

Re: GSViewer – PS2ASCII Pstotext Unsuccessful pdf_page failed

Source: <http://newsgroups.derkeiler.com/Archive/Comp/comp.lang.postscript/2007-08/msg00021.html>

- *From:* pstlou@xxxxxxxxxxxxxxxx
 - *Date:* Tue, 07 Aug 2007 10:12:58 -0700
-

On Aug 3, 3:16 pm, Ross Presser <rpres...@xxxxxxxx> wrote:

On Aug 3, 4:03 pm, pstl...@xxxxxxxxxxxxxxxx wrote:

On Aug 2, 2:58 pm, Ross Presser <rpres...@xxxxxxxx> wrote:

On Aug 1, 4:38 pm, pstl...@xxxxxxxxxxxxxxxx wrote:

On Aug 1, 2:51 pm, Ross Presser
<rpres...@xxxxxxxx> wrote:

On Aug 1, 3:15 pm,
pstl...@xxxxxxxxxxxxxxxx
wrote:

I am new to
Ghostscript
and
GSView. I
searched for
posts about
PS2ASCII
and found a
hefty 354.
However, I

Re: GSViewer – PS2ASCII Pstotext Unsuccessful pdf_page failed

have not
found, as
yet,
discussion(s)
related to
the error
message
that
generated
for me
when
I attempted
to convert a
PDF to text
within the
GSView
application.
Boiled
down the
message is:

GSview 4.8
2006-02-25
GPL
Ghostscript
8.56
(2007-03-14)

Scanning
PDF file

Warning:
File has a
corrupted
%%EOF
marker, or
garbage
after %
%EOF.

Ghostscript
returns error
code -8
9278
QS 2 47856
-49278 1 e
48388
-49278

QS 2 48389
-49278 1 d
48989
-49278
QS 2 48989
-49278 1 :
49322
-49278
QM 3

Warning:
An error
occurred
while
reading an
XREF
table.
**** The
file has
been
damaged.
This may
have been
caused
**** by a
problem
while
converting
or
transferring
the file.

Ghostscript
will attempt
to recover
the data.

Warning:
There are
objects with
matching
object and
generation

numbers.
The
accuracy of
the resulting
image is
unknown.
Ghostscript

Re: GSViewer – PS2ASCII Pstotext Unsuccessful pdf_page failed

```
returns error  
code -8  
3 35421  
-22776 1 :  
35754  
-22776  
QM 5
```

Has anyone experienced this? I wanted to test the application before I jumped into using a batch process on multiple files. Meanwhile, I'll read the Ghostscript manual. May be a darn good place to start.

Are you sure the PDF is not corrupt, as the top of the message says?
Can you post the PDF somewhere? Have you tried it with other PDFs?

What do you get when you try the commandline script pdftotext.cmd? – Hide quoted text –

– Show quoted text –

Hi Ross,

I suspect that either the PDF is corrupt, or the situation has to do with several imbedded Word tables. Unfortunately, I have no access to post the document to the web.

I had not tried pdftotext. I downloaded a copy. It works fine. Now I'll need to learn if possible to call it within a batch file and pass a macro containing 100 file names. Thanks for the lead....learning one small step at a time.

pdftotext works fine in a batch file, but only takes one file at a time, so you'll have to loop.

Confession: I mistyped my helpful leading question. I meant to ask what happens when you use pdf2ascii.bat --- the command-line version of converting to text using Ghostscript. As you now know pdftotext is a separate program (part of the xpdf package) and works quite differently.

You might also want to take a look at pdftohtml, which despite its name has the option for XML output. (pdftohtml is also derived from xpdf.)<http://pdftohtml.sourceforge.net>–Hidequoted text –

– Show quoted text –

Hi Ross,

I located the bat files and see one for ps2ascii.bat but not one for pdf2ascii.bat. This was in the gs\gs8.56\lib. So I'm stuck again. I went back to the authoring web site and checked to see if I could find it there, but no luck. Have a good weekend.

OK, fine, so my brain is as full of holes as Swiss cheese. ps2ascii is what I meant.

I have to learn to check my advice before hitting Send.– Hide quoted text –

– Show quoted text –

I want to thank you for helping and prodding me along to find the solution. Too often we take for granted the kind advice that people like you give freely. It is appreciated. I also thank Toby Dunn for his sage advice as well.

I found the solution to my problem. It has to do with SAS code used to send X commands to the system to execute the little PDFtoTEXT utility. For some reason, I couldn't pass the contents of the variable, only the variable name. I switched over to Call System commands and it worked like a charm. For anyone out there that may be using SAS and find themselves in a similar situation, here is the code that worked for me:

```
/* clean up previous files */
DATA _NULL_;
X %UNQUOTE(%STR(%'DEL "T:\CMD\PDFTOTEXT\G*.PDF" %'));
X=SLEEP(3);
X %UNQUOTE(%STR(%'DEL "T:\CMD\PDFTOTEXT\G*.TXT" %'));
X=SLEEP(3);
/* copy over new files to be converted */
X %UNQUOTE(%STR(%'COPY "G:\LIBRARY\nOTES\DES\08\G*.pdf" "T:\CMD
\PDFTOTEXT\*.pdf" %'));
X=SLEEP(3);
RUN;

/* pipe directory list */
OPTIONS NOXWAIT NOXSYNC;

FILENAME Q PIPE 'DIR "T:\CMD\PDFTOTEXT\*.PDF" /A-D /B' ;
```

```
Data _Null_ ;  
Infile Q TruncOver ;  
Input Text $200. ;  
  
/* replace .pdf with .txt as new variable NewFile */  
NewFile = CatS( Scan( Text , -2 , '.' ) , '.txt' ) ;  
  
Call System( 'cd t:\cmd\pdftotext' ) ;  
/* this is the line that makes it happen */  
Call System( 'pdftotext -layout ' || Text || ' ' || NewFile ) ;
```

Run ;

.