pyspider

github.com/binux/pyspider

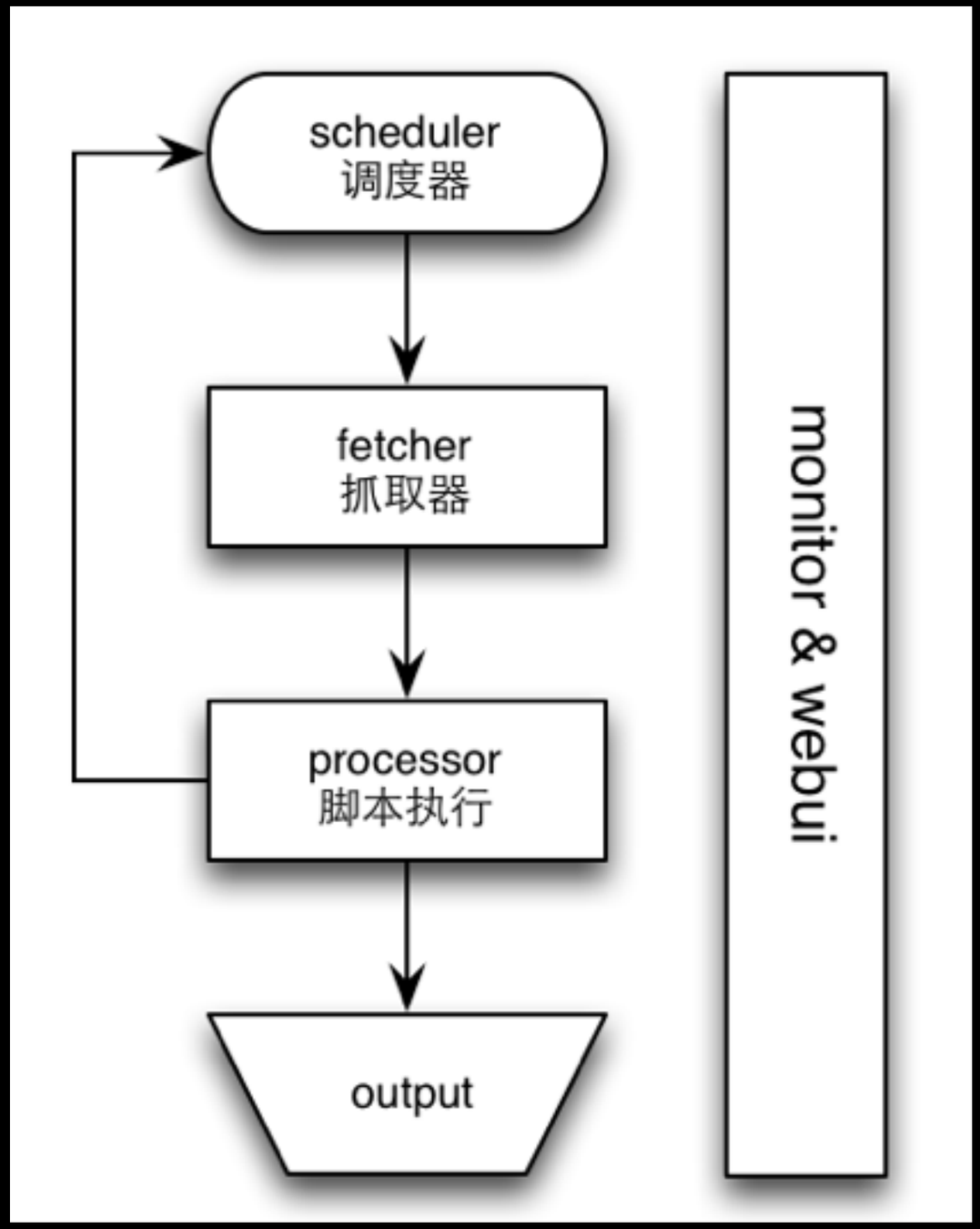
Binux(足兆叉虫)

