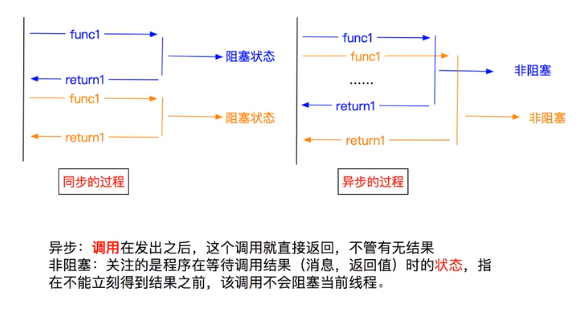
\*模块与框架

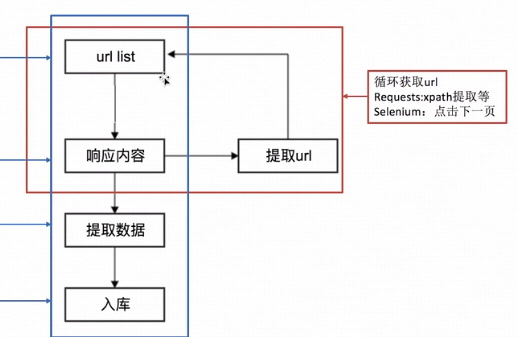
模块

框架内有许多模块，解决特定问题的解决方案

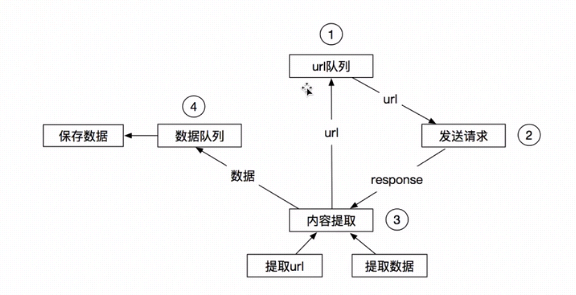
\*异步和非阻塞的区别

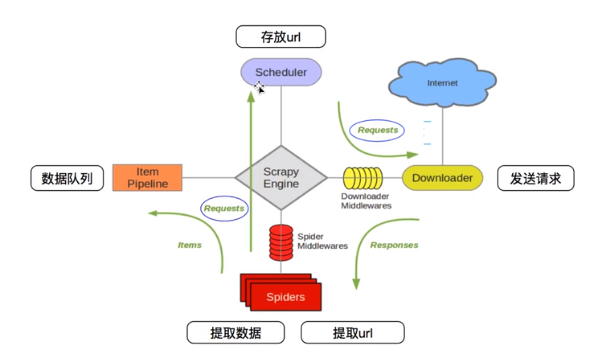


\*request爬虫流程



\*scrapy流程





\*Scrapy 入门

1. 创建一个项目

scrapy startproject projectName

1. 生成一个爬虫

scrapy genspider spider “targetURL”

1. 提取数据

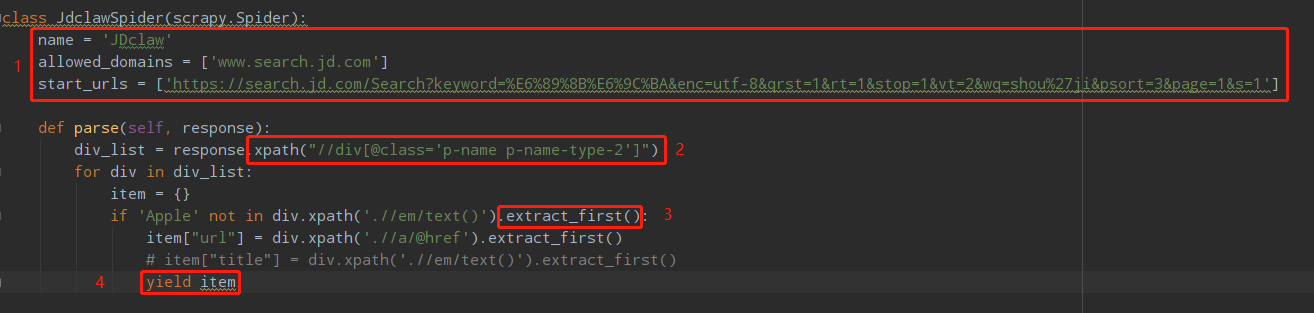
完善spider，使用xpath等方法

1. 保存数据

pipline中保存数据

1. 启动爬虫

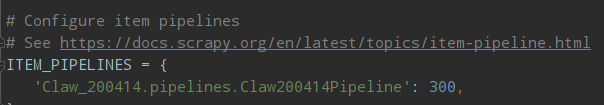
scrapy crawl spider



1. 定义爬虫名，允许爬取范围，起始爬取链接

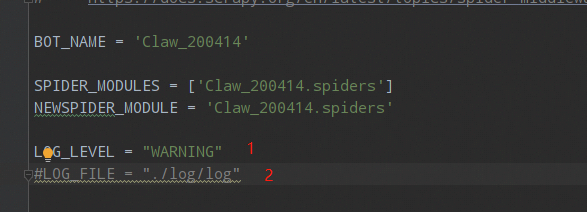
错误示范：  
 1 ）allowed\_domains = ['www.qiushibaike.com'] 主域名前面添加 www  
 2 ）allowed\_domains = ['qiushibaike.com/text']    主域名后面添加多余页码

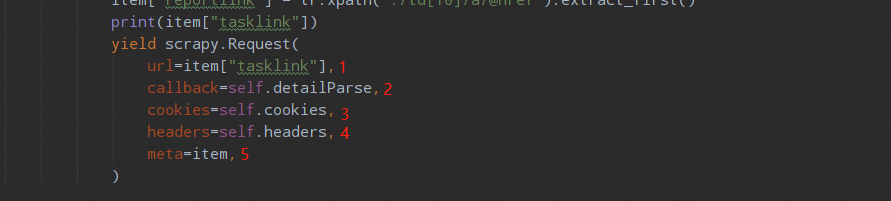
1. Scrapy提供Xpath方法返回selector对象
2. Selector提供extract方法返回含有字符串的列表,extract\_first方法返回列表内第一个元素，如没有返回None
3. 通过yield将item传递到piplines（需要先在settings.py 反注释下列代码，第一段为pipline方法名，第二段为优先级，值越小优先级越大）



日志设定

1. settings设置日志（1. 设置显示日志等级，2. 设置日志本地保存）





Scrapy.Request

1. url(str)必填:传入下一个请求URL
2. callback(function name):下一个解析函数方法名
3. cookie(dict):传入cookie参数
4. headers(dict):传入headers参数
5. meta(dict):实现在解析函数之间传输数据
6. dont\_filter(bool):是否筛选链接
7. method(str):请求方法