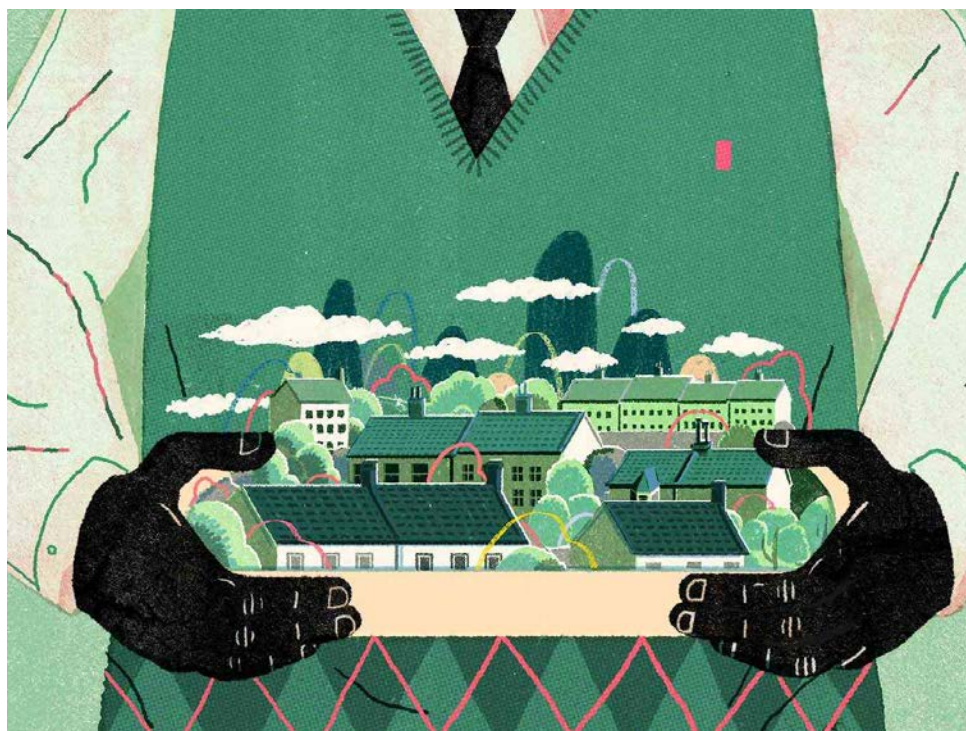


# 期末大作业 WhereToGo 爬虫说明文档

2017-03-31 指导老师 嵩天 作者 Bonfey



每到毕业季，学子们将纷纷离开难舍的校园去追求自己的理想。这时困扰很多同学的问题就出现了：如何选择最适合自己的生存和发展的城市？如何量化地比较几个候选城市以帮助自己做出最佳的选择？

可能每个同学心中都有那么几个理想的的城市作为目标，然而网上的数据和信息过于零散，碎片化地去查资料实在太低效。那么下面我利用刚刚从《Python 网络爬虫与信息提取》这门课中学到的知识设计一个能从多个网站定向爬取大量数据的爬虫，并对提取后的数据进行统计分析，帮助毕业班的同学们一键选择最宜居的城市。

年前北京的一轮重度雾霾刷爆了朋友圈，很多朋友表示开始认真考虑离开北京的问题了，更有知名互联网企业趁机把总部搬到了深圳，因此本设计把城市的气候作为首要参考条件。当然了，即便一年四季如春，每天都是花香鸟语，如果银包里没有 Money，还是开心不起来的，接下来就要参考跟自己专业相关的工作岗位有多少，薪资水平怎么样。最后一个重要的参考相信不说同学们也能猜到，对，就是住房，这里根据毕业生的需求着重爬取和分析了租房和二手房的数据。