来源于真实的垂搜引擎

- 100个站点
  - 脚本驱动
  - 任务管理、模板失效监控
  - 运行状态监控
- 5分钟内更新
  - 定时任务
  - 根据最近更新时间调度

pyspider 功能架构

- Python脚本驱动
- WebUI
- MySQL，MongoDB，SQLite 持久化后端
- 组件可替换、单机/分布式、Docker
- 强大的调度机制
- 支持JavaScript页面



```python
from libs.base_handler import *

class Handler(BaseHandler):

    @every(seconds=30)
    def on_start(self):
        self.crawl('http://www.douban.com/group/haixiuzu/discussion', callback=self.index_page)

    @config(age=10)
    def index_page(self, response):
        for each in response.doc('.title a[href^="http://"]').items():
            self.crawl(each.attr.href, callback=self.detail_page)

    @config(age=30*24*60*60)
    def detail_page(self, response):
        return {
            "url": response.url,
            "title": response.doc("#content h1").text(),
            "author": response.doc(".topic-content .from a").text(),
            "author_url": response.doc("DIV.topic-doc>H3>SPAN.from>A").attr.href,
            "imgs": [x.attr.src for x in response.doc('.topic-doc img').items()]
        }
```

demo.pyspider.org

processor - 脚本执行

- 完全的python
- Web下编写，Web下调试
- 通过API完全控制调度、抓取
- 脚本间通信、调用

fetcher - 抓取器

- 基于 tornado 的异步抓取
- 完整的抓取控制，从 method 到 timeout
- 支持JavaScript执行渲染（ 通过 phantomjs ）

scheduler - 调度器

- 任务优先级
- 流量控制
- 周期定时任务
- 按照过期时间调度
- 按照前链标记调度（例如更新时间）
- 失败重试

```
{
  "fetch": {
    "fetch_type": "js"
  },
  "process": {
    "callback": "detail_page"
  },
  "project": "js_test_sciencedirect",
  "taskid": "0915162322318ebba200ad0feb01d6ed",
  "url": "http://www.sciencedirect.com/science/article/pii/S0261560602000463"}
```

run

< | >


```
#!/usr/bin/env python
# -*- encoding: utf-8 -*-
# vim: set et sw=4 ts=4 sts=4 ff=unix fenc=utf8:
# Created on 2014-10-31 13:05:52

import re
from libs.base_handler import *

class Handler(BaseHandler):
    '''
    this is a sample handler
    '''
    def on_start(self):

self.crawl('http://www.sciencedirect.com/science/article/pii/S1568494612005741',
                   callback=self.detail_page)

    def index_page(self, response):
        for each in response.doc('a').items():
            if re.match('http://www.sciencedirect.com/science/article/pii/\w+$',
each.attr.href):
                self.crawl(each.attr.href, callback=self.detail_page)

    @config(fetch_type="js")
    def detail_page(self, response):
        self.index_page(response)
        self.crawl(response.doc('HTML>BODY>DIV#page-
area>DIV#rightPane>DIV#rightOuter>DIV#rightInner>DIV.innerPadding>DIV#recommend_rela
ted_articles>OL#relArtList>LI>A.viewMoreArticles.cLink').attr.href,
callback=self.index_page)

        return {
                "url": response.url,
                "title": response.doc('HTML>BODY>DIV#page-
area>DIV#centerPane>DIV#centerContent>DIV#centerInner>DIV#frag_1>H1.svTitle').text(),
                "authors": [{"name": x.text(), "url": x.attr.href} for x in
response.doc('HTML>BODY>DIV#page-
area>DIV#centerPane>DIV#centerContent>DIV#centerInner>DIV#frag_1>UL.authorGroup.noCo
llab>LI.smh5>A.authorName').items()],
                "abstract": response.doc('HTML>BODY>DIV#page-
area>DIV#centerPane>DIV#centerContent>DIV#centerInner>DIV#frag_2>DIV.abstract.svAbst
ract>P').text(),
 {'abstract': '',  "keywords": [x.text() for x in response.doc('HTML>BODY>DIV#page-
area>DIV#centerPane>DIV#centerContent>DIV#centerInner>DIV#frag_2>UL.keyword>LI.svKey
words>SPAN').items()],
  'authors': [{'name': 'J.R Lothian',
                'url': 'http://www.sciencedirect.com/science/article/pii/S02615606020004
  'keywords': [],
  'title': 'Editorial',
  'url': u'http://www.sciencedirect.com/science/article/pii/S0261560602000463'}
```

