

## 附：部分工作笔记展示

### 爬虫工作笔记目录

<b>urllib 模块的使用</b>	<b>26</b>
request.urlretrieve 函数的用法	26
urllib.parse 模块的使用	27
urlencode, unquote, parse_qs	27
urlparse 和 urlsplit	28
urllib.request 模块的使用	30
urllib.request.urlopen 发送请求	30
查看 Python3 中默认的 User-Agent	32
使用 urllib.request.Request 发送带有 User-Agent 的请求	32
添加更多的 Header 信息	36
拉勾网职位爬虫	37
urllib.request 使用代理	37
使用 urllib.request 发送携带 cookies 的请求	40
urllib.request 模块处理证书的错误	46
<b>requests 模块使用</b>	<b>46</b>
查看 requests 的源码	46
response 常用方法	48
查看 request 的所有属性和方法	51
response.text 和 response.content 的区别	53
url 地址的编码和解码	55
requests 发送带 headers 参数的请求	55
基本使用方法	55
随机选择请求头	55
fake_useragent 的使用	56
requests 发送带查询参数的 get 请求	58
requests 发送 post 请求	59
案例: 有道翻译	60
案例: 百度翻译	62
requests 代理地址的使用	65
使用简单代理地址	66
SOCKS 代理	66
使用 fiddler 代理	67
随机切换代理	67

定义代理开关	68
使用 requests 保存图片到本地	69
彻底弄清楚 response.cookies	70
cookie 的格式	70
cookeis 详解	71
requests.session 模拟登录	76
cookie 和 session 区别:	76
requests 处理 cookies 及 session 请求	77
session 模拟登录豆瓣	77
使用 session 模拟登录人人网	78
requests session 模拟登录总结	83
处理不信任的 SSL 证书	83
request 的异常和利用	84
GET 与 POST 功能整合	86
自动切换 ip 代理与 User-Agent	89
使用 logging 模块记录程序日志	89
使用 logging 提供的模块级别的函数记录日志	89
使用 logging 四大组件记录日志	91
logging 模块的四大组件	91
格式化输出日志	92
记录异常信息	92
requests 与 logging 模块的配合使用	93
<b>2. 第二章 非结构化数据与结构化数据提取</b>	<b>95</b>
页面解析和数据提取	95
非结构化的数据处理	95
结构化的数据处理	95
xpath 语法与 lxml 库	96
什么是 XPath?	96
认识 XML	96
XML 的节点关系	96
XML 的节点关系	97
XPath 开发工具	97
XPath 语法	97
lxml 库	102
使用 lxml 库对 html 文本及 html 文件的解析	103
使用 xpath 提取网页中的数据	105
推荐使用分组的方法对子节点内容进行提取	106
使用 lxml 解析 HTML 代码总结	108
lxml 结合 xpath 注意事项	108
lxml 和 xpath 的使用总结	109
lxml 案例	114

总结: 实现爬虫的套路	114
后续爬虫代码的建议	116
<b>BeautifulSoup4 解析库</b>	<b>116</b>
几大解析工具对比	116
使用 BeautifulSoup4 对网页进行解析	117
beautifulsoup 解析器	117
查看 bs 对象的方法和属性	118
BS 中四个常用的对象	121
BS 中四个对象之间的关系	127
总结: 四种对象的关系	129
遍历文档树	129
搜索文档树	133
总结: 获取标签属性的方法	143
bs4 总结	145
bs4 案例	160
<b>正则表达式和 re 模块</b>	<b>160</b>
什么是正则表达式	160
正则表达式常用匹配规则	160
匹配多个字符	166
其它特殊字符	167
数量词的贪婪模式和非贪婪模式	169
Python 的 re 模块	170
分组 group	188
转义字符和原生字符串	189
小案例	194
案例 匹配 0-100 之间的数字	195
使用正则表达式来匹配出生日期	196
正则表达式 古诗文爬虫	196
正则表达式 qiushibaike 爬虫	197
<b>3. 第三章: 数据存储</b>	<b>198</b>
<b>3.1. json 文件处理</b>	<b>198</b>
什么是 json	198
JSON 支持数据格式	199
python 中的 json 模块	200
json 使用注意点	205
在命令行中使用 python -m json.tool 对 json 进行美化输出	206
JsonPath	209
<b>3.2. csv 文件处理</b>	<b>211</b>
读取 csv 文件	211
写入数据到 csv 文件	215
<b>3.3. excel 文件处理</b>	<b>217</b>
<b>3.4. MySQL 数据库</b>	<b>217</b>
安装 mysql	217

