Website Input - Bug #1597

Parse failures not handled well

11/17/2016 07:05 AM - Luke Murphey

Status:	Closed	**Start date:**	11/17/2016
Priority:	Normal	**Due date:**	
Assignee:	Luke Murphey	**% Done:**	100%
Category:	Input: Web Spider	**Estimated time:**	0.00 hour
Target version:	4.0		

Description

Getting an exception like:

```
2016-11-17 01:04:33,894 ERROR An exception occurred when attempting to retrieve information from t
he web-page, stanza=web_input://Auth Fail Test
Traceback (most recent call last):
  File "/Users/lmurphey/Splunk/sp/etc/apps/website_input/bin/web_input.py", line 1035, in run
    result = WebInput.scrape_page(url, selector, username, password, timeout, name_attributes, pro
xy_type=proxy_type, proxy_server=proxy_server, proxy_port=proxy_port, proxy_user=proxy_user, proxy
_password=proxy_password, user_agent=user_agent, use_element_name=use_element_name, page_limit=pag
e_limit, depth_limit=depth_limit, url_filter=url_filter, include_raw_content=raw_content, text_sep
arator=text_separator, browser=browser, output_matches_as_mv=output_matches_as_mv, output_matches_
as_separate_fields=output_matches_as_separate_fields)
  File "/Users/lmurphey/Splunk/sp/etc/apps/website_input/bin/web_input.py", line 919, in scrape_pa
ge
    result = cls.get_result_single(http, urlparse(url), selector, headers, name_attributes, output
_matches_as_mv, output_matches_as_separate_fields, charset_detect_meta_enabled, charset_detect_con
tent_type_header_enabled, charset_detect_sniff_enabled, include_empty_matches, use_element_name, *
*kw)
  File "/Users/lmurphey/Splunk/sp/etc/apps/website_input/bin/web_input.py", line 622, in get_resul
t_single
    tree = lxml.html.fromstring(content_decoded)
  File "/Users/lmurphey/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 "/Users/lmurphey/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 1786, in lxml.etree._parseMemoryDocument (src/lxml/lxml.etree.c:102470)
  File "parser.pxi", line 1667, in lxml.etree._parseDoc (src/lxml/lxml.etree.c:101229)
  File "parser.pxi", line 1035, in lxml.etree._BaseParser._parseUnicodeDoc (src/lxml/lxml.etree.c:
96139)
  File "parser.pxi", line 582, in lxml.etree._ParserContext._handleParseResultDoc (src/lxml/lxml.e
tree.c:91290)
  File "parser.pxi", line 683, in lxml.etree._handleParseResult (src/lxml/lxml.etree.c:92476)
  File "parser.pxi", line 631, in lxml.etree._raiseParseError (src/lxml/lxml.etree.c:91904)
XMLSyntaxError: line 135: Element script embeds close tag
```

History

#1 - 11/17/2016 06:58 PM - Luke Murphey

- Status changed from New to Closed

- % Done changed from 0 to 100