Crowdsourcing in Question and Answer Systems

Question, Answer and SNS data spider

Project Details: **https://github.com/KeithYue/QA-spider**

Created by **WANG Yue**, HKUST

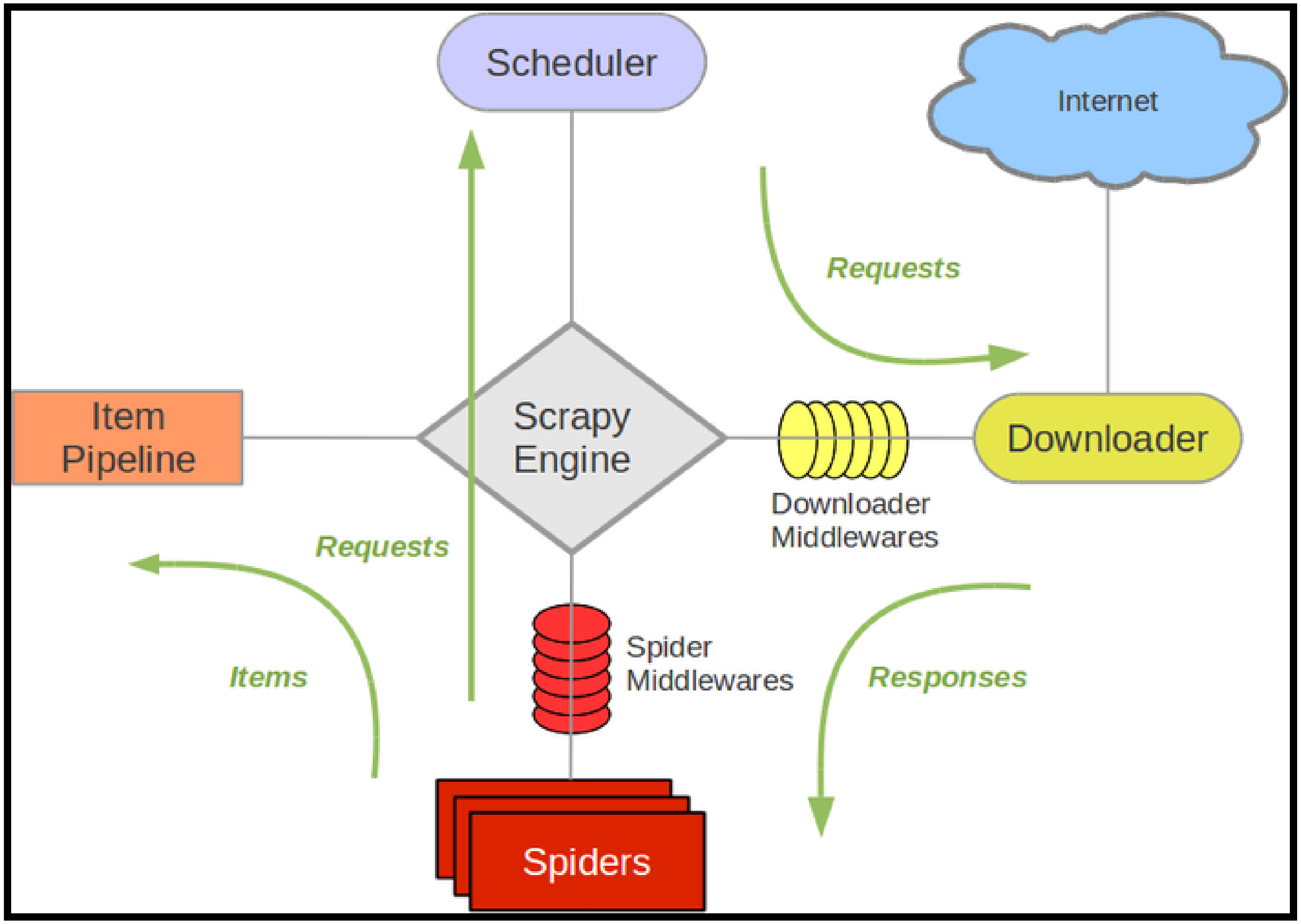
Done

- Yahoo! Answers Spider: crawls Yahoo! Answers data.
- Lazytweets Spider: crawls data of **http://www.lazytweet.com/**.
- Twitter-Archiever Script: archieve all tweets given a specific twitter user.
- Twitter-Following Sciprt: get all the following information of a user.

