面向知乎 QAWEB 的网络爬虫设计与实现

游浩然
黎张帆
郭金城

u201515429　u201514574　u201511174　@hust.edu.cn

January 2, 2018

Table of Contents

项目描述

- 爬虫控制端：启动爬虫，停止爬虫，监视爬虫的运行情况
- 爬虫运行模块：包含三个小模块，URL 管理器、网页下载器、网页解析器
  - URL 管理器：对需要爬取的 URL 和已经爬取过的 URL 进行管理，可以从其中取出待爬取的 URL 传递给网页下载器。
  - 网页下载器：网页下载器将 URL 指定的网页下载下来，存储成一个字符串，传递给网页解析器。
  - 网页解析器：网页解析器解析传递的字符串，解析器不仅可以解析出需要爬取的数据，而且还可以解析出每一个网页指向其他网页的 URL，这些 URL 被解析出来会补充进 URL 管理器。
- 数据输出模块：存储爬取的数据

开发平台

- 硬件平台：个人 PC，可连接 WEB 网络
- 软件环境：
  - 操作系统：WINDOWS 10 & Ubuntu
  - IDE：PyCharm Community Edition 2017.2.3
  - Python 版本：Python 3.6.3 |Anaconda, Inc.| (default, Oct 15 2017, 03:27:45)
- 项目所需库
  - Wxpthon：Python GUI 库
  - Selenium：Python 模拟浏览器操作库
  - Phantom：基于 selenium 的 Python 工业化模拟浏览器操作库，需设置运行路径或将 bin 文件夹放入系统路径中。

任务分工

姓名	负责模块	主要任务
黎张帆	爬虫控制器	启动、停止、监视爬虫的运行情况；实现多线程爬取。GUI 设计
游浩然	爬虫运行模块	知乎登陆；URL 管理器、网页下载器、网页解析器；动态加载
郭金城	数据输出模块	通过网页标签抓取问题 URL 下的相关信息；动态加载

