SGG爬虫设计方案

1. 项目需求

1.1 项目描述

1.1.1 数据源

来源	URL	备注
寄托社区offer榜	http://www.gter.net/offer/index.html	需要分析URL规律
微博留学大V	https://m.weibo.cn	崔钟博汶Cook; offer播报酱; 英国offer鸭
一亩三分地	...	...

1.1.2 字段

(1) 寄托社区

字段	备注
申请学校	已规范化
学位	已规范化
专业	已规范化
申请结果	已规范化
入学年份	已规范化
入学学期	已规范化
通知时间	已规范化
本科学校档次	匿去具体学校，保留规范化档次
本科专业	无规范，格式用户自定
本科成绩和算法、排名	无规范，用户自定

研究生专业	无规范，用户自定
研究生成绩和算法、排名	无规范，用户自定
研究生学校档次	无规范，用户自定
其他说明	无规范，用户自定

(2) 微博留学大V

字段	备注
微博ID	每条微博的标识符
发布时间	需要清洗，表示方法不统一
微博内容	文字信息

(3) 一亩三分地

字段	备注
申请学校	已规范化
学位	已规范化
专业	已规范化
申请结果	已规范化
入学年份	已规范化
入学学期	已规范化
通知时间	已规范化
本科学校档次	匿去具体学校，保留规范化档次
本科专业	无规范，格式用户自定
本科成绩和算法、排名	无规范，用户自定
研究生专业	无规范，用户自定
研究生成绩和算法、排名	无规范，用户自定
研究生学校档次	无规范，用户自定
其他说明	无规范，用户自定

1.2 网页分析

1.2.1 寄托社区

(1) URL分析

> 可绕开验证登录，分析每条offer播报的`URL`，直接获取相关信息
> offer播报`URL`模式：

- http://bbs.gter.net/offer_{code}.html

> `code`从4开始，到40020止（截至2020/3/25）

(2) 内容分析

> 所需获取的信息在`<div class="typeoption">`片段中
> 信息以`<table>`形式组织
> offer信息所在的`<table>`中`summary`属性为"`offer x`"（x为某个数字）
> 申请人信息所在的`<table>`中`summary`属性为"`个人情况`"

```
<table summary="offer 2" cellpadding="0" cellspacing="0" class="cgtl
mbm">
        <caption>offer </caption>
        <tbody>
            <tr>
                <th>申请学校:</th>
                <td>

                    <a
href="http://school.gter.net/index/show/id/125.html"
target="_blank">University of Guelph</a>

                </td>
            </tr>
            <tr>
                <th>学位:</th>
                <td>MS</td>
            </tr>
            <tr>
                <th>专业:</th>
                <td>

                    <a
href="http://school.gter.net/professional/lists/id/472.html"
target="_blank">Electrical and Electronics Engineering</a>

                </td>
            </tr>
            <tr>
                <th>申请结果:</th>
                <td>AD小奖</td>
            </tr>
            <tr>
                <th>入学年份:</th>
```

```
                <td>2014</td>
            </tr>
            <tr>
                <th>入学学期:</th>
                <td>Spring</td>
            </tr>
            <tr>
                <th>通知时间:</th>
                <td>2013-11-26</td>
            </tr>
        </tbody>
    </table>
```

```
<table summary="个人情况" cellpadding="0" cellspacing="0" class="cgtl
mbm">
        <caption>个人情况</caption>
        <tbody>
            <tr>
                <th>TOEFL:</th>
                <td>
                    Overall: 106,
                    R: 29 /
                    L: 29 /
                    S: 24 /
                    W: 24
                </td>
            </tr>
            <tr>
                <th>GRE:</th>
                <td>
                    Overall: 316,
                    V: 150 /
                    Q: 166 /
                    AW: 3
                </td>
            </tr>
            <tr><th>本科学校档次:</th><td>中山大学</td></tr>
            <tr><th>本科专业:</th><td>社会学  </td></tr>
            <tr><th>本科成绩和算法、排名:</th><td>3.8/4.0  </td></tr>
        </tbody>
    </table>
```