Spiders Included:

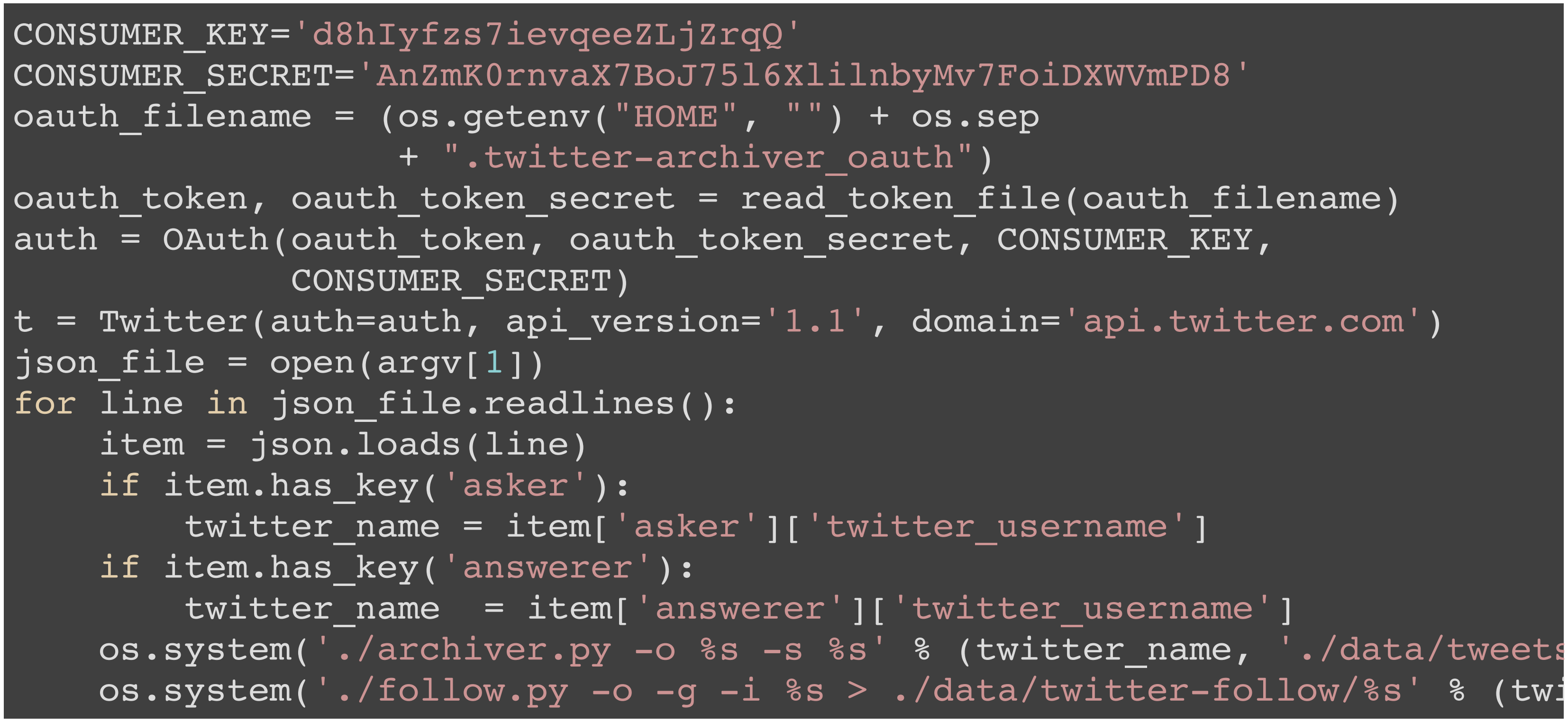
1. Define the item structure would be extracted
2. Wrote spider: define url patterns, rules and the function of parsing
3. Use XPath to parse the html structure to find item
4. Item Pipeline: valid scraped data, check for duplicates, store data in file or dataset

