Website Input - Bug #2859

Scrape page does not work for some pages

10/21/2020 03:34 PM - Luke Murphey

Status:	Closed	**Start date:**	10/21/2020
Priority:	Normal	**Due date:**	
Assignee:		**% Done:**	100%
Category:		**Estimated time:**	0.00 hour
Target version:			

Description

History

#1 - 10/21/2020 04:30 PM - Luke Murphey

Observations:

- \It works with the search command:
  - | webscrape selector=".u\.txt\:lg" url="https://splunkbase.splunk.com/app/1818/" depth_limit=25 empty_matches=0
- It works with Firefox
- I see an exception

```
2020-10-21 08:47:14,495 ERROR Exception generated while attempting to content for url=https://splunkbase.s
plunk.com/app/1818/
Traceback (most recent call last):
  File "/Users/lmurphey/Splunk/7291/etc/apps/website_input/bin/website_input_ops_rest_handler.py", line 28
5, in get_load_page
    content = web_client.get_url(url, 'GET')
  File "/Users/lmurphey/Splunk/7291/etc/apps/website_input/bin/website_input_app/web_client.py", line 492,
 in get_url
    raise ConnectionFailure(str(e), e)
ConnectionFailure: <urlopen error ('The read operation timed out',)>, caused by URLError(SSLError('The rea
d operation timed out',),)
```

*

Questions:

- What happens when I strip the endpoint down?
  - Still fails
- Does this fail in the unit tests?
  - It does
- Does the httplib2 client work?
  - DefaultWebClient is the MechanizeClient
  - It successfully returns but times out
- Does this work from the CLI?

```
import sys
sys.path.append(os.path.join("."))
import mechanize
browser = mechanize.Browser()
response = browser.open("https://splunkbase.splunk.com/app/1818/", timeout=15)
content = response.read(800)
```

  - `ImportError: This package should not be accessible on Python 3. Either you are trying to run from the p ython-future src folder or your installation of python-future is corrupted.`
    - This is because I tried the python 3 client on an old host

Possible causes:

- Python 3 issues

- Old mechanize client
  - Tried the new one: same issue
- SSL problem
- Low timeout
  - This was it

#2 - 10/21/2020 04:35 PM - Luke Murphey

- Status changed from New to Closed

- % Done changed from 0 to 100