Website Input - Bug #2190

Bad encoding causes the input to fail

01/26/2018 11:37 PM - Luke Murphey

Status:	Closed	**Start date:**	01/26/2018
Priority:	Normal	**Due date:**	
Assignee:	Luke Murphey	**% Done:**	100%
Category:	Input: Web Spider	**Estimated time:**	0.00 hour
Target version:	4.5.2		

Description

See https://answers.splunk.com/answers/597938/python-error-lookuperror-unknown-encoding-3dutf-8.html

Associated revisions

Revision 630 - 01/26/2018 11:54 PM - lukemurphey

The input will now continue even if it gets a bad encoding

Reference #2190

Revision 636 - 02/20/2018 09:37 PM - lukemurphey

Making the input continue even if an input fails to parse some content

Reference #2190

Revision 639 - 02/23/2018 09:52 PM - lukemurphey

Making the input more resistant to HTTP problems

Reference #2190

Revision 640 - 02/23/2018 11:16 PM - lukemurphey

Added additional logging

Reference #2190

History

#1 - 01/26/2018 11:55 PM - Luke Murphey

- Status changed from New to Closed

- % Done changed from 0 to 100

#2 - 02/07/2018 09:18 PM - Luke Murphey

- Status changed from Closed to In Progress

- % Done changed from 100 to 90

#4 - 02/07/2018 10:24 PM - Luke Murphey

Observations:

1. The matches are actually working with the exception of "http://www.mos-eisley.dk/dashboard/\\"

Questions:

1. Are results coming through?
    1. source="web_input://www_mos_eisley_dk" | table _time url match*
2. What platform and version of Splunk is this running on?

3. Is thus using authentication to connect?
    1. Yes
4. Why doesn't this repro locally?
    1. Perhaps because the limit isn't high enough
5. Is the scraper running without a filter?
    1. It indeed has no filter
    2. It may be running out of memory

#6 - 02/13/2018 01:16 AM - Luke Murphey

Observations:

1. Is the input running?
    1. The input isn't running on the host, despite being enabled
2. Does it work _and parse) when run from SPL?
    1. Yes: locally and on the host
    2. | webscrape selector="h1" url="http://www.mos-eisley.dk" page_limit=20 depth_limit=25 raw_content=1 empty_matches=0
3. What log messages exist?
    1. None that I can see that indicate why it isn't working

#7 - 02/23/2018 07:32 PM - Luke Murphey

Another error:

```
2018-02-23 20:14:25,239 ERROR An exception occurred when attempting to retrieve information from the web-page,
 stanza=web_input://www_mos_eisley_dk
Traceback (most recent call last):
  File "/splunk/etc/apps/website_input/bin/web_input.py", line 349, in run
    https_only=self.is_on_cloud(input_config.session_key))
  File "/splunk/etc/apps/website_input/bin/website_input_app/web_scraper.py", line 718, in scrape_page
    include_empty_matches, use_element_name,
  File "/splunk/etc/apps/website_input/bin/website_input_app/web_scraper.py", line 418, in get_result_single
    content = web_client.get_url(url.geturl())
  File "/splunk/etc/apps/website_input/bin/website_input_app/web_client.py", line 351, in get_url
    self.response = self.browser.open(url, timeout=self.timeout)
  File "/splunk/etc/apps/website_input/bin/mechanize/_mechanize.py", line 254, in open
    return self._mech_open(url_or_request, data, timeout=timeout)
  File "/splunk/etc/apps/website_input/bin/mechanize/_mechanize.py", line 284, in _mech_open
    response = UserAgentBase.open(self, request, data)
  File "/splunk/etc/apps/website_input/bin/mechanize/_opener.py", line 195, in open
    response = urlopen(self, req, data)
  File "/splunk/etc/apps/website_input/bin/mechanize/_urllib2_fork.py", line 352, in _open
    '_open', req)
  File "/splunk/etc/apps/website_input/bin/mechanize/_urllib2_fork.py", line 340, in _call_chain
    result = func(*args)
  File "/splunk/etc/apps/website_input/bin/mechanize/_urllib2_fork.py", line 1188, in http_open
    return self.do_open(httplib.HTTPConnection, req)
  File "/splunk/etc/apps/website_input/bin/mechanize/_urllib2_fork.py", line 1158, in do_open
```

```
    r = h.getresponse()
  File "/splunk/lib/python2.7/httplib.py", line 1121, in getresponse
    response.begin()
  File "/splunk/lib/python2.7/httplib.py", line 438, in begin
    version, status, reason = self._read_status()
  File "/splunk/lib/python2.7/httplib.py", line 402, in _read_status
    raise BadStatusLine(line)
BadStatusLine: ''
```

#8 - 02/25/2018 08:44 PM - Luke Murphey

- Status changed from In Progress to Closed

- % Done changed from 90 to 100