1.2.2 微博留学大V

(1) URL分析

> 直接对网页版微博进行爬取难度大、稳定性差，因此选择移动版
> 分析发现移动版存在微博内容接口，其中`uid`为大V用户标识码，`page`为页数

- https://m.weibo.cn/api/container/getIndex?uid={uid}&type=uid&page=

{page}&containerid=107603{uid}

> 微博内容接口所提供的微博文本可能需要"展开全文"
> 展开全文接口如下，其中 `wid` 为微博标识符，该URL的访问需要登录验证

- https://m.weibo.cn/statuses/extend?id={wid}

> 所需爬取的微博大V的 `uid` 如下

- 崔钟博汶Cook：1824301624
- offer播报酱：7330842801
- 英国offer鸭：7029497065

(2) 内容分析

> 微博接口返回的数据为 `json` 类型
> 需要提取 `mblog` 部分
> `mblog` 主要包含如下几个需要提取的字段

- `idstr`：该条微博的ID
- `created_at`：该条微博创建时间
- `text`：微博文字内容

> 展开全文接口部分，只需要关注 `text` 字段

1.2.3 一亩三分地

(1) URL分析

> 免费信息不需要登陆，从录取汇报板块 `URL` 获取相关信息。
> 每一页 `URL` 模式：
>
> - http://www.1point3acres.com/bbs/forum-82-{code}.html
>   `code` 从1开始，到1000止（截至2020/3/25），即共1000页。

(2) 内容分析

> 每一个页面包含若干个id为"normalthread_xxxxxx"的 `<tbody>`，每个片段包含一个录取案例信息。

```
<tbody id="normalthread_615904">
<tr>
    <td class="icn">
        <a href="forum.php?
mod=viewthread&amp;tid=615904&amp;extra=page%3D1%26filter%3Dsortid%26so
rtid%3D164%26sortid%3D164" title="新窗口打开" target="_blank">
            <img src="static/image/common/folder_common.gif">
        </a>
    </td>
```

