

专栏：015：重构“你要的实战篇”

E - 爬虫技术

爬虫知识

用理工科思维看待这个世界

系列爬虫专栏

初学者，尽力实现最小化学习系统

主题：重构专栏：014 + Scrapy 实战 + sqlalchemy

0：目标说明

- Scrapy 基础教程
[你要的最佳实战](#)
- 刘未鹏博客
[点我啊](#)
- 目标：获取刘未鹏博客全站博文
 - 文章标题：Title
 - 文章发布时间：Time
 - 文章全文：Content
 - 文章的链接：Url
- 思路：
 - 分析首页和翻页的组成
 - 抓取全部的文章链接
 - 在获取的全部链接的基础上解析需要的标题，发布时间，全文和链接

之前的逻辑是starts_url 包括全部的1,2,3,4页，在这个的基础上进行提取各个网页的文章的所需字段。

scrapy 可以编写Rule 规则抓取需要的url

1：目标分解

编写的规则：

```

start_urls = ["http://mindhacks.cn/"]
rules = (Rule(SgmlLinkExtractor(allow=(r'http://mindhacks.cn/page/\d
+/',))),
        Rule(SgmlLinkExtractor(allow=(r'http://mindhacks.cn/\d{4,}/\d{2,}/
\d{2,}/.*?-.*?-.*?-.*?/')), callback='parse_detail', follow = True)
        )
# 前一个Rule获取的是1,2,3,4页的网页组成: 如: http://mindhacks.cn/page/2/
# 后一个Rule获取的1,2,3,4网页下符合要求的文章的链接, 再在获取的文章链接的基础上进行解析
如: http://mindhacks.cn/2009/07/06/why-you-should-do-it-yourself/

```

解析文本函数：

```

def parse_detail(self, response):
    Item = LiuweipengItem()
    selector = Selector(response)
    title = selector.xpath('//div[@id="content"]/div/h1[@class="entry-titl
e"]/a/text()').extract()
    time = selector.xpath('//div[@id="content"]/div/div[@class="entry-inf
o"]/abbr/text()').extract()
    content = selector.xpath('//div[@id="content"]/div/div[@class="entry-con
tent clearfix"]/p/text()').extract()
    url = selector.xpath('//div[@id="content"]/div/h1[@class="entry-title"]/
a/@href').extract()
    for title, time, content, url in zip(title, time, content, url):
        Item["Title"] = title
        Item["Time"] = time
        Item["Content"] = content
        Item["Url"] = url
    yield Item
# 返回的Item 是需要抓取字段

```

2 : ORM

参见：[专栏：012](#)

数据表声明

```
from sqlalchemy import Column, String, Integer
from sqlalchemy.ext.declarative import declarative_base
Base = declarative_base()
class Article(Base):
    __tablename__ = "article"
    id = Column(Integer, primary_key=True)
    Title = Column(String)
    Time = Column(String)
    Content = Column(String)
    Url = Column(String)
```

3 : 储存

再次说明scrapy 文件目录结构和作用：

- items.py : 抓取的目标，定义数据结构
- pipelines.py : 处理数据
- settings.py : 设置文件，常量等设置
- spiders/: 爬虫代码

所以储存操作：pipelines.py

需要在本地先创建数据库表：

```
CREATE TABLE `article` (
  `id` INT(11) NOT NULL AUTO_INCREMENT,
  `Title` VARCHAR(255) COLLATE utf8_bin NOT NULL,
  `Content` VARCHAR(255) COLLATE utf8_bin NOT NULL,
  `Time` VARCHAR(255) COLLATE utf8_bin NOT NULL,
  `Url` VARCHAR(255) COLLATE utf8_bin NOT NULL,
  PRIMARY KEY (`id`)
) ENGINE=INNODB AUTO_INCREMENT=39 DEFAULT CHARSET=utf8 COLLATE=utf8_bin
```

```

def open_spider(self, spider):
    engine = create_engine("mysql://root:123456@localhost:3306/test?charset=utf8", echo = True)
    DBSession = sessionmaker(bind=engine)
    self.session = DBSession()
    pass

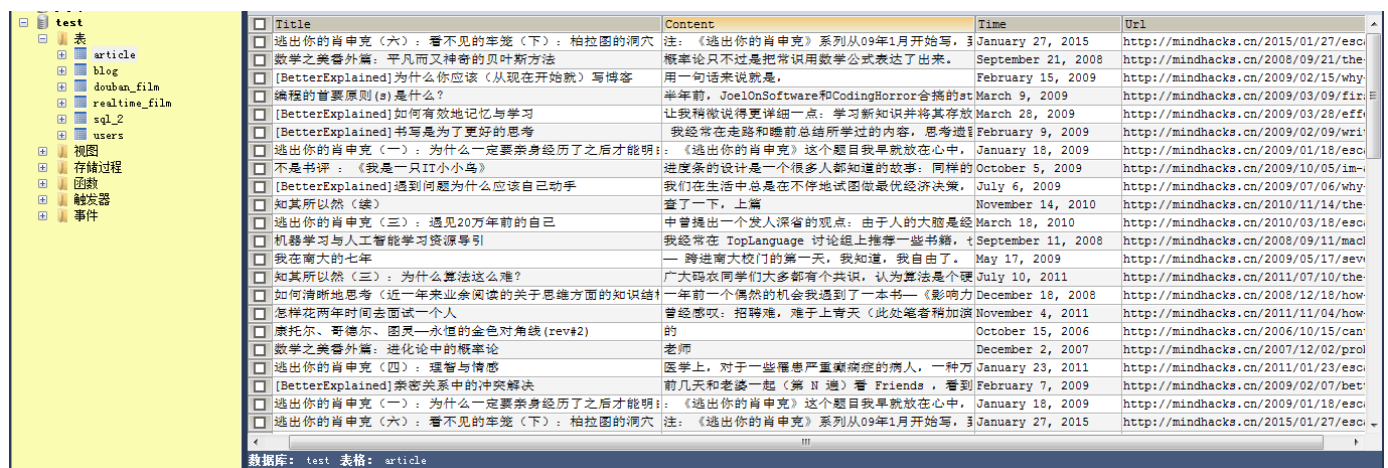
def process_item(self, item, spider):
    one = Article(Title=item["Title"],
                  Time=item["Time"],
                  Content=item["Content"],
                  Url=item["Url"])
    self.session.add(one)

    pass

def close_spider(self, spider):
    self.session.commit()
    self.session.close()
    pass

