

单个应用程序可以合理框架结构

② 用框架写高并发爬虫

③ scrapy 框架原理

[spiders] 爬取页面 方便爬取数据

1. 引擎开网站 → 2. 找链接/页面 → 3. 登录 爬取第一个页面

[scheduler] 优先队列

1. 拿到url, 请求/继续向下抓取 2. 把请求看返回的 cookie 保存返回的 cookie

[engine] scheduler

1. 调度, 把url给 engine (引擎)

[Downloader] 下载工具

1. 下载器真正下载, 把下载结果返回

[spiders]

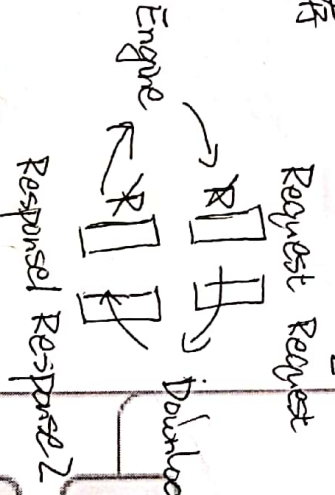
1. 把结果返回给 spiders, BS处理

1. 有读到磁盘

Scrapy C++ build tools

Settings 的功能 ① 动态延迟

② 缓存



不像是其他框架下可以爬取的 经过 SOLID 原则 (鸭子类型) 来设计 (鸭子类型)

scrapy 框架是什么

scrapy 框架的使用

url列表 回调函数 注册函数

callback 注册函数

Scrapy

深入 scrapy 框架剖析

运行爬虫 scrapy crawl 爬取

pipeline: 把请求的 items 通过 pipe

传输文件

Request Request

1

2

Engine

Response Response 2

Downloader

scrapy 的中间件

twisted 简介

Single Responsibility Principle - 一个类只有一个修改的理由

Open/Closed Principle - 对修改封闭 (封装)

Liskov Substitution Principle - 类的继承关系可替代

Interface Segregation Principle - 接口分离原则

Dependency Inversion Principle - 依赖倒置原则

scrapy 框架主要组件

引擎、调度器、下载器

调度器 scheduler → Downloader → spiders → pipeline

请求的url进入请求的队列里, 队列可以理解为列表, 列表按

请求的url和url处理成url后返回处理返回结果

框架的输出与调试

用 scrapy 框架爬豆瓣

选择器

setting 文件读取原理

非阻塞爬虫

系统, 根据请求

响应过程 scrapy crawl 爬取

登陆与 cookie Start-requests (self)

Cookies-Enabled = True

延迟请求与限速 Auto throttle-Enable

缓存 初始下载延迟 Auto throttle-Start

Filesystem Cache Storage 最大延迟 Auto throttle-Max

下载中间件 + 设置代理 IP + Scrapy!

爬虫中间件

爬虫中间件

爬虫中间件

爬虫中间件

爬虫中间件

爬虫中间件

爬虫中间件