本设计选取 Requests 库、Beautiful Soup 库和 re 库为主要技术路线，然后根据 Robots 协议确定了如下的数据网站：

天气网历史天气：<http://lishi.tianqi.com/>

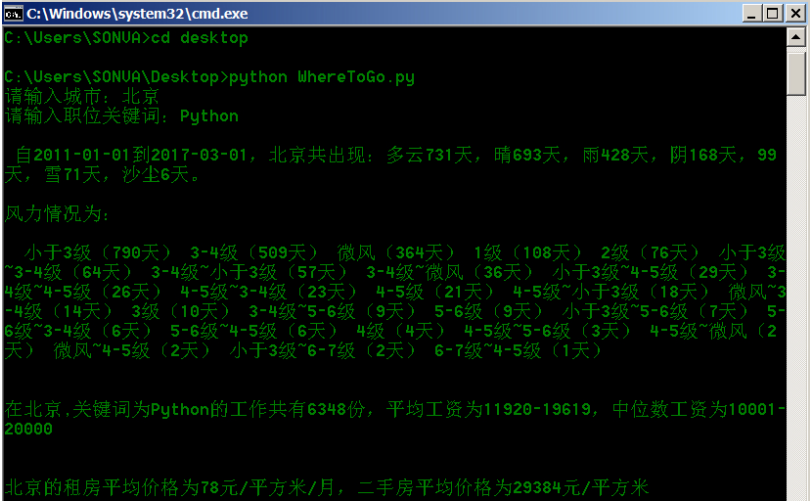
智联招聘：<http://sou.zhaopin.com/jobs/searchresult.ashx>

赶集网租房：<http://www.ganji.com/fang1>

赶集网二手房：<http://www.ganji.com/fang5>

程序代码采用了嵩老师示范的整体框架和接口设计，并在其中采用嵩老师多次强调的 try-except 结构确保爬虫的稳定可靠。作业中实践了 Beautiful 库、re、和 CSS Selector 三种网页解析的方法，也使用了一些在课程中和讨论区中学到一些小技巧。比如对爬取的 Response 对象的 encoding 属性默认使用 utf-8 编码，需要时再根据所爬取网页的 charset 字段手动修改以提高爬取速度。再比如，循环爬取几十个 HTML 页面时，用 time.sleep() 方法可以避免对服务器的性能造成骚扰，这样既符合 Robots 协议相关的精神，又能有效地提高爬取页面的成功率。

为了构造与网站服务接口相符的 URL，程序中引入了 Pypinyin 库实现从汉字到拼音的转换。同时为了实现一些统计功能，程序引入了 Python 的内建库 Statistics。下图为程序运行情况：



```
C:\Windows\system32\cmd.exe
C:\Users\SONUA>cd desktop
C:\Users\SONUA\Desktop>python WhereToGo.py
请输入城市：北京
请输入职位关键词：Python

自2011-01-01到2017-03-01，北京共出现：多云731天，晴693天，雨428天，阴168天，99天，雪71天，沙尘6天。

风力情况为：
  小于3级（790天） 3-4级（509天） 微风（364天） 1级（108天） 2级（76天） 小于3级~3-4级（64天） 3-4级~小于3级（57天） 3-4级~微风（36天） 小于3级~4-5级（29天） 3-4级~4-5级（26天） 4-5级~3-4级（23天） 4-5级（21天） 4-5级~小于3级（18天） 微风~3-4级（14天） 3级（10天） 3-4级~5-6级（9天） 5-6级（9天） 小于3级~5-6级（7天） 5-6级~3-4级（6天） 5-6级~4-5级（6天） 4级（4天） 4-5级~5-6级（3天） 4-5级~微风（2天） 微风~4-5级（2天） 小于3级~6-7级（2天） 6-7级~4-5级（1天）

在北京,关键词为Python的工作共有6348份，平均工资为11920-19619，中位数工资为10001-20000

北京的租房平均价格为78元/平方米/月，二手房平均价格为29384元/平方米
```

完整代码请参见 WhereToGo.py。

最后，谢谢嵩老师提供这么棒又这么实用的课程，连视频的配音、剪辑、课件的制作等细节上都非常地用心。谢谢袁炜佳和李天龙两位热心的助教，不但帮助扫除了很多学习上的障碍，还帮助拓展了相关的知识和视野。感谢！