Website Input - Task #1809

Refactor scrape_page call

04/04/2017 05:24 PM - Luke Murphey

Status:	Closed	**Start date:**	04/04/2017
Priority:	Normal	**Due date:**	
Assignee:	Luke Murphey	**% Done:**	100%
Category:	Input: Web Spider	**Estimated time:**	0.00 hour
Target version:	4.2		

Description

The scrape_page call takes a ton of arguments which makes it unwieldy.

Associated revisions

Revision 398 - 04/04/2017 06:26 PM - lukemurphey

Separating scraping from the input code

Reference #1809

Revision 401 - 04/04/2017 08:09 PM - lukemurphey

Moving parameters to constructor & function calls

Reference #1809

Revision 407 - 04/04/2017 08:09 PM - lukemurphey

Moving parameters to constructor & function calls

Reference #1809

Revision 408 - 04/05/2017 06:10 AM - lukemurphey

Refactored controller and input

Redesigning the controller and input to work with the refactored code

Reference #1809

History

#1 - 04/04/2017 05:25 PM - Luke Murphey

Some options:

1. Make a separate scraper class
2. Pass arguments as a config object
3. Pass arguments as *args or **kwargs

#2 - 04/04/2017 05:34 PM - Luke Murphey

Here are the functions invoked by scrape_page():

- resolve_proxy_type
- get_result_single
  - get_result_built_in_client
    - detect_encoding
  - get_result_browser
    - get_display
    - get_firefox_profile
    - add_auth_to_url
  - unescape
  - escape_field_name
  - extract_links
    - cleanup_link

#3 - 04/04/2017 05:37 PM - Luke Murphey

Here are the functions that are modular input specific:

- get_file_path
- run

I think the best approach is to

1. break out the modular input and leave everything else in the scraper
2. make the arguments to scrape_page class variables
3. updated the controller, search command, tests

#4 - 04/04/2017 05:41 PM - Luke Murphey

What should be in the constructor call versus the command call?

- **basics:** url, username, password, etc. (in scrape_page call)
- **output customization:** output_matches_as_mv, output_matches_as_separate_fields, etc. (in constructor)
- **charset detection:** charset_detect_meta_enabled, etc.
- **proxy info:** proxy_type, etc.
- **spider info:** page_limit, etc. (in constructor)

#5 - 04/04/2017 08:08 PM - Luke Murphey

Phases:

1. Break out WebScraper class
2. Add constructor options for proxy info and charset
    1. Modify get_result_browser, get_result_single, get_result_built_in_client, get_http_client
3. Update search command

4. Update modular input
5. Update controller

#6 - 04/05/2017 06:48 AM - Luke Murphey

- Status changed from New to Closed

- Target version set to 4.2

- % Done changed from 0 to 100