TextCritical.net - Bug #2616

Berean bible has weird characters

03/22/2020 09:25 PM - Luke Murphey

Status:	Closed	**Start date:**	03/22/2020
Priority:	Normal	**Due date:**	
Assignee:	Luke Murphey	**% Done:**	100%
Category:		**Estimated time:**	0.00 hour
Target version:	3.2.1		

Description

They are &#147

Related issues:

Related to TextCritical.net - Feature #2461: Add Berean Study Bible	**Closed**	**02/06/2020**

Associated revisions

Revision 1100 - 03/25/2020 07:51 PM - luke.murphey

Fixing Berean import failure

Reference #2616

Revision 1109 - 03/25/2020 07:51 PM - luke.murphey

Fixing Berean import failure

Reference #2616

Revision 1101 - 03/25/2020 08:56 PM - luke.murphey

Fixing test cases where the file got fixed in the test somehow

Reference #2616

Revision 1110 - 03/25/2020 08:56 PM - luke.murphey

Fixing test cases where the file got fixed in the test somehow

Reference #2616

History

#1 - 03/22/2020 09:30 PM - Luke Murphey

Obs:

- https://textcritical.net/work/berean-study-bible/Acts/6/11
- Surrogates not allowed happens when I use utf-8 when trying to persist to the database
  - https://github.com/mitmproxy/mitmproxy/issues/2075
- I was able to import with cp1252/surrogateescape; this fails in the tests though
- I tried importing via utf-8/surrogateescape; fails on database import which blows up complaining about

Qs:

- Is there something missing where the chars are?
  - Double quotes
- What char is this?
  - https://www.codetable.net/decimal/147

- What encoding is this file?
    - UTF-8 works with the double-quotes but complains about the start character: UnicodeDecodeError: 'utf-8' codec can't decode byte 0xa9 in position 40: invalid start byte
    - cp1252 cannot load: UnicodeDecodeError: 'charmap' codec can't decode byte 0x9d in position 614: character maps to <undefined>
        - https://superuser.com/questions/301552/how-to-auto-detect-text-file-encoding
        - bsb.txt: Windows-1252 with confidence 0.73
    - iso-8859-1 fails to load the double-quotes
    - windows-1252: UnicodeDecodeError: 'charmap' codec can't decode byte 0x9d in position 614: character maps to <undefined>
- What is the difference between cp1252 and windows-1252?
    - https://www.i18nqa.com/debug/table-iso8859-1-vs-windows-1252.html
    - This looks like utf-8 according to the table
- Where exactly is the problem in the file with cp1252?
    - That looks good actually
    - Tests fail on it though with surrogateescape. Strange because I imported it this way fine.

Refs:

- https://docs.python.org/2.4/lib/standard-encodings.html

#2 - 03/25/2020 06:23 PM - Luke Murphey

- Assignee set to Luke Murphey

- Target version set to 3.2.1

#3 - 03/25/2020 06:35 PM - Luke Murphey

- Related to Feature #2461: Add Berean Study Bible added

#4 - 03/25/2020 07:03 PM - Luke Murphey

```
python3 manage.py import_berean_bible -f /db/bsb.txt
```

#5 - 03/25/2020 08:59 PM - Luke Murphey

- Status changed from New to Closed

- % Done changed from 0 to 100