Table of Contents

框架

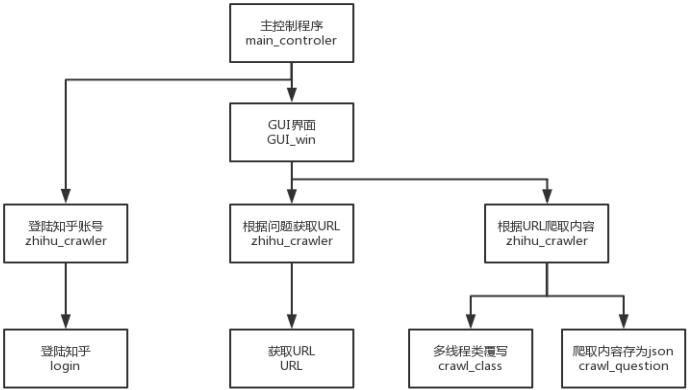

Figure 1: 整体框架设计

GUI 使用说明

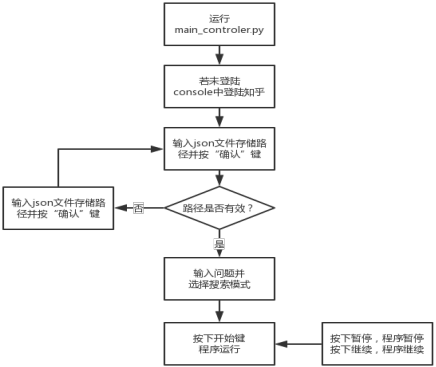

Figure 2: GUI 使用说明

Figure 3: GUI 界面

GUI 设计

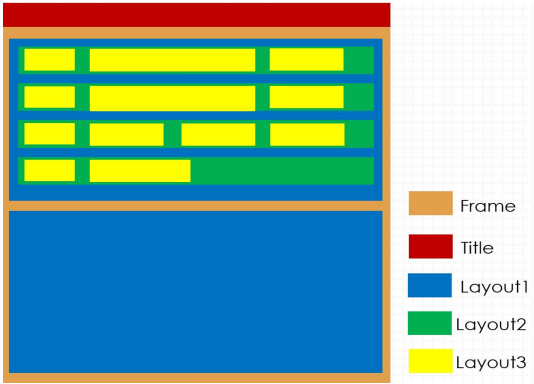

Figure 4: GUI 设计框图

Figure 5: GUI 界面

主程序流程

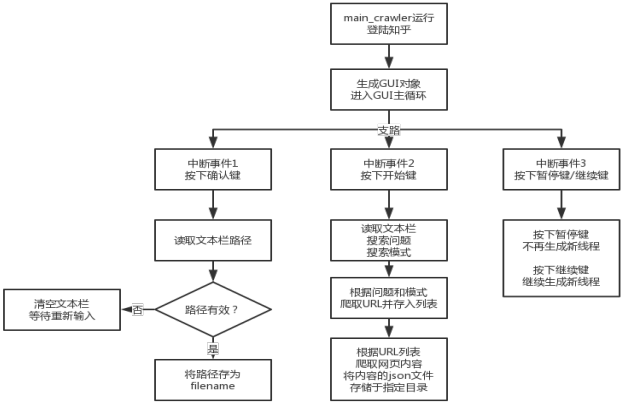

Figure 6: 主程序流程图

多线程控制

Thread 子类 Crawler

Class Crawler
Parent: threading.Thread

Method:
__init__(self, funcName, args=)#覆写父类方法
run(self)#覆写父类方法
_run(self)

Figure 7: Crawl

多线程控制

```
def run(self):
    try:
        self._run()
    except Exception as e:
        self.exitcode = 1
        self.exception = e
        self.exc_traceback = ''.join(traceback.format_exception(*sys.exc_info()))
        print("URL Crawler Exception: ", self.args[0])
        print(" self.exc_traceback")

def _run(self):
    try:
        self.funcName(self.args[0],self.args[1])
    except Exception as e:
        raise e
```

Figure 8: 覆写 run 方法

爬虫运行模块

本模块主要实现了爬虫的基本运行，分为两个主要部分：

- 知乎登陆
- 建立问题 URL 库
  - 针对具体问题（Question）搜索知乎的“综合”模块
  - 针对具体话题（Topic）搜索知乎的“话题”模块

其中对问题（Question）的爬取采用 PhantomJS 库来实现动态加载，对话题（Topic）的爬取采用模拟翻页来实现动态加载。

爬虫运行框图

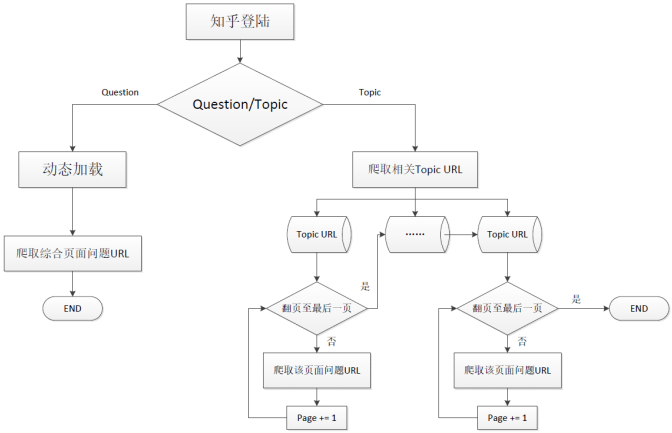

Figure 9: 爬虫运行框图

知乎登陆

使用手机的代理（User-Agent）来模拟登陆

```
1 headers = {
2     "Host": "www.zhihu.com",
3     "Referer" : url,
4     'X-Requested-With' : 'XMLHttpRequest',
5     'User-Agent' : 'Mozilla/5.0 (Linux; Android 6.0; Nexus 5 Build/MRA58N)
6                     AppleWebKit/537.36 (KHTML, like Gecko) Chrome/56.0.2924.87 Mobile Safari/537.36'
7 }
```

