

Re: More PDF IFilter problems

Source:

<http://www.tech-archive.net/Archive/SQL-Server/microsoft.public.sqlserver.fulltext/2007-09/msg00064.html>

- *From:* "Hilary Cotter" <hilary.cotter@xxxxxxxx>
 - *Date:* Fri, 28 Sep 2007 14:47:42 -0400
-

The entire pdf is an image (an image of text mind you), but there is not text in there.

You need to do ocr on the image to extract the text. =

--

RelevantNoise.com – dedicated to mining blogs for business intelligence.

Looking for a SQL Server replication book?

<http://www.nwsu.com/0974973602.html>

Looking for a FAQ on Indexing Services/SQL FTS

<http://www.indexserverfaq.com>

"LiveCycle" <livecycle@xxxxxxxxxxxxxxxxxxxxxxxx> wrote in message news:OaUtPGfAIHA.1208@xxxxxxxxxxxxxxxxxxxxxxxx

Hi Hilary,

Thank you for responding. I've tried a number of different PDFs, but this is one of the ones that did not work (http://www.cogenix.com/Registration_2007-2008.pdf). As you'll see, it's version 1.3 (Acrobat 4.x), and it's got quite a bit of text in it. I'm sorry I couldn't attach this file directly to this post...

Thanks again, Jim

"Hilary Cotter" <hilary.cotter@xxxxxxxx> wrote in message news:ucRJ11WAIHA.1212@xxxxxxxxxxxxxxxxxxxxxxxx

can you post a problem pdf here? Sometime PDFs only contain images and no text. The iFilter can only understand the text. Also what version of the PDF is it?

--

RelevantNoise.com – dedicated to mining blogs for business intelligence.

Looking for a SQL Server replication book?

Re: More PDF IFilter problems

<http://www.nwsu.com/0974973602.html>

Looking for a FAQ on Indexing Services/SQL FTS

<http://www.indexserverfaq.com>

"LiveCycle" <livecycle@xxxxxxxxxxxxxxxxxxxxxxxxxxxx> wrote in message
[news:%23eQ\\$P6VAIHA.3848@xxxxxxxxxxxxxxxxxxxxxxxxxxxx](mailto:news:%23eQ$P6VAIHA.3848@xxxxxxxxxxxxxxxxxxxxxxxxxxxx)

Sorry, scratch that last one, the DOC file I was looking at was

corrupted. PDF problem remains, however...

Thanks!

"LiveCycle" <livecycle@xxxxxxxxxxxxxxxxxxxxxxxxxxxx>
wrote in message

news:O1Tn0fVAIHA.5184@xxxxxxxxxxxxxxxxxxxxxxxxxxxx

OK, this gets stranger by the minute. I am able to successfully index Excel files, but I am not able to index Word documents! I get this message in my logs.

2007-09-27 15:41:04.59 spid19s Error
'0x8004170c: The document format is not recognized by the filter.'
occurred during full-text index population for table or indexed view '[RMSTest].[Template].[Content]' (table or indexed view ID '2142018762', database ID '12'), full-text key value 0x00000003.

Attempt will be made to reindex it.

2007-09-27 15:41:04.59 spid19s The component 'offfilt.dll' reported error while indexing. Component path 'C:\WINDOWS\system32\offfilt.dll'.

Please, I will lose all my hair soon, any ideas are welcome.

"LiveCycle"

<livecycle@xxxxxxxxxxxxxxxxxxxxxxxxxxxx>

wrote in message

news:eyhiy9UAIHA.1168@xxxxxxxxxxxxxxxxxxxxxxxxxxxx

So, I have found the log, and am receiving the following error information:

Re: More PDF IFilter problems

2007-09-27 14:40:23.13

spid21s Informational:

Full-text Full

population initialized for
table or indexed view

'[RMSTest].[Template].[Content]'

(table or indexed view ID

'2142018762', database ID

'12'). Population sub-tasks:

1.

2007-09-27 14:40:37.24

spid21s Error '0x80043651:

msftesql should

reprocess this document in

an isolated fashion to

confirm the error.'

occurred during full-text

index population for table or

indexed view

'[RMSTest].[Template].[Content]'

(table or indexed view ID

'2142018762', database ID

'12'), full-text key value

0x00000001.

Attempt will be made to

reindex it.

2007-09-27 14:40:37.24

spid21s The component

'MSFTE.DLL' reported

error while indexing.

Component path

'C:\Program Files\Microsoft

SQL

Server\MSSQL.1\MSSQL\Binn\MSFTE.DLL'.

2007-09-27 14:40:37.24

spid21s Warning: No

appropriate filter for

embedded object was found

during full-text index

population for table

or indexed view

'[RMSTest].[Template].[Content]'

(table or indexed

view ID '2142018762',

database ID '12'), full-text

key value

0x00000002. Some

embedded objects in the row

could not be indexed.

2007-09-27 14:40:37.24

spid21s Informational:

Re: More PDF IFilter problems

Full-text Full
population completed for
table or indexed view
'[RMSTest].[Template].[Content]'
(table or indexed view ID
'2142018762', database ID
'12'). Number of documents
processed: 2.
Number of documents
failed: 0. Number of
documents need retry: 1.

Clearly, it doesn't like my
IFilter. Any ideas how I can
make SQL
recognize this?

Thanks, Jim

"LiveCycle"
<livecycle@xxxxxxxxxxxxxxxxxxxxxxxx>
wrote in message
news:%23o1jMiUAIHA.3940@xxxxxxxxxxxxxxxxxxxxxxxx

Hi all,

I'm having
some
frustrating
issues with
the PDF
IFilter and
making it
work. I've
read the
other posts
here, and
still haven't
been able to
figure this
out. I am
running
SQL Server
2005
Standard 32
bit
edition on
Windows
Server 2003
Standard
Edition. I

Re: More PDF IFilter problems

performed
the
following:

- 1 –
Installed the
PDF IFilter
v 6.0
- 2 – Ran
EXEC
sp_fulltext_service
'load_os_resources',
1
- 3 – Stopped
and
restarted the
SQL Server
service
- 4 – Ran
sys.fulltext_document_types
and verified
that .pdf
was indeed
a valid
document
type
- 5 – Built a
new
full-text
catalog and
added my
table with
PDF &
other
files (stored
as image
data type) to
the catalog
- 6 – Fully
populated
the FT
index
- 7 – Ran my
CONTAINS
query
against that
table. I'm
able to
return
results
against

Re: More PDF IFilter problems

Office files,
but nothing
for PDF
files.

So, I'm not
sure what I
should do at
this point. I
even tried
restarting
the server
itself.

Somebody
(Hilary
Cotter?)
mentioned
that it might
be possible
to look at
gatherer
logs
somewhere,
but I'm
not clear
where those
would be. I
would
appreciate
any further
suggestions.

Thanks, Jim

Re: More PDF IFilter problems