安装驱动程序	217
python 操作 mysql 数据库	218
<b>3.5. MongoDB 数据库</b>	<b>225</b>
Windows 下安装 MongoDB 数据库	225
运行 MongoDB	225
连接 MongoDB	225
使用 Compass 软件连接 MongoDB	225
将 MongoDB 制作成 windows 服务	226
MongoDB 概念介绍	226
MongoDB 三元素	226
MongoDB 基本操作命令	227
Python 操作 MongoDB	227
<b>4. 第四章: 爬虫进阶</b>	<b>231</b>
<b>4.1. 多线程爬虫</b>	<b>231</b>
多线程介绍	231
threading 模块介绍	231
Lock 版本生产者和消费者模式	236
Condition 版的生产者与消费者模式	239
Queue 线程安全队列	242
多线程示意图	244
使用生产者与消费者模式多线程下载表情包	246
多线程下载百思不得姐段子	252
多线程下载汽车之家图片	253
<b>4.2. 动态网页爬虫</b>	<b>256</b>
<b>动态 HTML 技术了解</b>	<b>256</b>
什么是 AJAX	256
获取 ajax 数据的方式	257
<b>Selenium+chromedriver 获取动态数据</b>	<b>259</b>
selenium introduction	259
Drivers	260
安装 Selenium 和 chromedriver	261
selenium 基本使用方法	261
webdriver 的常用操作	262
实例化 driver 的常用操作	271
WebElement 常用的操作	304
selenium 使用的注意点	319
selenium 进阶使用	320
selenium 爬虫案例	321
<b>Selenium 总结</b>	<b>325</b>
<b>代理地址解决方案</b>	<b>326</b>
<b>代理方案一: 使用付费的独享代理</b>	<b>326</b>

代理方案二：自建 adsl 拨号服务器	327
自建服务器代理 ip 池架构原理	327
常用技术选型	327
服务器搭建	328
代理方案三：使用免费代理构建代理池	330
代理方案四：使用付费代理构建代理池	334
验证码识别	334
使用阿里云市场识别验证码	335
使用云打码平台处理验证码	335
识别知乎验证码 requests	336
识别豆瓣验证码 selenium	337
验证码的识别总结	337

## scrapy-redis+docker 实现分布式爬虫目录

伯乐在线全站爬虫	3
新建项目和爬虫	3
设置 pycharm 调试 scrapy 爬虫项目	4
修改 settings.py 进行设置	5
添加断点，进行调试，查看 response 响应的内容	5
使用 scrapy 提取网页的信息	6
使用 xpath 进行信息的提取	6
使用 css 选择器提取网页中的元素	8
使用 extract_first()代替 extract()	10
完善 jobbole.py，从帖子详情页中提取信息	11
调试工具 Stepping toolbar 的使用	12
从文章列表页爬取所有文章	13
提取下一页的链接，构造并发送请求	15
使用 item 类	17
列表页图片 url 的获取	21
下载文章的封面图片	23
自定义 pipeline 获取下载的图片保存的路径	25
查看 images.ImagesPipeline 的源码	25
修改 pipelines.py，自定义图片下载管道	28
debug 获取 results 的结构	28