save

脚本编辑和调试

Dashboard

group	project name	status	rate/burst	progress				actions
lock	douban	RUNNING	0.2/3.0	5m	1h	1d	all: 16451	Run Active Tasks Results
lock	js_test_sciencedirect	RUNNING	0.1/3.0	5m: 60	1h: 718	1d: 17218	all: 78689	Run Active Tasks Results
lock	huodongxing	RUNNING	0.2/3.0	5m	1h	1d: 1446	all: 1272	Run Active Tasks Results
lock	yongle	RUNNING	0.2/3.0	5m: 2	1h: 240	1d: 5283	all: 43	Run Active Tasks Results
[group]	fa	CHECKING	0.2/3.0	5m	1h	1d: 1	all: 66	Run Active Tasks Results
lock	haixiuzu	RUNNING	0.2/3.0	5m: 26	1h: 274	1d: 6725	all: 7830	Run Active Tasks Results
[group]	522	DEBUG	0.2/3.0	5m	1h	1d	all: 48	Run Active Tasks Results
xu	test	TODO	0.2/3.0	5m	1h	1d	all	Run Active Tasks Results
[group]	9sep	STOP	0.2/3.0	5m	1h	1d	all: 2	Run Active Tasks Results

Recent Active Tasks

Create

- 任务列表
- 任务状态
- 流量配额
- 最近5分钟、1小时、1天、总任务计数

SUCCESS js_test_sciencedirect > http://www.sciencedirect.com/science/article/pii/S0093691X09004713 **13 seconds ago** 3279.97ms +45

SUCCESS js_test_sciencedirect > http://www.sciencedirect.com/science/article/pii/S0378432006003800 **23 seconds ago** 2886.91ms +42

SUCCESS haixiuzu > http://www.douban.com/group/haixiuzu/discussion **26 seconds ago** 634.65ms +26

SUCCESS js_test_sciencedirect > http://www.sciencedirect.com/science/article/pii/S0093691X05004334 **33 second**

SUCCESS js_test_sciencedirect > http://www.sciencedirect.com/science/article/pii/S0093691X11003499 **43 second**

SUCCESS js_test_sciencedirect > http://www.sciencedirect.com/science/article/pii/S0378432003001143 **1 minute a**

SUCCESS haixiuzu > https://api.duoshuo.com/posts/import.json#65927264 **1 minute ago** 1378.66ms +0

SUCCESS haixiuzu > http://www.douban.com/group/topic/65927264/ **1 minute ago** 476.68ms +1

SUCCESS js_test_sciencedirect > http://www.sciencedirect.com/science/article/pii/S037842661000333X **1 minute a**

SUCCESS js_test_sciencedirect > http://www.sciencedirect.com/science/article/pii/S0922142504000027 **1 minute a**

SUCCESS js_test_sciencedirect > http://www.sciencedirect.com/science/article/pii/S0093691X06001452 **1 minute a**

SUCCESS js_test_sciencedirect > http://www.sciencedirect.com/science/article/pii/S0261560602000190 **2 minutes**

SUCCESS js_test_sciencedirect > http://www.sciencedirect.com/science/article/pii/S0093691X9700407X **2 minutes**

SUCCESS js_test_sciencedirect > http://www.sciencedirect.com/science/article/pii/S0093691X09002866 **2 minutes**

SUCCESS js_test_sciencedirect > http://www.sciencedirect.com/science/article/pii/S037843201000480X **2 minutes**

SUCCESS js_test_sciencedirect > http://www.sciencedirect.com/science/article/pii/S0093691X11002779 **2 minutes**

[redacted]