Overview of Scrapy



Twitter API 1.1

1. REST API
2. OAuth Signed

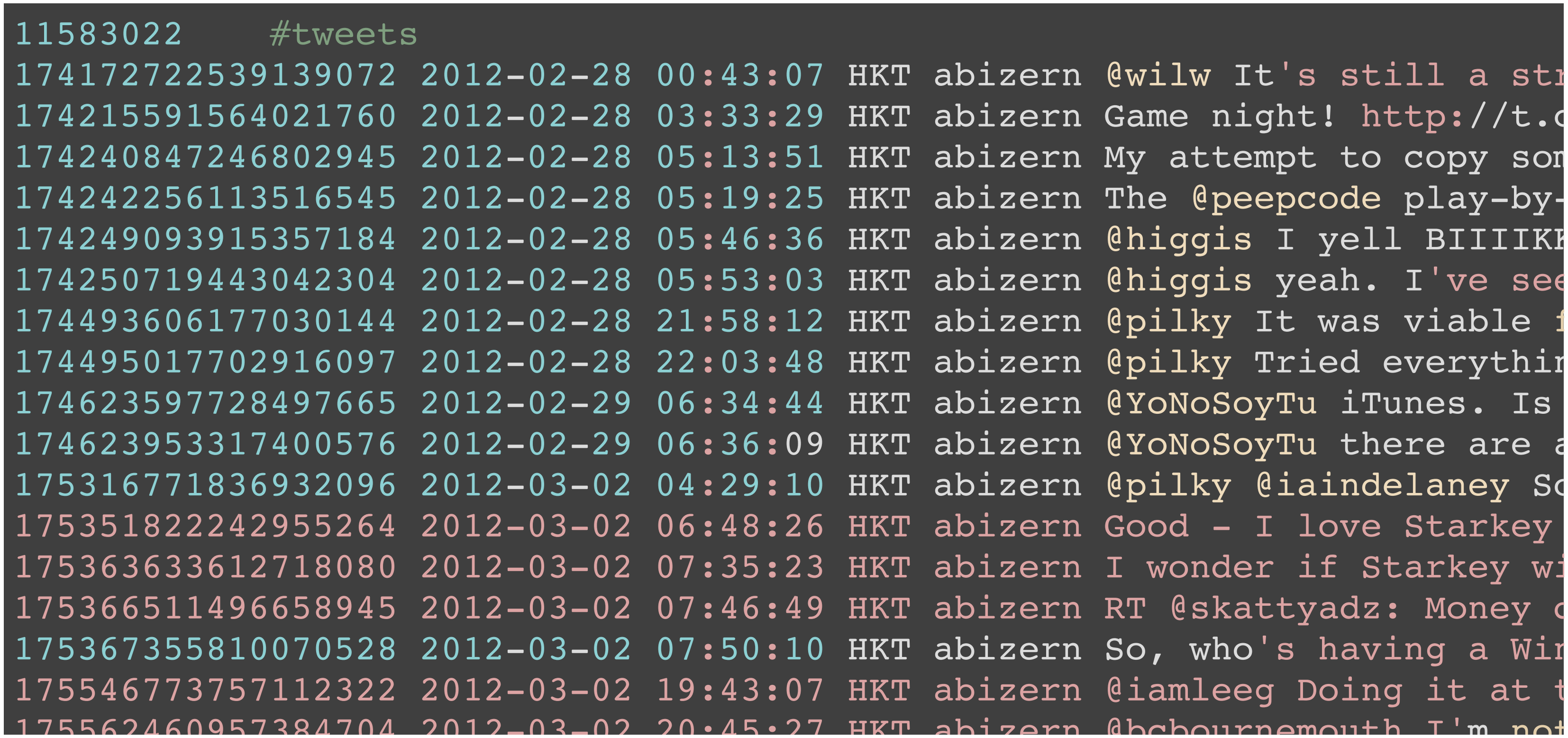


```
CONSUMER_KEY='d8hIyfzs7ievqeeZLjZrqQ'
CONSUMER_SECRET='AnZmK0rnvaX7BoJ75l6XlilnbyMv7FoiDXWVmPD8'
oauth_filename = (os.getenv("HOME", "") + os.sep
                  + ".twitter-archiver_oauth")
oauth_token, oauth_token_secret = read_token_file(oauth_filename)
auth = OAuth(oauth_token, oauth_token_secret, CONSUMER_KEY,
             CONSUMER_SECRET)
t = Twitter(auth=auth, api_version='1.1', domain='api.twitter.com')
json_file = open(argv[1])
for line in json_file.readlines():
    item = json.loads(line)
    if item.has_key('asker'):
        twitter_name = item['asker']['twitter_username']
    if item.has_key('answerer'):
        twitter_name  = item['answerer']['twitter_username']
    os.system('./archiver.py -o %s -s %s' % (twitter_name, './data/tweets
    os.system('./follow.py -o -g -i %s > ./data/twitter-follow/%s' % (twi
```

