AJAXなサイトをスクレイピングしてみた

中橋 研太郎

- twitter: @k_nakahashi
- github: @nakahashi
- qiita: @nakahashi

- 2003年: 組込みシステムの受託開発な会社に入社
- 2016年5月: UUUM（株）入社

単純なサイトでのスクレイピング

```
use GuzzleHttp\Client;
use Symfony\Component\DomCrawler\Crawler;

$response = $this->client->get($url);
$html = $response->getBody()->getContents();
$crawler = new Crawler($html);

$crawler = $crawler->filter('body > p');
dump($crawler->text());
```

ログイン後の画面でのスクレイピング

```
use Goutte\Client;

$crawler = $this->client->request('GET', 'ログイン後に行きたいサイトのURL');
$html = $crawler->html();

$crawler = new Crawler('', 'ログイン時にフォームのデータをポストしたいURL');
$crawler->addHtmlContent($html);

$form = $crawler->selectButton('signIn')->form();
$form->setValues(['Email' => 'hoge@fuga.com')]);
$crawler = $this->client->submit($form);
```

AJAXな画面でのスクレイピング

ブラウザになりきるしかない？

Selenium ? casperjs ?

```
$this->client->request('POST', 'https://www.hoge.com/api/fuga/fizz',
    [], [], ['HTTP_CONTENT_TYPE' => 'application/javascript; charset=UTF-
    json_encode([
        "method" => "Search",
        "params" => [
            null, hoge, $id, null, null, null, null, null, null,
            null, null, null, null, 3, 0, 25
        ],
        "xsrf" => 'jkl_9LZ5AJfZfcDwRcIPYrLXbWMdBTu2Vw:1468094843195'
    ])
);

$content = json_decode($this->client->getResponse()->getContent());
$mcnName = $content->result[2][0][8][0][13][11][1][0][2];
```

結論

どんどんサイトの構造への依存度が深くなる(>_<

できればスクレイピングやめたい