Converting MS Word .doc to LaTeX by command line

Asked 10 years, 3 months ago Modified 4 years, 9 months ago Viewed 49k times

45

24

There are many guides to convert `.doc` to `.tex` with text editor softwares (mostly under Windows); and this popularity has made it hard to search for command-based converting methods.

I am looking for a command line to convert MS Word to LaTeX. This will be part of a PHP script running on Linux server. I mean a shell command to be executed by PHP.

texlive msword word-to-latex

Share Improve this question Follow

asked Feb 27, 2012 at 17:27

Googlebot

3,895 6 35 61

3 Pandoc is a good place to start, although I'm not sure how well it understands `.doc`. You should be prepared; conversion between the two is notoriously patchy (whatever the tool), so you're going to have big problems beyond the most basic of documents. – qubyte Feb 27, 2012 at 18:25

4 You may also try converting `doc` to `html` first, maybe clean it up a bit (`Tidy`) and then running one of the html2latex convertors. There are plenty of those. Hope it helps. – Ondrej Feb 27, 2012 at 18:37

@MarkS.Everitt I understand what you mean. I hope to find a way to ignore messy markups instead of converting them. Converting a very basic structure is better than a bunch of messy markups. – Googlebot Feb 27, 2012 at 18:39

@Ondrej I tried this before (actually my first choice), but conversion of doc to html is not very clean. Do you know a good tool to do so? – Googlebot Feb 27, 2012 at 18:42

2 As the maintainer of `rtf2latex2e` I suggest that the best way is to use Word to save the document as `.rtf` and use `rtf2latex2e` (rtf2latex2e.sf.net). All the MathType equations get translated to more-or-less normal LaTeX. All the figures get extracted and the proper `\includegraphics{file.pdf}` mark-up added. Of course, if you don't have equations then all the other suggestions are just fine. – Scott Prahl Mar 25, 2012 at 6:38

8 Answers

Sorted by:

Highest score (default)

24

Antiword is going to do reasonable good job converting .doc to .tex files . It makes every effort to preserve not only the content but formating as well. It is well suited for batch processing that you want to do.

Edit: Several people asked me privately about LaTeX switch in Antiword and the latest version of Antiword. The latest version is indeed 0.37. As of LaTeX output I think I mixed up things a bit. I used Antiword for formated ASCII output. I think it is capable of PostScript output but not of LaTeX output. As Jon observed you can use pandoc to convert well formated ASCII into LaTeX. However, wvWare (wv and wv2) are capable of outputting LaTeX. A bit of warning. wvWare is depreciated in favor of AbiWord but can be used for batch processing (I have no clue if AbiWord can be used from the command line). It is still a bit younger program (dormant since 2006) than Antiword(dormant since 2004).

Finally there is a tool called catdoc which is great for batch processing but will not preserve format (great for extracting content though and supports MS Excel format).

Share Improve this answer Follow edited Mar 3, 2012 at 2:06 answered Feb 27, 2012 at 18:47

Predrag Punosevac
9,047 1 51 65

I second that: `antiword myfile.doc > myfile.txt ; pandoc -o myfile.tex myfile.txt` works surprisingly well. – jon Feb 27, 2012 at 19:49

@jon Antiword can convert directly into .tex format. Please refer to man pages. – Predrag Punosevac Feb 27, 2012 at 20:03

1 Maybe in newer versions, but the version that I have (0.37) makes no mention of (La)TeX. – jon Feb 27, 2012 at 20:12

Thanks for sharing this useful tool; but I expect to save basic formatting. I do not expect to convert all MS Word markups (and I do not need), but I hope to keep basic ones such as Bold and Italic – Googlebot Feb 27, 2012 at 20:15

3 Is there a newer version than 0.37? That's the only version I can find, and I don't see any options for (La)TeX either. – imnothere Feb 28, 2012 at 9:41

28

This answer is specific to OS X...