SUCCESS haixiuzu.detail_page > http://www.douban.com/group/topic/65926822/ (2 minutes ago crawled)

taskid

```
9d9d5ca868a54c04bd43820729d987bc
```

lastcrawltime

```
1415540551.57 (2 minutes ago)
```

updatetime

```
1415540551.57 (2 minutes ago)
```

track.fetch ✔ 227.26ms


```
{
  "content": null,
  "encoding": "utf-8",
  "headers": {
    "Cache-Control": "must-revalidate, no-cache, private",
    "Connection": "keep-alive",
    "Content-Encoding": "gzip",
    "Content-Length": "9469",
    "Content-Type": "text/html; charset=utf-8",
    "Date": "Sun, 09 Nov 2014 13:42:31 GMT",
    "Expires": "Sun, 1 Jan 2006 01:00:00 GMT",
    "Keep-Alive": "timeout=10",
    "P3p": "CP=\"IDC DSP COR ADM DEVi TAIi PSA PSD IVAi IVDi CONi HIS OUR IND CNT\"",
    "Pragma": "no-cache",
    "Server": "nginx",
    "Set-Cookie": "bid=\"O4/TBTgolH0\"; path=/; domain=.douban.com; expires=Mon, 09-Nov-2015 13:42:31 GMT",
```



```
",
lTgolH0",
ovie.tv-chart-gaia"
```


[redacted]

js_test_sciencedirect - Results

JSON | URL-JSON | CSV

url	abstract	authors	keywords	title	url	...
http://www.sciencedirect.com/science/article/pii/S037843200700070X	"Skim milk (SM) is considered to be the most widely employed extender for goat sperm used for ...	[{"name": "L. Mara", "url": "http://www.sciencedirect.com/science/article/pii/S037843200700070X"}, ...	["buck semen", "dilution", "storage", "fertility"]	"Effect of different diluents on goat semen fertility"	"http://www.sciencedirect.com/science/article/pii/S037843200700070X"	{}
http://www.sciencedirect.com/science/article/pii/S104402839690006X	""	[{"name": "Geraldo M. Vasconcellos", "url": ...	[]	"Factors affecting cross-border mergers and acquisitions: The Canada-U.S. experience"	"http://www.sciencedirect.com/science/article/pii/S104402839690006X"	{}
http://www.sciencedirect.com/science/article/pii/S0378426612001185	"We use the CoVaR approach to identify the main factors behind systemic risk in a set of large ...	[{"name": "Germán López-Espinosa", "url": ...	["C30", "G01", "G20", "Systemic importance", "Liquidity risk", "Macroprudential regulation"]	"Short-term wholesale funding and systemic risk: A global CoVaR approach"	"http://www.sciencedirect.com/science/article/pii/S0378426612001185"	{}
http://www.sciencedirect.com/science/article/pii/S0093691X06000914	"Plasma concentrations of luteinizing hormone (LH) and follicle stimulating hormone (FSH) were ...	[{"name": "J. de Gier", "url": ...	["Canine", "Gonadotropins", "Estrous cycle", "Progesterone", "Dog"]	"Differential regulation of the secretion of luteinizing hormone and follicle-stimulating hormone ...	"http://www.sciencedirect.com/science/article/pii/S0093691X06000914"	{}
http://www.sciencedirect.com/science/article/pii/S1090023304002618	"We have investigated the effects of the timing of progesterone supplementation on early embryo ...	[{"name": "G.E. Mann", "url": ...	["Cow", "Progesterone", "Embryo", "Pregnancy", "Interferon-τ"]	"Effects of time of progesterone supplementation on embryo development and interferon-τ production ...	"http://www.sciencedirect.com/science/article/pii/S1090023304002618"	{}
http://www.sciencedirect.com/science/article/pii/S0019850108001661	"This study utilized structural equations modeling (SEM) to explore the positive effects of ...	[{"name": "Yu-Shan Chen", "url": ...	["Relationship learning", "Absorptive capacity", "Innovation performance", "Competitive advantage"]	"The positive effects of relationship learning and absorptive capacity on innovation performance ...	"http://www.sciencedirect.com/science/article/pii/S0019850108001661"	{}
http://www.sciencedirect.com/science/article/pii/S0378432005001831	"There are several hormones and local testicular factors involved in the initiation and control of ...	[{"name": "Monna F. Hess", "url": ...	["Equine", "Leydig cells", "LH", "GH", "IGF-I"]	"A comparison of the effects of equine luteinizing hormone (eLH), equine growth hormone (eGH) and ...	"http://www.sciencedirect.com/science/article/pii/S0378432005001831"	{}
http://www.sciencedirect.com/science/article/pii/S0093691X0900288X	"In mammals, sperm ascension within the female reproductive tract involves a transient adhesion to ...	[{"name": "R. Talevi", "url": ...	["Oviduct", "Sperm", "Sperm reservoir", "Adhesion", "Release"]	"Molecules involved in sperm-oviduct adhesion and release"	"http://www.sciencedirect.com/science/article/pii/S0093691X0900288X"	{}
http://www.sciencedirect.com/science/article/pii/	"Although controversy surrounds cloning efforts, the cloning of animals to a	[{"name": "Oliver A Ryder", "url": ...	["cloning", "endangered species", "Mouflon, nuclear transfer", "oocyte	"Cloning advances and challenges for conservation"	"http://www.sciencedirect.com/science/article/pii/	{}