Figure 10: headers 头文件信息

知乎登陆

采用验证码方式登陆，测试结果如下：

```
1 C:\Users\dell-pc\Anaconda3\python.exe C:/Users/dell-pc/Desktop/zhihu_crawler_py3/login.py
2 Cookie cannot be load!
3 your name: None
4 please input your account:******
5 please input your secret:******
6 please input the captcha:f8dk
7 the status code returned by server: 200
8 {'r': 0, 'msg': '登录成功'}
9 your name: <span class="name">游浩然</span>
```

Figure 11: login

网页下载和解析

网页下载函数即为下代码所示，使用 requests 库来抓取指定 URL 的内容：

```
1 def getHTMLText(url):
2     try:
3         r = requests.get(url, headers=headers)
4         # r.status_code  200
5         # r.encoding  'utf-8'
6         return r.content
7     except:
8         return "Error"
```

Figure 12: 网页下载

网页下载和解析

网页解析函数即为下代码所示，使用 BeautifulSoup 库来抓取解析网页：

```
1 demo = demo_temp.pop()
2 soup = BeautifulSoup(demo, 'html.parser')
3 for meta in soup.find_all('meta', attrs={'itemprop': 'url'}):
4     url_temp = meta.get('content')
5     url_list.append(url_temp)
```

Figure 13: 网页解析

URL 爬取和管理

- 对知乎“综合”模块进行爬取
- 对知乎“话题”模块进行爬取

存储为 urllist[] 列表格式供数据爬取模块调用

Question 页面动态加载

- 问题的完整描述需要执行点击操作，使用 selenium 模拟
- 通过鼠标的下滑不断地动态加载，使用 selenium 模拟

Question 页面动态加载

```
# 点击操作
try:
    click_btn = driver.find_element_by_xpath('//button[@class="Button QuestionRichText-more Button--plain"]')
    ActionChains(driver).click(click_btn).perform()
    time.sleep(0.5)
    html_text = driver.page_source
except:
    html_text = driver.page_source
# 下滑操作
while(True):
    driver.execute_script('window.scrollTo(0,document.body.scrollHeight)')
    time.sleep(0.8)
    if(html_text == driver.page_source):
        break
    html_text = driver.page_source
```

