

1.engine : 控制数据流的流向,将任务进行统一的管理

2.spider: 对相应进行解析,然后,将内容返回给引擎

3.scheduler:调度器,从engine接受请求,并将请求排队,返回给引擎

4.downloader:下载器

5.item pipelines:数据项下载管道

使用scrapy startproject 项目名称 在命令所在的目录创建一个scrapy工程

推荐使用xpath规则.

// 从匹配选择的当前节点选择文档中的节点,而不考虑他们的位置

[1].[last()],[last()-1] 选取当前元素列表的第一个元素,最后一个元素,倒数第二个元素

[position() < 3] ,[position() > 5] 选取当前元素列表的前两个元素,

| 选取复合"或"选择两端的元素

response.xpath('lxml的路径').extract()

extract()可以将所有的文本提取出来

extract\_first()提取出来第一个元素

官方文档

<https://doc.scrapy.org/em/lastert/topics/spiders.html>

代码提供了很多简便的功能