- 最近活动的任务
- 任务历史
- 产出结果

github.com/binux/pyspider

demo.pyspider.org

f.binux.me/ia/

基于多页面的模板规则自动学习

Adobe Flash Player 15.0.0.328 多语版 -

软件介绍

很多电脑刚装是没这个flash支持插件 必

软件名

大小

更新时间

简介

下载地址页链接

主配图URL

站内导航

TAG标签

访问次数

新建 TEXT 类型属性

新建 URL 类型属性

新建 NUM 类型属性

新建 NONUM 类型属性

新建 DATETIME 类型属性

新建 IMAGE 类型属性

新建 面包屑 类型属性

0 1 2 3 4 5 > gen_tpl test all

test test test test test test

```
http://www.xdowns.com/soft/1/16/2006/Soft_33666.html
      大小: 5.34 MB
      更新时间: 2014-10-18 14:50:55
      软件名: Adobe Flash Player 15.0.0.328 多语版 - 让你的浏览器能播放flash

http://www.xdowns.com/soft/1/16/2006/Soft_33713.html
      大小: 321 KB
      更新时间: 2006-10-24 0:00:03
      软件名: HTMLConvert V1.0.5 Build 0280_汉化绿色版_转换HTML文件到图像文件

http://www.xdowns.com/soft/1/16/2006/Soft_33881.html
      大小: 1.08 MB
      更新时间: 2010-10-12 5:41:43
      软件名: null
```

f.binux.me/ia/

拉手网 Lashou.com 北京 切换城市

自助餐 搜索

美食 火锅 自助餐 蛋糕 体检 美发 ktv 电影票 烤肉 旅游

全部商品分类 首页 身边团购 今日新单 美食 电影 酒店 旅游 购物

您的位置：北京团购 » 美食团购 » 火锅团购 » 海底捞1000元储值卡

【24店通用】海底捞

仅售880元，市场价1000元的海底捞官方储值卡，全国通用，全国包邮，刷卡不限次数，品味特色
您真正体验做上帝的感觉！

HaiDiLaoHotPot
海底捞火锅

¥880 ¥1000 8.8折

已售 178 | 3 人已评价 ★★★★★

有效期 2014-12-31

服务 随时退 过期退

数量 — 1 +

立即购买

商家地址
body>main>div>div.detail-content>div.detail-right>div.detail-info>div.de
海底捞（牡丹园店）

body>main>div>div.detail-content>div.detail-right>div.detail-info>div.de
北京海淀区花园东路2号

营业时间
body>main>div>div.detail-content>div.detail-right>div.detail-info>div.de
10:00-22:00

电话
body>main>div>div.detail-content>div.detail-right>div.detail-info>div.de
010-57895151

本单用户评价，消费评价
body>main>div>div.detail-content>div.detail-right>div.d
5.0

评价人次，满意占比
body>main>div>div.detail-content>div.detail-right>div.det
3，100%

满意占比
body>main>div>div.detail-content>div.detail-right>div.detail-info>div.d
100%

f.binux.me/ia/

github.com/binux/pyspider

f.binux.me/ia/