修改 pipelines.py, 从 results 中取出图片保存的路径 .....	30
<b>对 article_url 进行 md5 处理 .....</b>	<b>32</b>
<b>保存 item 数据到 json 文件中 .....</b>	<b>33</b>
使用自定义方法将数据保存到 json 文件中 .....	33
使用 scrapy 的 JsonItemExporter 保存 json 文件 .....	35
<b>通过 pipeline 保存数据到 mysql .....</b>	<b>38</b>
安装 python 的 mysql 驱动程序 .....	38
建立数据库和数据表 .....	40
修改 jobbole.py, 把发表时间修改为日期格式 .....	42
将数据保存到 mysql 中 .....	46
同步化将 Item 保存入数据库 .....	46
异步化将 Item 保存入数据库 .....	47
把 django 的 ORM 集成到 scrapy 中 .....	49
<b>scrapy itemloader 机制 .....</b>	<b>50</b>
ItemLoader 的使用方法 .....	50
修改 jobbole_it.py, 使用 itemloader .....	52
修改 items.py, 对 itemloader 提取到的字段进行处理 .....	54
input_processor 的使用 .....	54
处理 create_date 字段 .....	59
TakeFirst()的使用 .....	61
自定义 itemloader 对 item 中的信息进行处理 .....	63
修改 item.py. 自定义 ItemLoader .....	63
修改 jobbole_it.py, 导入并使用自定义的 JobboleArticleItemLoader .....	64
修改 items.py, 改进 input_processor .....	65
最终完成的代码 .....	72
<b>通过 downloadmiddleware 随机更换 user-agent .....</b>	<b>83</b>
实现 user-agent 的自动更换的方法 .....	83
查看 scrapy 自带的 UserAgentMiddleware 的源码 .....	83
启用自定义的 downloader middleware .....	84
使用 Fake Useragent .....	85
Installation .....	85
Usage .....	85
Notes .....	86
Experiencing issues??? .....	87
使用 fake useragent 来随机切换 user-agent .....	88
基本使用方法 .....	88
使用本地的 fake_useragent.json 文件 .....	88
<b>scrapy 实现 ip 代理池 .....</b>	<b>90</b>
爬虫代理哪家强? 十大付费代理详细对比评测出炉! .....	90
使用 2 个代理获取代理地址 .....	91
修改 middlewares.py, 同时切换代理和 ua .....	93
在 settings.py 中设置启动 middleware .....	96
解决图片下载出错引起的频繁更换代理的问题 .....	99
解决 mysql 4 字节 utf-8 字符的问题 .....	99

<b>把项目部署到 ubuntu server 中 .....</b>	<b>101</b>
环境配置 .....	101
阿里云 ubuntu 基本配置 .....	101
安装 mysql-server 并配置 .....	101
安装 redis 并配置 .....	103
安装 python3.6 并创建虚拟环境 .....	107
把项目部署到 ubuntu server 中 .....	109
pip install mysqlclient: mysql_config not found 错误 .....	110
错误 2: Could not run curl-config: [Errno 2] 'curl-config': 'curl-config' .....	111
Failed building wheel for mysqlclient .....	112
错误 3: Failed to build pycurl .....	113
mysql_config not found .....	113
运行爬虫 .....	114
<b>修改为分布式爬虫 .....</b>	<b>114</b>
<b>使用 docker 部署爬虫 .....</b>	<b>116</b>
安装 Docker .....	116
Get Docker CE for Ubuntu .....	116
添加到用户组（可选项） .....	119
使用阿里云 docker 加速器 .....	121
Docker 的使用 .....	123
在 aliyun ubuntu 中创建容器并进行配置 .....	123
docker 基本操作 .....	123
安装 mysql-server .....	125
安装 redis 并配置 .....	128
安装 python3.6 并创建虚拟环境 .....	133
配置物理机中的 mysql 和 redis-server .....	136
将容器封装为 docker 镜像 .....	139
把保存的 tar 镜像文件复制到 Yunzhuji 的 ubuntu 物理机中 .....	140
在 yunzhuji 的容器中运行爬虫 .....	142
scrapy 日志处理 .....	144
分布式爬虫爬取结束时自动结束爬虫 .....	144

## R 语言基础笔记目录

<b><i>R 语言入门基础</i> .....</b>	<b>2</b>
1. 统计软件比较 .....	2
2. R 语言入门和安装 .....	3
3. R 的基本数据类型 .....	5
4. R 的数据结构 .....	8
5. 外部数据导入 .....	27
6. 程序控制 .....	31

<i>R</i> 进行描述性分析	36
1. 变量统计特征探索	36
2. 变量的可视化探索	49
<i>R</i> 进行推断性分析	61
1. 参数估计	61
2. 假设检验	66
3. 两个样本检验	72
4. 方差分析	79
<i>R</i> 结语	86

## 数据抽样与预处理笔记目录

市场调研与数据预处理技术	2
课程章节标题	2
一. 数据采集	2
1.1 数据的间接来源	2
1.2 数据的直接来源	3
二. 数据抽样	6
抽样调查分类	6
抽样调查中的误差来源	7
样本量的确定	8
软件实现	8
三. 数据预处理	14
3.1 数据错误	14
3.2 缺失值	28
3.3 离群值	41
四. 变量关系探索	49
4.1 连续变量相关系数筛选法	50



4.2 分类变量之间相关系数	55
4.3 连续变量与分类变量之间	61
五. 数据变换	64
5.1 标准化	64
5.2 Tukey's Ladder of Powers	65
5.3 Box - Cox transformation	66
六.数据降维	72
背景	72
主成分分析的思路	72
三维变量之间的关系	73
主成分分析公式化表述	75
基于相关系数矩阵的主成分分析	77