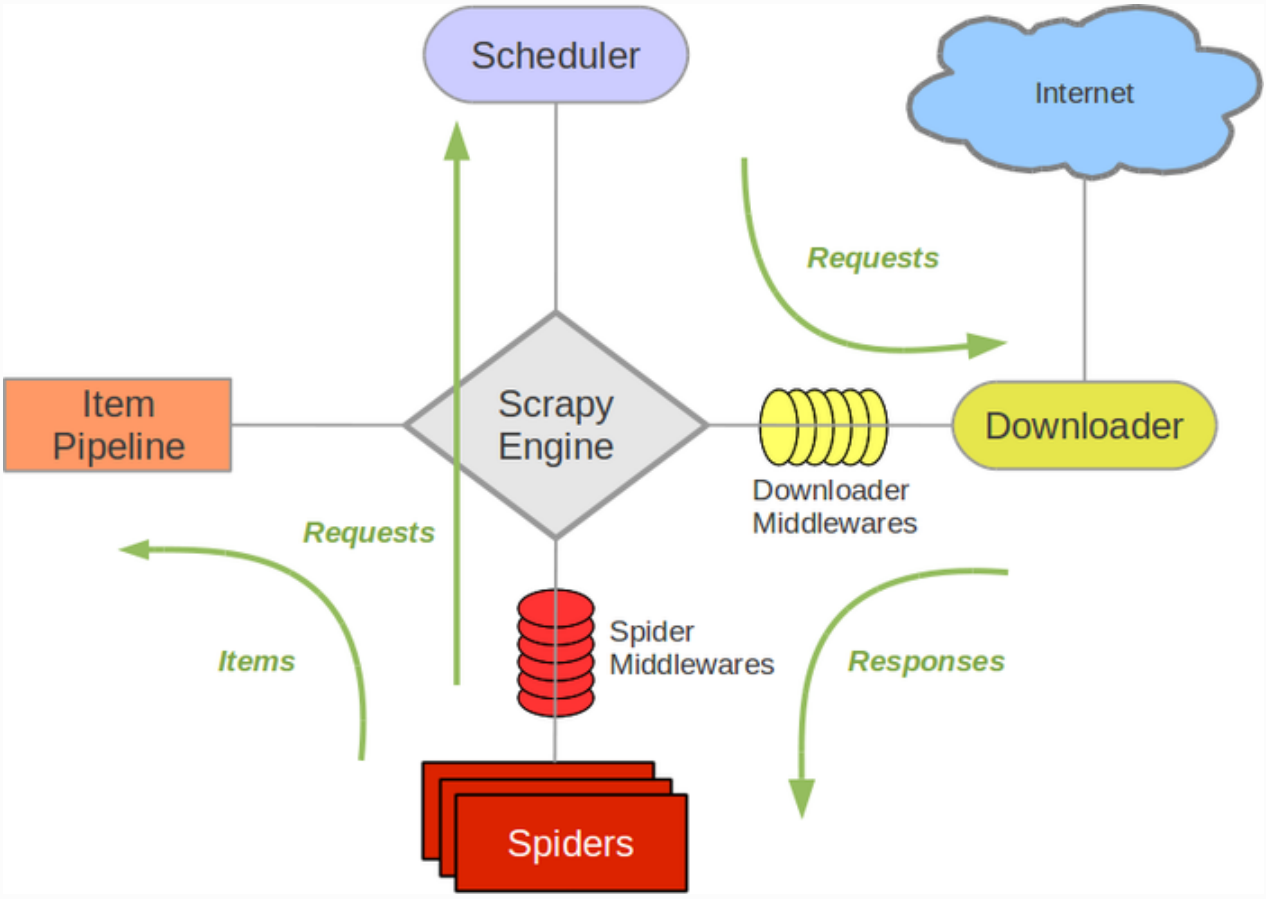
```
    <th class="common">
        <em>
            [<a href="forum.php?
mod=forumdisplay&amp;fid=82&amp;filter=sortid&amp;sortid=164">结果汇报
</a>]
        </em>
        <a href="forum.php?
mod=viewthread&amp;tid=615904&amp;extra=page%3D1%26filter%3Dsortid%26so
rtid%3D164%26sortid%3D164" onclick="atarget(this)" class="s xst">UFL cs
正式录取</a>
        <br>
        <span style="margin-top: 3px">
            <u>
                <font color="#666">[20Fall</font>.
                <font color="blue">MS</font>.
                <font color="black">
                    <b>AD无奖</b>
                </font>
                [<font color="#F60">
                    <b>CS</b>
                </font>
                <font color="#00B2E8">UFL</font>
            </u>
            <font color="brown">2020-03-25</font>
            <font color="green"></font>
            <font color="purple">本科其他985/211</font>
            <font color="hotpink"></font>
            <font color="brown"></font>
        </span>
        <span class="tps"> ...<a href="forum.php?
mod=viewthread&amp;tid=615904&amp;extra=page%3D1%26filter%3Dsortid%26so
rtid%3D164%26sortid%3D164&amp;page=2">2</a></span>
    </th>
    <td class="by">
        <cite>
            <a href="space-uid-595158.html" c="1"
mid="card_690">Passion.</a></cite>
        <em><span><span title="2020-3-25">昨天 23:58</span></span>
</em>
    </td>
    <td class="num"><a href="forum.php?
mod=viewthread&amp;tid=615904&amp;extra=page%3D1%26filter%3Dsortid%26so
rtid%3D164%26sortid%3D164" class="xi2">13</a><em>778</em></td>
    <td class="by">
        <cite><a href="space-username-%F4%94%C9%DF%B5%DB%CD%F5.html"
c="1" mid="card_4377">鲁蛇帝王</a></cite>
        <em><a href="forum.php?
mod=redirect&amp;tid=615904&amp;goto=lastpost#lastpost"><span
title="2020-3-26 19:54">26 分钟前</span></a></em>
      </td>
  </tr>
</tbody>
```

1.3 运行环境要求

- 操作系统：`Windows/Linux/Mac`
- 编程语言：`python 3.5` 以上版本
- 爬虫框架：`Scrapy`
- 数据解析库：`bs4` ；`lxml`

2. 爬虫项目结构

2.1 总体结构



2.2 简要说明

2.2.1 下载器中间件

> 下载器中间件是介于 `Scrapy` 的 `request/response` 处理的钩子框架。 是用于全局修改 `Scrapy request` 和 `response` 的一个轻量、底层的系统
> 对于微博数据源，构造 `request` 时需要附带 `cookie`
> 对于寄托社区与一亩三分地，构造 `request` 时需要附带 `UA`
> 本期项目需要构建一个 `UA` 池和一个 `cookie` 池