Figure 14: selenium 模拟点击与下滑

Question 数据存储

效果展示


```
{"question_text": "", "question_commet": "19 条评论", "question_id": "62968639", "question_followers": "427", "question_title": "文学是否是抵御 AI 的最后一个阵地？", "answer_number": "95 个回答", "br
{"answer_id": 252689132, "answer_comment": "7 条评论", "answer_author": "知乎用户", "answer_votes": "27", "answer_text": "文学不是，但文学批评（literacy criticism）估计是AI第一个取代的领域"}
{"answer_id": 205125138, "answer_comment": "3 条评论", "answer_author": "吕学平", "answer_votes": "6", "answer_text": "文学的意义在于“记录”和“表达”AI的记录功能不用怀疑，强悍的信息检索与汇总能力是远远在人
{"answer_id": 252710799, "answer_comment": "9 条评论", "answer_author": "Karl Ernst", "answer_votes": "27", "answer_text": "我知道AI搞文学创作肯定是有可能的。但是，我想说的不是这个问题，也不是这个“抵御”
{"answer_id": 252231747, "answer_comment": "66 条评论", "answer_author": "景略集智", "answer_votes": "89", "answer_text": "<img src=\"https://pic2.zhimg.com/50/v2-ca629e3baa8a45f800c24f7a65c39e76_
{"answer_id": 204802622, "answer_comment": "4 条评论", "answer_author": "Farhi Al", "answer_votes": "28", "answer_text": "AI的最后一个障碍只会是自我编译"}
{"answer_id": 204566871, "answer_comment": "83 条评论", "answer_author": "蔡骏", "answer_votes": "314", "answer_text": "谢邀。比起下围棋，写作者的生活变量太多，无法预料，曹雪芹早上多吃个馒头，今晚贾府就
{"answer_id": 252739979, "answer_comment": "8 条评论", "answer_author": "李怡", "answer_votes": "4", "answer_text": "我觉得单纯从文学水平上去聊AI文学毫无意义，就算AI将来能一字不差写出一个能得诺贝尔的文学
{"answer_id": 204789891, "answer_comment": "8 条评论", "answer_author": "Sirius", "answer_votes": "17", "answer_text": "结论：几乎不能创作纯粹的AI文学，不是机器的原因，是人的原因。先看看文学是怎么来的。
{"answer_id": 204659547, "answer_comment": "36 条评论", "answer_author": "五百击", "answer_votes": "101", "answer_text": "AI早就开始文学创作，而且还差点获奖了2016年“星新一文学奖”（以日本著名科幻小说家命
{"answer_id": 252725645, "answer_comment": "12 条评论", "answer_author": "理想的蘑菇", "answer_votes": "45", "answer_text": "别的我不知道现代诗肯定挡不住而且不是在未来就现在已经被干掉了甚至都没有使用ai"}
{"answer_id": 204796511, "answer_comment": "9 条评论", "answer_author": "尽黑丶", "answer_votes": "25", "answer_text": "在2014年翻一下知乎里围棋相关的回答，基本上没有认为AI可以达到职业顶尖水平的。而现在
{"answer_id": 257483906, "answer_comment": "添加评论", "answer_author": "王汪汪", "answer_votes": "0", "answer_text": "文学也分严肃文学和通俗文学，那些遍布网上的爽文甚至是畅销一时的通俗小说，有朝一日ai一
{"answer_id": 204678305, "answer_comment": "23 条评论", "answer_author": "三当家的", "answer_votes": "64", "answer_text": "阿尔法。先生，我在这里。我是不是要死了？我很遗憾，是的，先生。我死之前可不可以给
{"answer_id": 253198967, "answer_comment": "添加评论", "answer_author": "文艺范花样作死君", "answer_votes": "0", "answer_text": "ai可能永远搞不懂文学，但当ai成为主流以后，文学可能就没有存在的价值了"}
{"answer_id": 204934130, "answer_comment": "5 条评论", "answer_author": "谢丹", "answer_votes": "4", "answer_text": "不能。因为图灵判断。一个AI写出的文字，不需要超越人类才能生存，只要大家读起来意识不到差
{"answer_id": 204946842, "answer_comment": "12 条评论", "answer_author": "M3小蘑菇", "answer_votes": "9", "answer_text": "<img src=\"https://pic4.zhimg.com/50/v2-cf6d5054c234c428c75d39bb780cde2e_h
{"answer_id": 204805251, "answer_comment": "12 条评论", "answer_author": "王玄策", "answer_votes": "31", "answer_text": "&lt;img data-rawheight=\"222\" src=\"https://pic3.zhimg.com/50/v2-348b256cb
{"answer_id": 205080032, "answer_comment": "添加评论", "answer_author": "稻草人", "answer_votes": "1", "answer_text": "回答这个问题前，必须先回答另一个问题：意识的是一种功能吗？这里的功能指的是物质的某种
