

Website Input - Bug #987

Failure to parse content due to encoding problem

03/24/2015 02:42 AM - Luke Murphey

Status:	Closed	Start date:	03/24/2015
Priority:	Normal	Due date:	
Assignee:	Luke Murphey	% Done:	100%
Category:		Estimated time:	0.00 hour
Target version:	1.0.4		
Description			
See http://answers.splunk.com/answers/223557/how-to-troubleshoot-why-the-website-input-app-is-n.html			
The app generates the following exceptions:			
<pre>2015-03-23 21:40:06,607 ERROR A general exception was thrown when executing a web request Traceback (most recent call last): File "/Library/Splunk/splunk_sp/etc/apps/website_input/bin/web_input.py", line 314, in scrape_page tree = lxml.html.fromstring(content) File "/Library/Splunk/splunk_sp/lib/python2.7/site-packages/lxml/html/__init__.py", line 706, in fromstring doc = document_fromstring(html, parser=parser, base_url=base_url, **kw) File "/Library/Splunk/splunk_sp/lib/python2.7/site-packages/lxml/html/__init__.py", line 600, in document_fromstring value = etree.fromstring(html, parser, **kw) File "lxml.etree.pyx", line 3032, in lxml.etree.fromstring (src/lxml/lxml.etree.c:68121) File "parser.pxi", line 1781, in lxml.etree._parseMemoryDocument (src/lxml/lxml.etree.c:102435) ValueError: Unicode strings with encoding declaration are not supported. Please use bytes input or XML fragments without declaration.</pre>			
<pre>2015-03-23 21:40:06,607 ERROR An exception occurred when attempting to retrieve information from the web-page Traceback (most recent call last): File "/Library/Splunk/splunk_sp/etc/apps/website_input/bin/web_input.py", line 470, in run result = WebInput.scrape_page(url, selector, username, password, timeout, name_attributes, proxy_type=proxy_type, proxy_server=proxy_server, proxy_port=proxy_port, proxy_user=proxy_user, proxy_password=proxy_password) File "/Library/Splunk/splunk_sp/etc/apps/website_input/bin/web_input.py", line 314, in scrape_page tree = lxml.html.fromstring(content) File "/Library/Splunk/splunk_sp/lib/python2.7/site-packages/lxml/html/__init__.py", line 706, in fromstring doc = document_fromstring(html, parser=parser, base_url=base_url, **kw) File "/Library/Splunk/splunk_sp/lib/python2.7/site-packages/lxml/html/__init__.py", line 600, in document_fromstring value = etree.fromstring(html, parser, **kw) File "lxml.etree.pyx", line 3032, in lxml.etree.fromstring (src/lxml/lxml.etree.c:68121) File "parser.pxi", line 1781, in lxml.etree._parseMemoryDocument (src/lxml/lxml.etree.c:102435) ValueError: Unicode strings with encoding declaration are not supported. Please use bytes input or XML fragments without declaration.</pre>			

History

#1 - 03/24/2015 02:42 AM - Luke Murphey

See <http://stackoverflow.com/questions/25023237/html-scraping-using-lxml-and-requests-gives-a-unicode-error>

#2 - 03/27/2015 02:52 AM - Luke Murphey

- *Status changed from New to In Progress*

#3 - 03/27/2015 03:45 AM - Luke Murphey

I have a unit test reproducing this.

#4 - 03/27/2015 10:58 PM - Luke Murphey

<http://rss.slashdot.org/Slashdot/slashdot>

#5 - 03/28/2015 03:06 AM - Luke Murphey

- *Status changed from In Progress to Closed*

- *% Done changed from 0 to 100*