Command line utility `textutil`

There's a nice command line utility called `textutil` included in OS X that will convert among common document formats:

Word docx to txt

```
$ textutil -convert txt worddoc.docx
```

txt to Word docx

```
$ textutil -convert docx mytextdoc.txt
```

txt to Word, using Times New Roman 12pt

```
$ textutil -convert docx -font "Times New Roman" -fontsize 12 blah.txt
```

Also works with html, rtf, doc, odt, and others...

word2latex and latex2word by using `textutil` with Pandoc

If you use Pandoc in combination with `textutil` you have can have a decent Word-to-LaTeX and LaTeX-to-Word roundtrip. For docx support you need the latest version of Pandoc (1.9+).

word2latex

```
$ textutil -convert html worddoc.docx -stdout | pandoc -s -f html -t latex -o
latexdoc.tex
```

latex2word

```
$ pandoc -t docx -f latex -o backtoword.docx latexdoc.tex
```

Share Improve this answer Follow

answered Feb 27, 2012 at 21:01
Paul M.
1,471 10 17

1 So much pity that I'm on Linux ;) – Googlebot Feb 27, 2012 at 21:47

1 I just saw this amazing post! I would like to mention that pandoc runs on Mac, Linux, and Windows! see johnmacfarlane.net/pandoc/installing.html and tex.stackexchange.com/questions/44236/... – Jonathan Komar Aug 6, 2012 at 14:21

1 @macmadness86 -- I think the point is that `textutil` is for OS X. `Pandoc` is indeed cross-platform and very useful. – jon Aug 10, 2012 at 1:03

1 Works well except for the Math formulas – lauhub May 11, 2014 at 21:12

3 `pandoc -t latex -f docx in.docx -o out.tex` is the way to go in 2020. – 0x90 Jul 2, 2020 at 17:30

11

A lot depends on how complicated the Word document formatting is. I have had very good success with rtf2latex2e, which converts RTF formatted text to LaTeX. It has various levels of matching the RTF formatting. I have mainly used its "minimal LaTeX markup mode", which is ideal for a document that will be subsequently hand-edited (which I understand is not the same conditions as you require.)

Share Improve this answer Follow edited Feb 28, 2012 at 3:14

answered Feb 28, 2012 at 2:25
Alan Munn
202k 40 511 822

1 Wow, the output of rtf2latex2e is awesome. – mik01aj Mar 29, 2015 at 14:49

8

I have tried most of the suggestions mentioned here at some point. The best workflow I have developed for this starts from first converting the MS Word document to an ODT file using Libre Office on the command line:

```
loffice --headless --convert-to odt msword.doc && cp msword.odt loffice.odt
```

After that, Writer2LaTeX does a pretty good job of converting the basic structure into TeX. I almost always still need to tweak some things manually, so I go with the ultraclean formatting option in order to minimize the output:

```
w2l -ultraclean loffice.odt tweak_this.tex
```

At this point, I normally convert the file to a PDF and visually inspect it, fixing anything that is wrong. When I convert many similar files, I often write short scripts in Python to deal with the most common tweaks.

If you are using Debian/Ubuntu, you can install Writer2LaTeX with `apt-get install writer2latex`, of course. Otherwise, see their website for installation and usage details.

Share Improve this answer Follow edited May 21, 2014 at 19:55 answered Apr 18, 2014 at 18:02

Karol
181 1 3

6

A multiplatform solution is in the works. Rob Oakes is implementing this feature in LyX. Once implemented, I think that using it on the commandline would be straightforward. However, (1) I'm not sure you want to install a somewhat large application for just this conversion and (2) the work is in alpha stages right now. Rob is looking for testers. He's mainly looking for .doc files to test on. He just gave an update on the features that his tool supports: http://marc.info/?l=lyx-users&m=133070969217214&w=2 Here is a list that of features already implemented (copied from the url):

1. Translating Word paragraph and character styles to LyX paragraph and character styles. In the case of character styles that aren't defined, it will write entries for them into the local layout (including basic LaTeX commands).
2. Importing Word tables, including those with merged rows or columns. It will also do its best with the table borders.
3. Enumerated and itemized lists.
4. Importing images from the Word document. (It skips over embedded objects, such as charts from Excel.)
5. The use of custom templates, which allows you to fine tune importing your documents. I've created templates for article.cls and book.cls. I'll also probably create one for

memoir.cls as well.