{"answer_id": 204645686, "answer_comment": "1 条评论", "answer_author": "华祝迎恩", "answer_votes": "7", "answer_text": "文学不只由读者组成。写作的愉悦也是文学的一部分。"}
{"answer_id": 204908494, "answer_comment": "3 条评论", "answer_author": "喆喆喆", "answer_votes": "4", "answer_text": "以现代人的阅读习惯，说不定这一块会比你想象的更快攻克。这里有一位著名作家回答了，我知
{"answer_id": 204625024, "answer_comment": "2 条评论", "answer_author": "Grotrupida", "answer_votes": "4", "answer_text": "微软小冰已经开始写诗了"}
{"answer_id": 205176195, "answer_comment": "添加评论", "answer_author": "imawesome", "answer_votes": "3", "answer_text": "https://arxiv.org/pdf/1705.08807.pdf这篇文章里作者调查了AI研究者对未来AI取代
{"answer_id": 204673094, "answer_comment": "1 条评论", "answer_author": "AABb", "answer_votes": "3", "answer_text": "还有村东头50一晚的红房子&lt;img src=\"https://pic1.zhimg.com/50/v2-fcf96b4a113da
{"answer_id": 204573387, "answer_comment": "添加评论", "answer_author": "果果大树小树", "answer_votes": "3", "answer_text": "不是还有音乐、绘画、雕塑等艺术另外，我见过有人说艺术要的就是人创造出来的东西，
{"answer_id": 204600313, "answer_comment": "添加评论", "answer_author": "匿名用户", "answer_votes": "4", "answer_text": "基础学科研究才是"}
{"answer_id": 205445454, "answer_comment": "1 条评论", "answer_author": "blank-shiki", "answer_votes": "2", "answer_text": "不好说啊。比如：脑洞一下……假如人类的思维与行为也是按照有逻辑性的算法形成，那AI
{"answer_id": 204701404, "answer_comment": "4 条评论", "answer_author": "匿名用户", "answer_votes": "2", "answer_text": "还记得《锦绣未央》（原名《庶女有毒》）是如何成书的吗？而那还不是AI，不过是区区软件
{"answer_id": 205012767, "answer_comment": "添加评论", "answer_author": "苏晗若", "answer_votes": "2", "answer_text": "以艺术文学家的智商，怕不是早就被AI爆了十条街。至于应用文学家(哲学家，科学家)，早就已
{"answer_id": 253874917, "answer_comment": "添加评论", "answer_author": "周剑", "answer_votes": "1", "answer_text": "你们平时看到的部分新闻稿已经是AI自动编写自动发表的了。平时有感觉到吗？国内外很多媒体都
{"answer_id": 252628580, "answer_comment": "添加评论", "answer_author": "匿名用户", "answer_votes": "1", "answer_text": "最后一个阵地。。。把基础科学研究放哪?"}
{"answer_id": 205591314, "answer_comment": "添加评论", "answer_author": "匿名用户", "answer_votes": "1", "answer_text": "人类写作的意义应该不只在于吸引别人来看，还可以自我愉悦，Al应该不能时常翻出自己从前
{"answer_id": 205168203, "answer_comment": "添加评论", "answer_author": "如是", "answer_votes": "1", "answer_text": "01.这问题把文学和人工智能联合在一起了，文学是经久不衰的，人工智能又是时代的宠儿，挺有
{"answer_id": 204848234, "answer_comment": "添加评论", "answer_author": "辣鸡翔", "answer_votes": "1", "answer_text": "微软小冰不是出了一部诗集吗&lt;img src=\"https://pic2.zhimg.com/50/v2-3232de7828
{"answer_id": 204640795, "answer_comment": "5 条评论", "answer_author": "EVANGELION", "answer_votes": "1", "answer_text": "文学拿什么来抵御……"}
```


Figure 15: json 效果展示

Question 数据存储


```
{
    "question_followers": "15,436",
    "question_title": "如何在四小时内学会用 Ai 做 UI？",
    "answer_number": "86 个回答",
    "question_text": "整个问题的初衷在于看到了@黎翱 的回答，产品经理新人是否有必要学习Photoshop？ 我这个问题主要在于希望一些新人能速度的进入AI操作，而不必花时间去找教程，需要这个问题回答的人不少。再次谢谢 @黎翱。",
    "question_comnet": "15 条评论",
    "brower_number": "989,872",
    "question_id": "21378038"
}
```


Figure 16: json 效果展示


```
{
    "answer_votes": "0",
    "answer_author": "mac mico",
    "answer_text": "工欲善其事必先利其器，效率一定是首要的。UI设计之效率为王 - 设计与开发之效率 - 知乎专栏",
    "answer_id": 61575458,
    "answer_comment": "添加评论"
}
```


Figure 17: json 效果展示

Table of Contents

- Thank You.