简述：

实现对微信小程序社区，文章列表每个详情的爬取。

从列表第一页开始，对每个文章详情里的文章标题，作者，发布时间，内容四个字段进行爬取。然后下载到wxapp.json文件。

1. 创建项目

scrapy startproject wxapp

1. 创建CrawlSpider爬虫

scrapy genspider -t crawl wxapp\_spider "wxapp-union.com"

1. 设置setting.py的USER\_AGENT，ROBOTSTXT\_OBEY, DOWNLOAD\_DELAY

, ITEM\_PIPELINES参数

1. 设置items.py模板字段
2. 设置pipelines.py的下载位置，下载到项目根目录下的wxapp.json文件，代码如下

from scrapy.exporters import JsonLinesItemExporter  
  
class WxappPipeline(object):  
 def \_\_init\_\_(self):  
 self.fp = open("wxapp.json","wb")  
 self.exporter = JsonLinesItemExporter(self.fp, ensure\_ascii=False, encoding='utf-8')  
 def open\_spider(self,spider):  
 print("爬虫开始。。。")  
 def process\_item(self, item, spider):  
 self.exporter.export\_item(item)  
 return item  
 def close\_spider(self,spider):  
 self.fp.close()  
 print("爬虫结束。。。")

1. 编写爬虫wxapp\_spider

代码如下：

from scrapy.linkextractors import LinkExtractor  
from scrapy.spiders import CrawlSpider, Rule  
  
from wxapp.items import WxappItem  
from scrapy.http.request  
  
class WxappSpiderSpider(CrawlSpider):  
 name = 'wxapp\_spider'  
 allowed\_domains = ['wxapp-union.com']  
 start\_urls = ['http://www.wxapp-union.com/portal.php?mod=list&catid=2&page=1']  
  
 rules = (  
 Rule(LinkExtractor(allow=r'.+mod=list&catid=2&page=\d'),  
 follow=True),  
 Rule(LinkExtractor(allow=r'.+article-.+\.html'),callback="parse\_detail",  
 follow=False),  
 )  
  
 def parse\_detail(self, response):  
 print("\033[0;37;40m\t=============URL===============\033[0m")  
 print("\033[0;37;40m\t" + response.request.url + "\033[0m")  
  
 title = response.xpath("//div[@class='h hm cl']/div[@class='cl']/h1/text()").get().strip()  
 author = response.xpath("//div[@class='avatar\_right cl']//p[@class='authors']/a/text()").get().strip()  
 time = response.xpath("//div[@class='avatar\_right cl']//p[@class='authors']/span/text()").get().strip()  
 content = response.xpath("//div[@class='content\_middle cl']/div[@class='d']/table//td[@id='article\_content']//text()").getall()  
 content = "".join(content).strip()  
 item = WxappItem(title=title,author=author,time=time,content=content)  
 yield item

重点：

LinkExtractor:对匹配这个正则的url进行抓取。

Callback:回调函数，对抓取的url内容进行回调处理。本例中，对列表没有回调。

Follow:是否跟进匹配。当到达第十页的时候，是否继续匹配后面的页码，进行爬取。



