Acrobat • Fixed an issue that is abnormally terminated with 0 bytes of PDF size
1.28	2008/03/18	• Corresponding to a PDF that was created in PDFMaker8.1
1.27	2008/01/24	• Fixed an issue where the path name of the input file is more than 256 bytes "error in the file name" and can not be processed are displayed
1.26a	2007/10/21	• Modify the new line is 0x0D, there is a case of a buffer overrun in the HTML not 0x0A
1.26	2007/05/11	• bug fixes that part of the column does not appear in the Microsoft Office Excel2007 • bug fixes 2 reviews about PDF (display leakage, abnormal termination measures)
1.25	2007/08/13	• Corresponding to the display of the insertion field name in Microsoft Word
1.24	2007/02/18	• Microsoft Office Word2007 / Excel2007 / PowerPoint2007, OpenOffice.org, corresponding to the Open Document • bug fixes abnormal completion, with a number of the large number of digits as 1E + 275 in EXCEL.
		• Corresponding to a problem that can not be text

1.23	2006/08/29	extraction from PDF that was created in AntenaHouse PDF Driver2.0 (Corresponding to the PDF1.5 or later Cross-Reference Streams) • By PDF, fix the problems that there is a case that can not be text extraction to the end of the file
1.22	2006/05/28	• bug fixes by PDF encoding " "" is garbled
-	2006/05/10	• Terms and conditions change for commercial use
1.21	2006//05/08	• bug fixes When you search for a password with a document of Ichitaro Ver6 out of memory
1.20	2006/02/17	• Corresponding to the Unicode mapping ligatures (ff, fi, etc.) in PDF
1.19	2006/02/08	• Preventive modification of the buffer overrun in PDF
1.18	2006/02/04	• PowerPoint95 correspondence • bug fixes that there is that the contents of the line does not appear all in EXCEL
1.17	2005/09/19	• Add the character spacing adjustment parameters -g of PDF
		• If the huge figure in the PDF has been compressed with / FlateDecode, fix the bug that fail to allocate memory • Additional PDF options -o = 0 in the PDF -? - Do not display the form page number of

1.16	2005/05/02	-o = 1 Delete the line breaks in PDF (If the vertical writing is a new line for every one letter) • HTML ruby output options -r = 0 None -r = 1 () -r = 2 "" Aozora Bunko format • Fixed a bug that space disappears immediately after the tag in the text of the HTML
1.15	2005/04/23	• Modify some text of which can not be the case in a PDF that was created in Acrobat4 • Fixed a bug that with the PDF can not text the stamp even once in Acrobat
1.14	2005/01/31	• bug fixes that abnormal termination in the PDF that was created in Justsystem PDF Creator • bug fixes that image only a case of abnormal termination with no text PDF at
1.13	2004/05/30	• Adjust the calculation of between PDF of character • bug fixes that there is an abnormal termination to the case WK4 (123)
1.12	2004/05/05	• Option to ignore the setting of the access rights of the PDF document (-n) • bug fixes that do not see the half-width of the CID in PDF • When you output to standard output, bug fixes that extra carriage return is displayed
		• Encrypted without a password the PDF support (up to 128bit encryption). ***** However, the need to download a separate cryptlib.dll there • Corresponding to the PDF that you created in the

1.11	2004/04/04	"easyPDF 3.1" "Jaws PDF Creator" • Addresses the problem of tab characters are deleted in Ichitaro V7 or later • Add the -p option. View the properties of the Office document
1.10	2004/03/13	• Corresponding to a PDF that was created in OpenOffice.org.1.1
1.09	2004/02/25	• Corresponding to the PDF output in the ActiveReports 2.0J • bug fixes that abnormal termination in a particular PDF
1.08	2004/01/28	• If the formula of the result string of EXCEL double-byte "± × ÷" appeared, bug fixes garbled to the half-width Kana • Removal of the extra line breaks
1.07	2004/01/26	• Word, EXCEL, bug fixes of the full-size in PowerPoint "± × ÷" there is a case to be garbled in the half-width Kana
-	2004/01/18	• To stipulate the "Copyright and Terms of Use".
1.06	2003/11/09	• bug fixes that the first row of the RTF that you saved in WordPad does not appear • If the extension is not included in the Word document in the OLE document of .doc, bug fixes that abnormal termination
1.05	2003/07/15	• PDF of display correspondence that was created in Acrobat6.0

1.04	2003/03/26	• Improvement of calculation between the Japanese PDF of character
1.03	2002/11/23	• Of Unicode encoding PDF support
1.02	2002/10/18	• Corresponding to mht / html
1.01	2002/9/9	• -c option added
1.00	2002/7/8	• To separate the text extraction part from KWIC Finder, published as a filter.

Go to Home