2.2.2 Spiders

> `Spider` 是 `Scrapy` 用户编写用于分析 `response` 并提取 `item` (即获取到的 `item`)或额外跟进的URL的类。 每个 `spider` 负责处理一个特定(或一些)网站
> 本期项目包含三个 `spider` ，分别对寄托社区、一亩三分地、微博进行爬取

2.2.3 Item

> `Item` 用于存储数据解析得到的信息，传递至 `Pipeline` 处理
> 本期项目包含三个 `item` ，分别为 `offers` 、 `applicants` ， `weibo`
> `offer` 存储offer和拒信的相关数据
> `applicants` 存储申请人相关数据
> `weibo` 存储微博相关数据

2.2.4 Pipeline

> `Pipeline` 主要用于数据清洗、检查item包含某些字段、查重、存储数据
> 本期项目 `Pipeline` 主要任务为:

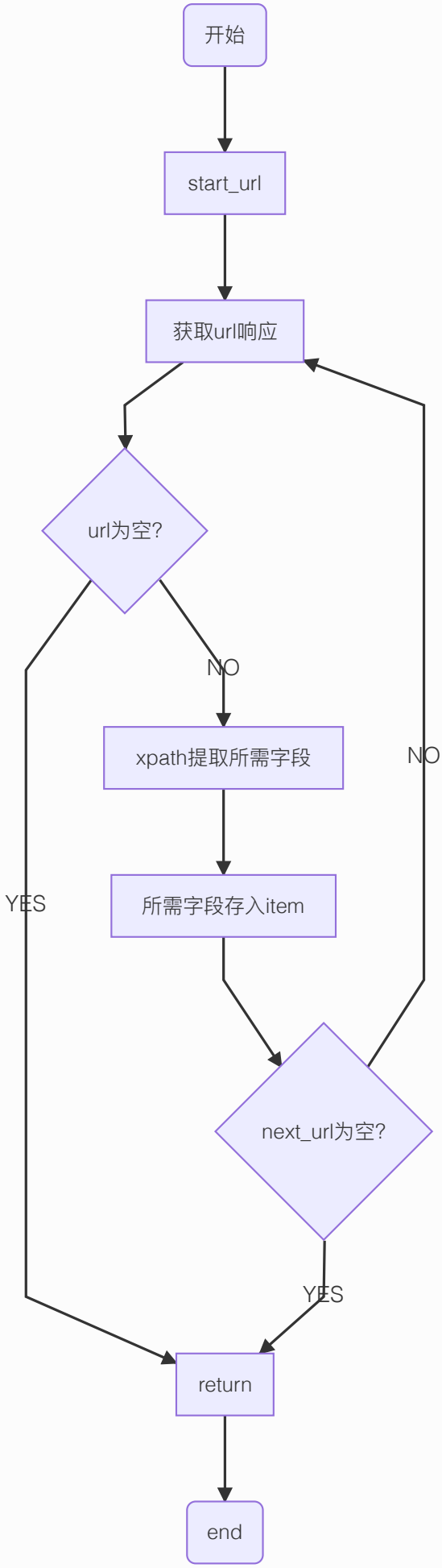
- 将微博 `created_at` 转化为日期（不含时间）
- 对微博 `text` 、一亩三分地与寄托社区文本数据进行文本清洗与分词操作
- 将数据存入csv文件中

3. 爬虫逻辑

3.1 寄托爬虫

```
step 1  获取待解析的html文本
step 2  提取<table summary="offer x">与<table summary="个人情况">的tag
step 3  if 所提取tag的个数小于等于1:
step 4      结束
step 5  else:
step 6      获取需要的字段，存入item
step 7  end if
```

3.2 一亩三分地爬虫


开始
start_url
获取url响应
url为空?
NO
xpath提取所需字段
所需字段存入item
next_url为空?
NO
YES
YES
return
end

3.3 微博爬虫

```
step 1   获取Json
step 2   提取mblog字段数据
step 3   if 不存在mblog字段:
step 4       结束
step 5   else:
step 6       将除了text字段的其他数据存入item
step 7       if ">全文<" 存在与text字段中:
step 9           访问全文接口获取text字段并存入item
step 10      else:
step 11          将text字段存入item中
step 12      end if
step 13  end if
```