Share Improve this answer Follow

answered Mar 3, 2012 at 6:26

scottkosty
12.4k 2 31 52

5

I was dismayed that the Linux user above expressed envy of `textutil` on Mac OS X. Although I am on Mac OS X, I don't use `textutil`. As far as I know, it is not capable of reading any formatting information from files anyway. I believe it just extracts text. There are many utilities for that purpose. For instance `antiword` does it for pre-docx MSWord files and is cross platform. `catdoc`, mentioned above, works better on some files. I'm sure I'm forgetting a few good ones ...

Instead of `textutil`, I use a shell script I inherited (I don't know who posted it originally or where) to decode `.docx` files. Here it is:

```
#!/bin/sh
unzip -p "$1" word/document.xml \
| sed -e 's/<\/w:p>/\
/g' -e 's/<[^>]*>//g'
```

I named it `docx`, put it in `/usr/local/bin` and said `sudo chmod 700 docx` to make it executable. Then I say

```
docx mystupidlittlewordfile.docx | less
```

to read the contents of any `.docx` file quickly and conveniently.

All it does is three things one per line (not counting the first line that tells which shell to use). First, it unzips the xml document containing text. This xml document is currently usually `word/document.xml` in the `.docx` file which, if you have not guessed, is a structured `.zip` file. If this script does not work, the first thing I do is to say

```
unzip -l mystupidwordfile.docx
```

so I can see where the text is. You should look inside `word/document.xml` anyway, to understand the next two lines of the shell script above. The second line pipes the output to a `sed` script that replaces MSWord's symbol for newline with an actual newline. The third line pipes this output to another `sed` script that sends everything to `stdout` except whatever is between angle brackets.

There are actually a number of other useful items between angle brackets, such as table layouts, that would be easy to add with intermediate `sed` scripts. You could replace many of these useful items, for instance, with `markdown` equivalents. I just haven't gotten around to it. The main reason I use this script is so that I can see the content of an *average* `.docx` file on a network volume to see if I can understand its content without having to run MSWord to read it. Nine out of ten times, this is all I need to read and understand `.docx` files. Formatting in *average* `.docx` files is usually decorative rather than meaningful.

Presumably, Microsoft will change these angle bracket directives with successive versions of MSWord to make it harder to use any non-Microsoft tools on them. I have heard that Microsoft's original attempt at an *open* format was essentially a MSWord document surrounded by angle brackets.

Nevertheless, as long as public pressure remains for them to use XML, at least some of the directives can be read and modified from version to version. My view is that the end user should be able to modify the tool when it fails because intentional incompatibility has been introduced by the monopolist. I'm not sure how to do that. My solution requires that the user know `sed` and regular expressions and be less lazy than, evidently, I am. This will only fit a small subset of people who would like, perhaps for ethical reasons, to avoid owning a copy of MSWord.

Share Improve this answer Follow

answered Aug 13, 2012 at 20:25
Mick
59 1 1

3 I like your comment a lot! As you observed docx is a glorified xml file + few style files which are zipped. A simple sed script is perfectly capable of dealing with it reasonable well. For people who look more serious tool I recommend docx2txt (sourceforge.net/projects/docx2txt) written in Perl. – Predrag Punosevac Aug 13, 2012 at 22:00

2

I have had a lot of success with an XSL stylesheet called WordML2Latex. It does an excellent of converting formatting to sensible equivalents. Much more than your one-liner. You first save your Word doc as word XML2003, then apply the stylesheet to it.

Being XSL it is cross platform and you can use it from the command line :-). It is also open source ad modifiable.

It's on CTAN http://www.ctan.org/tex-archive/support/WordML2LaTeX

Share Improve this answer Follow

answered Mar 19, 2013 at 13:42
DaveG
91 3

0

I think this site can help you.

https://www.docx2latex.com/

Share Improve this answer Follow edited Sep 13, 2017 at 17:19

answered Sep 13, 2016 at 20:01
aaryan
413 3 10

3 ...but are you sure...? – Werner ♦ Sep 13, 2016 at 20:19

Isn't that for `.docx` ? The question asks about `.doc` which is rather different. – cfr Sep 13, 2016 at 23:18

@cfr you can first convert .doc file to .docx and then try. – aaryan Sep 27, 2016 at 11:09

So you convert to DOC how ...? Note the command line requirement in the question. – cfr Sep 27, 2016 at 22:39