牛人计划-高级项目课（12）

牛客网
NOWCODER

- 爬虫简介
- pyspider安装
- CSS选择器
- MySQLdb
- pyspider实践

NOWCODER.COM
牛客网-中国最大IT笔试/面试题库

爬虫简介

手动：requests，beautifulsoup，urllib2，httpclient
入门：pyspider，webmagic
专业：scrapy，nutch

爬虫目标：

1. 搜索引擎
2. 新闻/图片聚合
3. 数据监控（微博）
4. 羞羞的目的

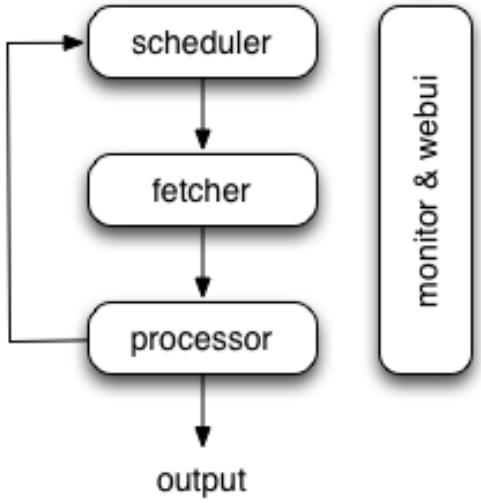
爬虫可以做什么： https://www.zhihu.com/question/27621722
BT种子爬虫：https://github.com/78/ssbc

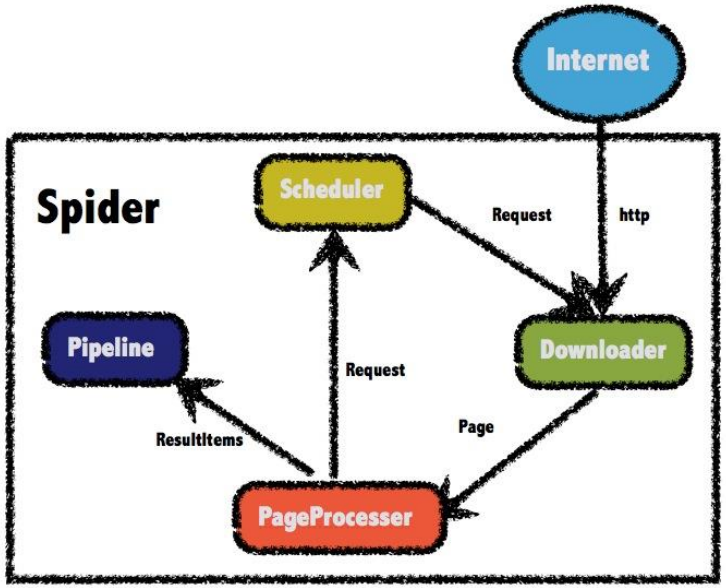
pyspider/webmagic

官网：http://docs.pyspider.org/en/latest/
在线测试：http://demo.pyspider.org/

安装：pip install pyspider
依赖包下载 http://www.lfd.uci.edu/~gohlke/pythonlibs/

启动 pyspider –c conf.json





NOWCODER.COM
牛客网-中国最大IT笔试/面试题库

pyspider例子

```
class Handler(BaseHandler):
    crawl_config = {
    }

    @every(minutes=24 * 60)
    def on_start(self):
        self.crawl('http://scrapy.org/', callback=self.index_page)

    @config(age=10 * 24 * 60 * 60)
    def index_page(self, response):
        for each in response.doc('a[href^="http"]').items():
            self.crawl(each.attr.href, callback=self.detail_page)

    @config(priority=2)
    def detail_page(self, response):
        return {
            "url": response.url,
            "title": response.doc('title').text(),
        }
```

NOWCODER.COM
牛客网-中国最大IT笔试/面试题库

Response/PyQuery

方法/属性	描述
doc	Html查找PyQuery对象
url	对应链接
text	文本
header	返回的header
cookies	下发cookie

参考资料: http://docs.pyspider.org/en/latest/apis/Response/

NOWCODER.COM
牛客网-中国最大IT笔试/面试题库

Processor/CSS选择器

选择器	描述
.class	class=“class”
#id	<p id=“id”>
div.inner	
a[href^=“http://”]	带http开头href的a元素
p div	p元素下的div元素（不必父子）
p>div>span	p元素下的div元素下的span
[target=_blank]	Target=_blank

PyQuery： https://pythonhosted.org/pyquery/api.html

参考资料：http://www.w3school.com.cn/cssref/css_selectors.asp

NOWCODER.COM
牛客网-中国最大IT笔试/面试题库

MySQLdb-insert

```
db = MySQLdb.connect('localhost', 'root', 'nowcoder', 'wenda',
charset='utf8')
try:
    cursor = db.cursor()
    sql = 'insert into question(title, content, user_id, created_date,
comment_count) values ("%s","%s",%d, %s, %d)' % (
    'title', 'content', random.randint(1, 10), 'now()', 0);
    # print sql
    cursor.execute(sql)
    qid = cursor.lastrowid
    db.commit()
    print qid
db.commit()

except Exception, e:
    print e
    db.rollback()
db.close()
```

NOWCODER.COM
牛客网-中国最大IT笔试/面试题库

MySQLdb-read

```
db = MySQLdb.connect('localhost', 'root', 'nowcoder', 'wenda',
charset='utf8')
try:
    cursor = db.cursor()
    sql = 'select * from question order by id desc limit 2'
    cursor.execute(sql)
    for each in cursor.fetchall():
        for row in each:
            print row
    db.commit()

except Exception, e:
    print e
    db.rollback()
db.close()
```

NOWCODER.COM
牛客网-中国最大IT笔试/面试题库

pyspider实践（代码演示）

课后练习

1. PyQuery练习
2. MySQLdb练习
3. 用户数据抓取

Thanks