31日记录

问题

1. 如何动态生成类定义或实例[Python中动态创建类的方法 - 腾讯云开发者社区-腾讯云 (tencent.com)](https://cloud.tencent.com/developer/article/1176455)
   1. 使用 type 动态生成类
   2. 使用元类（metaclass）
   3. 使用\_\_new\_\_方法

工作

1. 进一步提高爬虫系统的可扩展性与可维护性
2. 进一步学习xpath，css等选择器

1日记录

注意

1. 字典按键赋值dict[‘’] = value，当见不存在时自动添加新键值对
2. 字典按键取值dict.get(‘’)，
3. 返回可遍历的(键, 值) 元组数：dict.items()，可用 for key, value in 遍历
4. 在 Item 中使用的 Field() 类继承于 dict
5. # Item 类自带一个属性-Field 字典，相当于该 Item的所有属性及其类型为 Field，修改这个字典相当于改变 Item 的属性名或数量[Scrapy 动态创建 Item - 知乎 (zhihu.com)](https://zhuanlan.zhihu.com/p/72711337)
6. Item 中只有Field 类型的属性才能被打印或装载

工作

1. 改变流程：粗爬取（爬虫系统）->细调整（管理系统）
2. 使Item通用化

问题

1. 数据中有许多多余的信息，在哪部分将其删除？如何将其删除？
2. 一个字段可能对应多个数据（之前没有考虑到），因此需要用数组存储，在哪部分拼接？如何拼接？答：1. 在爬虫系统拼接2. 在先全部用数组储存，然后再在管理界面拼接
3. 暂时无法爬取的网站：1. 有反扒机制2. 滚动刷新

2日记录

工作

1. 框架爬取链接的功能不灵活，重新编写换页方法，不使用scrapy提供的rule，实现控制爬取区间，可爬取无‘下一页’的网站
2. 在一个进程中运行多个spider[Common Practices — Scrapy 2.6.2 documentation](https://docs.scrapy.org/en/latest/topics/practices.html)
3. 将自定义爬虫环境配置从run.py（全局自定义）转移到spider（局部自定义）中（废弃，局部自定义环境配置要设置在spider的类变量custom\_settings中，而类变量要在类实例化前创建，与通用化相违背（类变量硬编写在spider中，而我需要动态改变自定义环境变量的值））

问题

1. 如何输入一段字符，自动生成匹配的通用正则表达式
2. 无法指定爬取的页数。解决方案：1.用户可构造分页链接，然后把分页链接当作start\_url来存放，爬取多少页就有多少个分页链接，可实现控制爬取区间[(11条消息) 关于Scrapy crawlspider rules的规则——翻页\_可乐乐乐乐乐的博客-CSDN博客](https://blog.csdn.net/joe8910/article/details/85159059)
3. 对于无“下一页“的页面无法使用rule进行递归爬取。解决方案：1.把分页链接当作start\_url来存放2.分页链接爬取规则设置为当前页的标签的下一个[(11条消息) 基于scrapy框架的爬虫详细步骤（含没有“下一页”按钮的href抓取）\_Dear Slim.的博客-CSDN博客\_scrapy 下一页](https://blog.csdn.net/weixin_43525427/article/details/96487294)
4. 无法爬取由ajax产生的内容
5. 通用性只提高到新闻列表页[(11条消息) 使用scrapy爬取新浪新闻\_silence cc的博客-CSDN博客](https://blog.csdn.net/qq_19339041/article/details/81229962)
6. 用scrapy编写通用化爬虫难道只能是幻想？
7. 如何让每个spider自定义爬虫环境配置？答：1.将自定义爬虫环境配置从run.py（全局自定义）转移到spider（局部自定义）中（废弃，局部自定义环境配置要设置在spider的类变量custom\_settings中，而类变量要在类实例化前创建，与通用化相违背（类变量硬编写在spider中，而我需要动态改变自定义环境变量的值））2.创建多个爬虫进程，每个进程的使用不一样的自定义settings

注意

1. Feed：换页，走纸
2. Yield可以返回值且不中断函数运行
3. Spider接收到engine发来的response后解析其中内容，会获得爬取结果item或要继续爬取的页面的request，然后会将这些item与request发送给engine，engine分别转发到itempipeline或scheduler，request会等待scheduler的下一次调度
4. range()是一个可迭代对象，不是列表，不可打印，一般用来生成for in 迭代器的序列数
5. list()函数是对象迭代器，可以将range()返回的可迭代对象转换成一个列表
6. 迭代：指不停重复一个行为或过程，上一个迭代的结果是下一个迭代的初始值
7. 深拷贝，浅拷贝：[Python 直接赋值、浅拷贝和深度拷贝解析 | 菜鸟教程 (runoob.com)](https://www.runoob.com/w3cnote/python-understanding-dict-copy-shallow-or-deep.html)
8. Idea ctrl F4 关闭当前页面
9. Dict()：用于创建一个字典，可接受不定长键值对数组\*\*kwarg，映射mapping与iterable可迭代对象
10. Python 类属性与实例属性：类属性：在类实例化之前就定义好的属性，与java中的属性概念一致，形式：在类中，方法外，a = b。实例属性：在类实例化时或执行方法时动态创建的属性，形式：在“构造函数”或方法中，self.a = b
11. 字符串比较：str1 == str2，str1 is str2为对比两个字符串的值，当字符串相等时，其id值也相等，is用来判断id值是否相等
12. Var = None, del var 释放变量内存

3日记录

工作

1. 多爬虫同时运行：创建多个爬虫进程，每个进程的使用不一样的自定义settings，然后同步执行
2. 找到开源markdown编辑器marktext，试用
3. 开始编写后台界面，使用nodejs（npm包管理）驱动vite构建vue

问题

1. 对以下形式的新闻网站完成通用化



1. 尚未完成代理接入等反扒措施（除伪装浏览器），获取滚动界面的新闻等

注意

1. 列表合并：listC = listA + listB[python中如何合并两个列表-Python教程-PHP中文网](https://www.php.cn/python-tutorials-422181.html)
2. 开发环境dev，测试环境test，生产环境prod [(11条消息) 项目开发中的dev, test, prod , staging 环境是什么意思\_橘子y的博客-CSDN博客\_staging环境](https://blog.csdn.net/yupyuping/article/details/112131236?spm=1001.2101.3001.6661.1&utm_medium=distribute.pc_relevant_t0.none-task-blog-2%7Edefault%7ECTRLIST%7Edefault-1-112131236-blog-86700038.pc_relevant_multi_platform_featuressortv2dupreplace&depth_1-utm_source=distribute.pc_relevant_t0.none-task-blog-2%7Edefault%7ECTRLIST%7Edefault-1-112131236-blog-86700038.pc_relevant_multi_platform_featuressortv2dupreplace&utm_relevant_index=1)
3. json 文件：{} 是字典（键值对），[] 是列表

4日记录

工作

1. 编写后台管理界面

问题

注意

Json 配置文件各键描述

rules：爬取规则表，第一个linkextractor定义详细新闻页链接的规则，第二个linkextractor定义翻页链接的规则，最后都要用Rule类包装

linkextractor：链接提取器，