```

效果显示：



Title	Content	Time	Url
逃出你的肖申克（六）：看不见的牢笼（下）：柏拉图的洞穴	注：《逃出你的肖申克》系列从09年1月开始写，	January 27, 2015	http://mindhacks.cn/2015/01/27/esc
数学之美番外篇：平凡而又神奇的贝叶斯方法	概率论只不过是把常识用数学公式表达了出来。	September 21, 2008	http://mindhacks.cn/2008/09/21/the
[BetterExplained]为什么你应该（从现在开始就）写博客	用一句话来说就是，	February 15, 2009	http://mindhacks.cn/2009/02/15/why
编程的首要原则(s)是什么？	半年前，JoelOnSoftware和CodingHorror合搞的st	March 9, 2009	http://mindhacks.cn/2009/03/09/fir
[BetterExplained]如何有效地记忆与学习	让我稍微说得再详细一点：学习新知识并将其存放	March 28, 2009	http://mindhacks.cn/2009/03/28/eff
[BetterExplained]书写是为了更好的思考	我经常在走路和睡前总结所学过的内容。思考进	February 9, 2009	http://mindhacks.cn/2009/02/09/wri
逃出你的肖申克（一）：为什么一定要亲身经历了之后才能明	：《逃出你的肖申克》这个题目我早就放在心中。	January 18, 2009	http://mindhacks.cn/2009/01/18/esc
不是书评：《我是一只IT小小鸟》	进度条的设计是一个很多人都知道的故事：同样的	October 5, 2009	http://mindhacks.cn/2009/10/05/im-
[BetterExplained]遇到问题为什么应该自己动手	我们在生活中总是在不停地试图做最优经济决策，	July 6, 2009	http://mindhacks.cn/2009/07/06/why
知其所以然（续）	查了一下，上篇	November 14, 2010	http://mindhacks.cn/2010/11/14/the
逃出你的肖申克（三）：遇见20万年前的自己	中曾提出一个发人深省的观点：由于人的大脑是经	March 18, 2010	http://mindhacks.cn/2010/03/18/esc
机器学习与人工智能学习资源指引	我经常在 TopLanguage 论坛上推荐一些书籍，	September 11, 2008	http://mindhacks.cn/2008/09/11/maci
我在南大的七年	一 跨进南大校门的第一天，我知道，我自由了。	May 17, 2009	http://mindhacks.cn/2009/05/17/sev
知其所以然（三）：为什么算法这么难？	广大码农同学们大多都有个共识，认为算法是个硬	July 10, 2011	http://mindhacks.cn/2011/07/10/the
如何清晰地思考（近一年来业余阅读的关于思维方面的知识结	一年前一个偶然的机会我遇到了一本书——《影响力	December 18, 2008	http://mindhacks.cn/2008/12/18/how
怎样花两年时间志面一个新人	曾经感叹：招聘难，难于上青天（此处笔者稍加演	November 4, 2011	http://mindhacks.cn/2011/11/04/how
康托尔、哥德尔、图灵—永恒的金色对角线 (rev#2)	的	October 15, 2006	http://mindhacks.cn/2006/10/15/can
数学之美番外篇：进化论中的概率论	老师	December 2, 2007	http://mindhacks.cn/2007/12/02/prol
逃出你的肖申克（四）：理智与情感	医学上，对于一些罹患严重抑郁症的病人，一种万	January 23, 2011	http://mindhacks.cn/2011/01/23/esc
[BetterExplained]亲密关系中的冲突解决	前几天和老婆一起《第N遍》看 Friends，看到	February 7, 2009	http://mindhacks.cn/2009/02/07/bet
逃出你的肖申克（一）：为什么一定要亲身经历了之后才能明	：《逃出你的肖申克》这个题目我早就放在心中。	January 18, 2009	http://mindhacks.cn/2009/01/18/esc
逃出你的肖申克（六）：看不见的牢笼（下）：柏拉图的洞穴	注：《逃出你的肖申克》系列从09年1月开始写，	January 27, 2015	http://mindhacks.cn/2015/01/27/esc

- Tips

IDE下运行启动scrapy 爬虫：

新建任意一个文件：比如：main.py

```

from scrapy.cmdline import execute
execute("scrapy crawl name".split())

```

运行这个文件，就可以启动爬虫，其中name，是spiders文件下编写爬虫所对应的那个name

完整代码: [点不点都是代码](#)

4：总结和说明

参考文献：

强烈建议：[1](#)

强烈建议：[2](#)

Scrapy 爬虫框架还存在许多的未知...

[Scrapy各种实例](#)