Web Scraper Documentation

Async Scraper

A modular and scalable asynchronous web scraper built on top of beautifulsoup.

Installation

```
from scraper import Scraper
```

Documentation

```
Scraper(initial_base_url, pool_number)
```

Params

- String: `initial_base_url` : base url of first page to be scrapped.
- Integer: `pool_number` : number of websites to visit concurrently at a time.

Return

- Object: instance of Scraper class.

```
Scraper.scrape(initial_url, body_config, nav_config, data_config)
```

Params

- String `initial_url` : the webpage that the scraper will first visit.
- Dictionary `body_config` : dictionary pointing the scraper to where the thread posts are.

```
body_config = {
    'path': [ # ancestors of where the data needed is.
        {
            'selector': 'table',
            'options': {'class': 'forumline'} # specific css / html attribute selection.
        },
    ],
    'selector': 'tr', # where the data is specifically after traversing ancestors.
    'options': {}
}
```

- Dictionary `nav_config` : dictionary pointing the scraper to where the navigation is.

```
nav_config = {
    'path': [
        { 'selector': 'td', 'options': {'align': 'right'} },
        { 'selector': 'span', 'options': {'class': 'nav'} },
    ],
    'selector': 'a',
    'options': {},
    'attribute': 'href' # specific part of the element to extract information.
}
```

- Dictionary `data_config` : dictionary pointing the scraper to each needed data point in the body.

```
data_config = {
    'post_id': { # field the data will be saved in result record object
        'path': [
            { 'selector': 'span', 'options': { 'class': 'name' } }
        ],
        'selector': 'a',
        'options': {},
        'attribute': 'name',
        'transform': lambda res: res[0] # optional custom function to
manipulate response
    },
}
```

Return

- List: list of all records scrapped from the website.

```
[
  {
      'post_body': [u'message', u'messageB'], # list of all requested data
found within thread
      'post_date': [u'tues'],
      'post_id': [u'87120'],
      'user_name': [u'Rick']
  }
]
```

Example

Implementation

```
from scraper import Scraper

body_config = {
    'path': [
        {
            'selector': 'table',
            'options': {'class': 'forumline'}
        },
    ],
    'selector': 'tr',
    'options': {}
}

data_config = {
    'post_id': {
        'path': [
            { 'selector': 'span', 'options': { 'class': 'name' } }
        ],
        'selector': 'a',
        'options': {},
        'attribute': 'name',
        'transform': lambda res: res[0]
    },
    'user_name': {
        'path': [
            { 'selector': 'span', 'options': { 'class': 'name' } }
        ],
        'selector': 'b',
        'options': {},
        'attribute': 'text'
    },
    'post_date': {
        'path': [
            { 'selector': 'td', 'options': { 'width': '100%' } }
        ],
        'selector': 'span',
        'options': {'class': 'postdetails'},
        'attribute': 'text'
    },
    'post_body': {
        'path': [],
        'selector': 'span',
        'options': { 'class': 'postbody' },
        'attribute': 'text'
    },
    'quote': {
        'path': [],
        'selector': 'td',
        'options': { 'class': 'quote' },
        'attribute': 'text'
    }
}

nav_config = {
    'path': [
        { 'selector': 'td', 'options': {'align': 'right'} },
        { 'selector': 'span', 'options': {'class': 'nav'} },
    ],
    'selector': 'a',
    'options': {},
    'attribute': 'href'
}



scraper = Scraper('http://www.oldclassiccar.co.uk/forum/phpbb/phpBB2/', 5)
data = scraper.scrape('viewtopic.php?t=12591', body_config, nav_config,
data_config)
```

Response

```
[
  {
      'post_body': [u'message', u'messageB'], # list of all requested data
found within thread
      'post_date': [u'tues'],
      'post_id': '87120', # transformed response
      'user_name': [u'Rick']
  }, ...
]
```