Website Input - Feature #758

Handle forms authentication

07/07/2014 09:03 PM - Luke Murphey

Status:	Closed	**Start date:**	07/30/2017
Priority:	Normal	**Due date:**	
Assignee:	Luke Murphey	**% Done:**	100%
Category:	Input: Web Spider	**Estimated time:**	0.00 hour
Target version:	4.4		

Description

Add support for forms authentication.

Subtasks:

Task # 1949: Use a wrapper around the http client	**Closed**
Task # 1950: Include mechanize in the app	**Closed**
Task # 1951: Create a web client wrapper around mechanize	**Closed**
Task # 1952: Add options for forms authentication (backend)	**Closed**
Task # 1953: Use form authentication in web client	**Closed**
Task # 1954: Add ability to define form authentication on front-end	**Closed**
Task # 1955: Update controller to work with web-client wrapper	**Closed**
Task # 1956: Implement proxy support for mechanize	**Closed**
Feature # 1961: Update controller to support previews with forms authentication	**Closed**
Bug # 1962: Deal with unit test that responds with a 401	**Closed**
Task # 1966: Update documentation and text regarding HTTP basic authentication and form...	**Closed**
Feature # 1967: Improve handling for form fields	**Closed**

Related issues:

Related to Website Input - Feature #1963: Auto-discover form fields	**Closed**	**08/02/2017**
Related to Website Input - Feature #1968: Add browser support for forms authe...	**Closed**	**08/07/2017**

History

#1 - 07/07/2014 09:03 PM - Luke Murphey

"Sometimes you might need to create an account and login to access the information you need. If you have a good HTTP library that handles logins and automatically sending session cookies (did I mention how awesome Requests is?), then you just need your scraper login before it gets to work."

http://blog.hartleybrody.com/web-scraping/

#2 - 08/26/2014 04:14 PM - Luke Murphey

http://wwwsearch.sourceforge.net/mechanize/
http://stackoverflow.com/questions/11685235/login-using-python-in-basic-html-form

#3 - 09/03/2015 10:16 PM - Luke Murphey

Might be able to support with this: https://github.com/lorien/grab

#4 - 09/03/2015 11:18 PM - Luke Murphey

http://answers.splunk.com/answers/301772/website-input-how-do-i-monitor-a-forum-that-requir.html#answer-305006

#5 - 01/15/2016 06:07 PM - Luke Murphey

https://answers.splunk.com/answers/339200/website-input-how-far-off-is-support-for-forms-bas.html

#6 - 01/15/2016 06:08 PM - Luke Murphey

Might be able to make this work using mechanize (http://wwwsearch.sourceforge.net/mechanize/)

See http://stackoverflow.com/questions/4847226/form-based-authentication-with-python

#7 - 01/15/2016 06:22 PM - Luke Murphey

- Priority changed from Normal to High

#8 - 05/04/2016 06:56 PM - Luke Murphey

http://wwwsearch.sourceforge.net/mechanize/

#9 - 07/26/2017 05:18 PM - Luke Murphey

httplib2 doesn't seem to persist sessions very well, see:

- https://github.com/jcgregorio/httplib2/wiki/Examples
- https://stackoverflow.com/questions/923296/keeping-a-session-in-python-while-making-http-requests

Request is likely a better option: https://stackoverflow.com/questions/19566645/python-httplib2-https-login-fails

#10 - 07/26/2017 05:27 PM - Luke Murphey

The user is going to need to define several things:

1. Authentication URL
2. Username
3. Password
4. Username field name
5. Password field name

Concerns:

1. How do I get the built-in client and the browser to both support this?
    1. Can the cookies be transferred?
2. What about CSRF protection?

#11 - 07/26/2017 05:33 PM - Luke Murphey

A conf file might look like this:

```
authentication_url=https://domain.com/auth
username_field=username
password_field=password
authentication_action=POST
```

#12 - 07/29/2017 06:32 AM - Luke Murphey

Tried mechanize. It actually works. Here is a Redmine login:

```
import mechanize

url = "http://Lukemurphey.net/login"
br = mechanize.Browser()
br.set_handle_robots(False) # ignore robots
br.open(url)

br.select_form(nr=2)

br.form['username'] = 'Luke'
br.form['password'] = 'OPENSESAME'

res = br.submit()
content = res.read()
```

To use it, you will need:

- mechanize: https://github.com/python-mechanize/mechanize
- html5lib 0.999999999: https://pypi.python.org/pypi/html5lib
- six 1.10: https://pypi.python.org/pypi/six
- webencodings 0.5.1: https://pypi.python.org/pypi/webencodings

#13 - 07/29/2017 06:43 AM - Luke Murphey

To convert this over, I would need:

1. The ability to obtain the raw HTML in get_result_built_in_client()
    1. https://stackoverflow.com/questions/9552773/raw-html-vs-dom-scraping-in-python-using-mechanize-and-beautiful-soup
2. The ability to obtain the response code in get_result_built_in_client()
    1. https://stackoverflow.com/questions/11809696/python-mechanize-browser-openurl-status-code
3. The ability to assign a proxy server in get_http_client()
    1. Note that get_http_client() is not currently used by scrape page
    2. https://stackoverflow.com/questions/1997894/pythons-mechanize-proxy-support
4. The ability to assign HTTP credentials
    1. https://stackoverflow.com/questions/5291576/basic-and-form-authentication-with-mechanize-ruby
    2. https://stackoverflow.com/questions/40919543/python-mechanize-implementation-of-http-basic-auth
    3. https://stackoverflow.com/questions/1097380/can-python-mechanize-handle-http-auth
5. The ability to set the user-agent
    1. http://stockrt.github.io/p/emulating-a-browser-in-python-with-mechanize/

I think I should also:

- Break out authentication from scrape_page() into a setAuthentication() call
- Put the HTTP client under an abstraction layer so that I can switch it out as I see fit

#14 - 07/29/2017 06:44 AM - Luke Murphey

- Category set to Input: Web Spider

- Target version set to 4.4

#15 - 07/31/2017 04:21 PM - Luke Murphey

http://mechanize.readthedocs.io/en/latest/browser_api.html

#16 - 08/02/2017 07:44 AM - Luke Murphey

- Related to Feature #1963: Auto-discover form fields added

#17 - 08/03/2017 08:00 AM - Luke Murphey

Need to have some logging if the fields cannot be determined.

#18 - 08/04/2017 04:45 AM - Luke Murphey

- Related to Feature #1968: Add browser support for forms authentication added

#19 - 08/04/2017 08:22 AM - Luke Murphey

- Status changed from New to Closed