Website Input - Feature #762

Web spider support

07/09/2014 09:47 PM - Luke Murphey

Status:	Closed	**Start date:**	04/29/2016
Priority:	Normal	**Due date:**	
Assignee:	Luke Murphey	**% Done:**	100%
Category:		**Estimated time:**	0.00 hour
Target version:	2.0		

Description

Add ability to have input spread to multiple pages. To do this, the input would need to be updated to support:

- A seed URL (that indicates where to start looking for matching URLs)
- A URL filter to limit what to look at
- Depth limit (limits how deep to keep looking for content)
- URL limit (limits how many files to evaluate)
- Selector to use for grabbing content
- (optional) content-type filter

This may not be done as a modular input but rather as a one-time input to get data into Splunk.

Subtasks:

Task # 1303: Convert scrape_page to use more helper functions	**Closed**
Task # 1304: Update input to output multiple results	**Closed**
Task # 1305: Update input to recursively spider websites	**Closed**
Task # 1309: Implement extracted link processing	**Closed**
Task # 1310: Implement domain_limit	**Closed**
Task # 1306: Update search command to include options for scraping	**Closed**
Task # 1307: Update input page to include options for scraping	**Closed**
Task # 1308: Update preview to output scraping results	**Closed**

History

#1 - 07/23/2014 09:31 PM - Luke Murphey

- Assignee deleted (Luke Murphey)

#2 - 01/03/2015 09:14 AM - Luke Murphey

This may be a good search command. With a search command, users could set up some processing to do things like get certain page numbers. For example, if I wanted to pull notes from Bible.com (https://www.bible.com/users/LukeMurphey/notes), then I would need to increment the page number variable to get the next page.

#3 - 05/26/2015 07:24 PM - Luke Murphey

- Assignee set to Luke Murphey

- Target version set to 2.0

#4 - 05/26/2015 07:25 PM - Luke Murphey

See http://answers.splunk.com/answers/239251/what-is-the-best-way-to-monitor-a-web-page-contain.html

#5 - 06/12/2015 04:54 AM - Luke Murphey

- Status changed from New to In Progress

#6 - 06/22/2015 11:21 PM - Luke Murphey

http://www.xavierdupre.fr/blog/2013-10-27_nojs.html

#7 - 09/18/2015 06:35 AM - Luke Murphey

To simplify, I think I could just use the main URL as the seed URL and then have a separate section for spidering.

The spider section would include:

- URL filter
- Depth limit
- URL limit

#8 - 09/18/2015 06:40 AM - Luke Murphey

A good test case is Ryobi Tools. In that case, I want it to extract the content from https://www.ryobitools.com/outdoor/products/list/family/one-plus and then go to the pages. I need to have the app distinguish between pages that are identical but just have a different page number.

#9 - 09/18/2015 06:43 AM - Luke Murphey

Could use Etags but these are not necessarily provided by all servers.

#10 - 01/15/2016 06:23 PM - Luke Murphey

- Priority changed from Normal to High

#11 - 01/15/2016 06:23 PM - Luke Murphey

- Status changed from In Progress to New

#12 - 04/29/2016 06:18 PM - Luke Murphey

Could you use Scrapy: http://scrapy.org/download/

#13 - 04/29/2016 06:59 PM - Luke Murphey

I will likely need to break up the scrape_page function up. It currently:

1. Preps and validates the arguments
2. Resolves the proxy configuration
3. Resolves the user-agent string
4. Performs the HTTP request, records the stats regarding load time, size
5. Resolves the encoding and decodes the content
6. Parses the HTML
7. Runs the selector against the HTML
8. Outputs the matches

#14 - 04/29/2016 10:35 PM - Luke Murphey

Question about using scrapy:

- Can apply a limit?
  - http://stackoverflow.com/questions/19160594/scrapy-limit-the-number-of-request-or-request-bytes
- How are links extracted?
- How to handle proxies?
  - http://stackoverflow.com/questions/4710483/scrapy-and-proxies
- How to set user-agent string?
  - http://stackoverflow.com/questions/18920930/scrapy-python-set-up-user-agent
  - http://doc.scrapy.org/en/latest/topics/spiders.html#scrapy.spiders.Spider.custom_settings

- How is encoding detected?
  - http://stackoverflow.com/questions/10735836/scrapy-spider-dealing-with-pages-that-have-incorrectly-defined-character-encodi
- What types of proxy servers does scrapy support?
  - How can authentication be handled?

#15 - 04/29/2016 11:02 PM - Luke Murphey

To do mine own spider I would just need to:

- Write extractors
- Loop on the results

#16 - 05/02/2016 04:13 AM - Luke Murphey

- Status changed from New to Closed