Website Input - Bug #1739

Figure why selector doesn't match

02/11/2017 08:37 PM - Luke Murphey

Status:	Closed		**Start date:**	02/11/2017
Priority:	Normal		**Due date:**	
Assignee:	Luke Murphey		**% Done:**	100%
Category:	Input: Web Spider		**Estimated time:**	0.00 hour
Target version:	4.1			

Description
Try a selector of "#siteTable .title" on https://www.reddit.com/r/Splunk/new/. It shows matches and it matches in a browser manually. However, it produces no results.

Associated revisions

Revision 357 - 02/26/2017 06:20 AM - lukemurphey

Fixing no matches on some IDs

Fixing no matches on selectors using non-lowercase IDs

Reference #1739

History

#1 - 02/11/2017 08:38 PM - Luke Murphey

Repro:

```
| webscrape selector="#siteTable .title" url="https://www.reddit.com/r/Splunk/new/"
```

This doesn't work either:

```
| webscrape selector="#siteTable .title" url="https://www.reddit.com/r/Splunk/new/" browser="firefox"
```

#2 - 02/26/2017 05:32 AM - Luke Murphey

To check:

1. Is the selector getting passed through
2. Can this be reproduced via a simple Python script?

#3 - 02/26/2017 05:42 AM - Luke Murphey

I cannot match on #siteTable.

#4 - 02/26/2017 06:03 AM - Luke Murphey

The issue is that I am making the selector lowercase in the SelectorField class. This is because CSS selectors ought to match without respect to case ("DIV" and "div" should match identically).

See http://stackoverflow.com/questions/1734125/is-it-possible-for-lxml-to-work-in-a-case-insensitive-manner

#5 - 02/26/2017 06:05 AM - Luke Murphey

Some options:

1. See if a newer version of CSS select addresses this
2. Use BeautifulSoup
3. Change the selector to match both (given and lowercase)

#6 - 02/26/2017 06:07 AM - Luke Murphey

Can try https://pypi.python.org/pypi/cssselect#downloads

#7 - 02/26/2017 06:10 AM - Luke Murphey

- % Done changed from 0 to 30

#8 - 02/26/2017 06:15 AM - Luke Murphey

cssselector 1.0.1 does not change the behavior.

#9 - 02/26/2017 06:22 AM - Luke Murphey

- Status changed from New to Closed

- % Done changed from 30 to 100