```
#展示模板
scrapy genspider -1
```

```
PS C:\Users\23303\Desktop\scrapy相关\fang> scrapy genspider -l
Available templates:
basic
crawl
csvfeed
xmlfeed
PS C:\Users\23303\Desktop\scrapy相关\fang>
```

scrapy genspider -t crawl 爬虫名 首个请求的网址

这里还是以

```
https://www.xiaoshuopu.com
```

为例子

```
from scrapy.linkextractors import LinkExtractor
from scrapy.spiders import CrawlSpider, Rule
class BaiduSpider(CrawlSpider):
    name = "baidu"
    allowed_domains = ["xiaoshuopu.com"]
    start_urls = [f"https://www.xiaoshuopu.com/class_{type}/" for type in
range(1,4)
 Rule(
            LinkExtractor(
                allow=r'https://www.xiaoshuopu.com/xiaoshuo/\d+/\d+/',
                deny=r'https://www.xiaoshuopu.com/xiaoshuo/\d+/\d+/\d+\.html',
                # restrict_xpaths='//*[@id="at"]'
            ),
            callback='parse_books',
            follow=True
        ),
        Rule(
            LinkExtractor(
                allow=r'https://www.xiaoshuopu.com/xiaoshuo/d+/d+.html',
                restrict_xpaths='//*[@id="at"]'
            callback='parse_item',
            follow=False
        ),
    def parse_books(self, response):
        book_name = response.xpath('//h1/text()').extract()
        print('book_name:',book_name,response.url)
    def parse_item(self, response):
```

```
title = response.xpath('//*[@id="amain"]/dl/dd[1]/h1/text()').extract()
content = response.xpath('//*[@id="htmlContent"]//text()').extract()
print('title:',title,response.url)
# item = {}
#item["domain_id"] = response.xpath('//input[@id="sid"]/@value').get()
#item["name"] = response.xpath('//div[@id="name"]').get()
#item["description"] = response.xpath('//div[@id="description"]').get()
# return item
```

这里

allow=r'https://www.xiaoshuopu.com/xiaoshuo/\d+/\d+/',

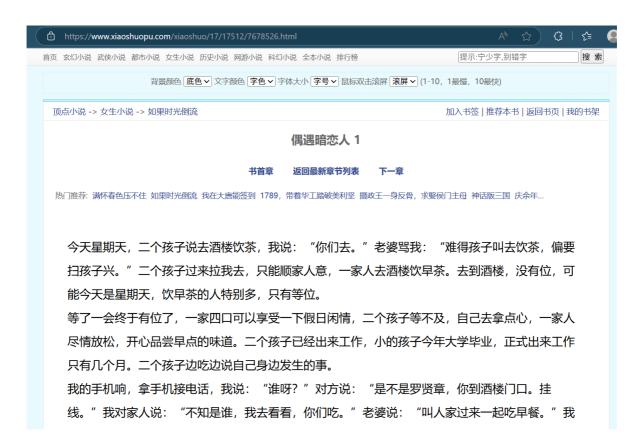
是打算匹配所有书对应的页面, 比如这种

https://www.xiaoshuopu.	com/xiaoshuo/17/17512/		A ^N 🖒 🕻 -	ć 🚇 ··
首页 玄幻小说 武侠小说 都市小说 女生小说 历史小说 网游小说 科幻小说 全本小说 排行榜 顶点小说-> 女生小说 ->如果时光倒流简介 加入			提示:宁少字,别错字	搜 索
			入书架 推荐本书 返回书页 我的书架 手机阅读	
	加里时光度	刨流 最新章节更新列表		
	TIOCCH TO THE	3/// 政机学 1-32417348		
	如果时光倒流作者: 风霜	雪伴我行 更新时间: 2025-05-09	15:00	
偶遇暗恋人 1	偶遇暗恋人 2	偶遇暗恋人 3	偶遇暗恋人 4	
接管工厂1	接管工厂 2	接管工厂 3	接管工厂 4	
人生百态 1	人生百态 2	人生百态 3	人生百态 4	
人生百态 5	人生百态 6	人生百态 7	人生百态 8	
亲情 1	亲情 2	亲情 3	亲情 4	
亲情 5	亲情 6	亲情 7	亲情 8	
女儿的婚事 1	女儿的婚事 2	女儿的婚事 3	女儿的婚事 4	
女儿的婚事 5	女儿的婚事 6	女儿的婚事 7	女儿的婚事 8	
女儿的婚事 9	女儿的婚事 10	女儿的婚事 11	女儿的婚事 12	
女儿的婚事 13	女儿的婚事 14	女儿的婚事	女儿的婚事 16	
女儿的婚事 17	女儿的婚事 18	女儿的婚事 19	女儿的婚事 20	
女儿的婚事 21	女儿的婚事 22	女儿的婚事 23	女儿的婚事24	
女儿的婚事 25	女儿的婚事 26	女儿的婚事 27	女儿的婚事 28	
女儿的婚事 29	女儿的婚事 30	女儿的婚事 31	女儿的婚事 32	
女儿的婚事 33	女儿的婚事 34	女儿的婚事 35	女儿的婚事 36	
女儿的婚事 37	女儿的婚事 38	女儿的婚事 39	女儿的婚事 40	
女儿的婚事 41	女儿的婚事 42	女儿的婚事 43	女儿的婚事 44	
同学之间的恩怨 1	同学之间的恩怨 2	同学之间的恩怨 3	同学之间的恩怨 4	
同学之间的恩怨 5	同学之间的恩怨 6	同学之间的恩怨 7	同学之间的恩怨 8	
同学之间的恩怨 9	同学之间的恩怨 10	同学之间的恩怨 11	同学之间的恩怨 12	
同学之间的恩怨 13	同学之间的恩怨 14	同学之间的恩怨 15	同学之间的恩怨 16	
同学之间的恩怨 17	同学之间的恩怨 18	同学之间的恩怨 19	同学之间的恩怨 20	

follow=True 代表可以继续跟进,也就是后面的规则允许在对应的页面进行规则匹配,这里我们还想获取这个页面上章节的链接然后请求,所以肯定是要跟进的

deny=r'https://www.xiaoshuopu.com/xiaoshuo/\d+/\d+/\d+.html',这个是说不允许访问什么,可以看到文章的详情页,这个链接符合了 allow=r'https://www.xiaoshuopu.com/xiaoshuo/\d+/\d+/\d+/',这个表达式的要求,但是实际我们不想要,那可以用deny排除,deny的优先级要比allow高。

https://www.xiaoshuopu.com/xiaoshuo/17/17512/7678526.html



之后我们可以单起一个rule来匹配详情页内的东西

```
Rule(
    LinkExtractor(
        allow=r'https://www.xiaoshuopu.com/xiaoshuo/\d+/\d+/\d+\.html',
        restrict_xpaths='//*[@id="at"]'
    ),
    callback='parse_item',
    follow=False
),
```

由于小说的详情页里没有具体的链接要继续抓取,所以这里的follow就改为false即可