Sample of crawled question data

```
{"question_tags": ["apple", "osx"], "question_content": "Anyone know of a
{"question_tags": ["vine"], "question_content": "Why is my activity tab a
{"question_tags": ["angularjs"], "question_content": "Dear Lazyweb, is te
```

Sample of twitter user data



```
11583022    #tweets
174172722539139072 2012-02-28 00:43:07 HKT abizern @wilw It's still a str
174215591564021760 2012-02-28 03:33:29 HKT abizern Game night! http://t.c
174240847246802945 2012-02-28 05:13:51 HKT abizern My attempt to copy son
174242256113516545 2012-02-28 05:19:25 HKT abizern The @peepcode play-by-
174249093915357184 2012-02-28 05:46:36 HKT abizern @higgis I yell BIIIIKK
174250719443042304 2012-02-28 05:53:03 HKT abizern @higgis yeah. I've see
174493606177030144 2012-02-28 21:58:12 HKT abizern @pilky It was viable f
174495017702916097 2012-02-28 22:03:48 HKT abizern @pilky Tried everythin
174623597728497665 2012-02-29 06:34:44 HKT abizern @YoNoSoyTu iTunes. Is
174623953317400576 2012-02-29 06:36:09 HKT abizern @YoNoSoyTu there are a
175316771836932096 2012-03-02 04:29:10 HKT abizern @pilky @iaindelaney So
175351822242955264 2012-03-02 06:48:26 HKT abizern Good - I love Starkey
175363633612718080 2012-03-02 07:35:23 HKT abizern I wonder if Starkey wi
175366511496658945 2012-03-02 07:46:49 HKT abizern RT @skattyadz: Money c
175367355810070528 2012-03-02 07:50:10 HKT abizern So, who's having a Win
175546773757112322 2012-03-02 19:43:07 HKT abizern @iamleeg Doing it at t
175562460957384704 2012-03-02 20:45:27 HKT abizern @bcbournemouth I'm not
```

Todo

- Improve the twitter-tool set to crawl Quora dataset in a more general way.
- Find another dataset besides Lazytweets.
- Build Java Development Environment.(Python is not suitable for time-consuming calculation)

Issue

- Spider Persistence: pause and resume of a spider, state storage.
- Twitter API Rate limit exceeded
  - Alternative way: screen scraping(Ajax situation)
  - Twitter Streaming API

Open

- Deploy spiders in server.
- Get a XPath expression of a item using developer tools in Chrome or Firefox

End