



This repository

Explore Gist Blog Help

 russkel + - ✕

.IC  jcushman / pdfquery

Watch 8 Star 40 Fork 10

unicode problem when processing doc.info #11

[New issue](#)

Open **xuewei4d** opened this issue on Mar 29 · 11 comments



xuewei4d commented on Mar 29

When I use pdfquery processing a scholar pdf, I found a unicode problem in Line 305, pdfquery.py The variable 'v' is a str type, but stores unicode character. For example, v could be 'xǹc'. Since 'v' is a str type, it is literally '\', 'x', 'ǹ', 'c'.

Line 305,

```
root.set(k, unicode(v))
```

would get a 'UnicodeDecodeError'. I suggest to use

```
root.set(k, v.decode('unicode-escape'))
```



xuewei4d commented on Mar 29

Another problem is Line 358 in pdfquery.py

```
branch.text = node.get_text()
```

I suggest remove illegal xml characters here.



russkel commented 22 hours ago

What's the go with this issue? I just ran into it as well.

Changing the line to unicode-escape leads to issues such as:

Traceback (most recent call last):

```
File "/Users/russ/PycharmProjects/solar_inspection_report/si_rename.py", line 11, in <module>
    pdf.load(2) # load only the 2nd page to save CPU time
File "/usr/local/lib/python2.7/site-packages/pdfquery/pdfquery.py", line 230, in load
    self.tree = self.get_tree(*_flatten(page_numbers))
File "/usr/local/lib/python2.7/site-packages/pdfquery/pdfquery.py", line 307, in get_tree
    root.set(k, v.decode('unicode-escape'))
File "lxml.etree.pyx", line 746, in lxml.etree._Element.set (src/lxml/lxml.etree.c:42970)
File "apihelpers.pxi", line 547, in lxml.etree._setAttributeValue (src/lxml/lxml.etree.c:19025)
File "apihelpers.pxi", line 1395, in lxml.etree._utf8 (src/lxml/lxml.etree.c:26485)
ValueError: All strings must be XML compatible: Unicode or ASCII, no NULL bytes or control characters
```



xuewei4d commented 16 hours ago

@**russkel** I suggest to debug around Line 305, like printing something out. Maybe in your PDF file, key `k` also contains illegal characters.



jcushman commented 15 hours ago

Owner

Unicode issues are tricky! If you can point me to a PDF that causes the issue I may be able to debug. Even better if you can give me a patch/pull request that fixes the issue with the PDF you point me to ...

Thanks,
Jack

Labels

None yet

Milestone

No milestone

Assignee

No one assigned

Notifications

Unsubscribe

You're receiving notifications because you were mentioned.

3 participants

