

《信息安全》实验报告

**（2019）**



题目 对旅游景点信息的抓取和数据分析

作者姓名 何勇 阮天奇

学科(专业) 网络工程

所在学院 计算机科学与技术学院

提交日期 2019年 1月 15

1. **实验设计**

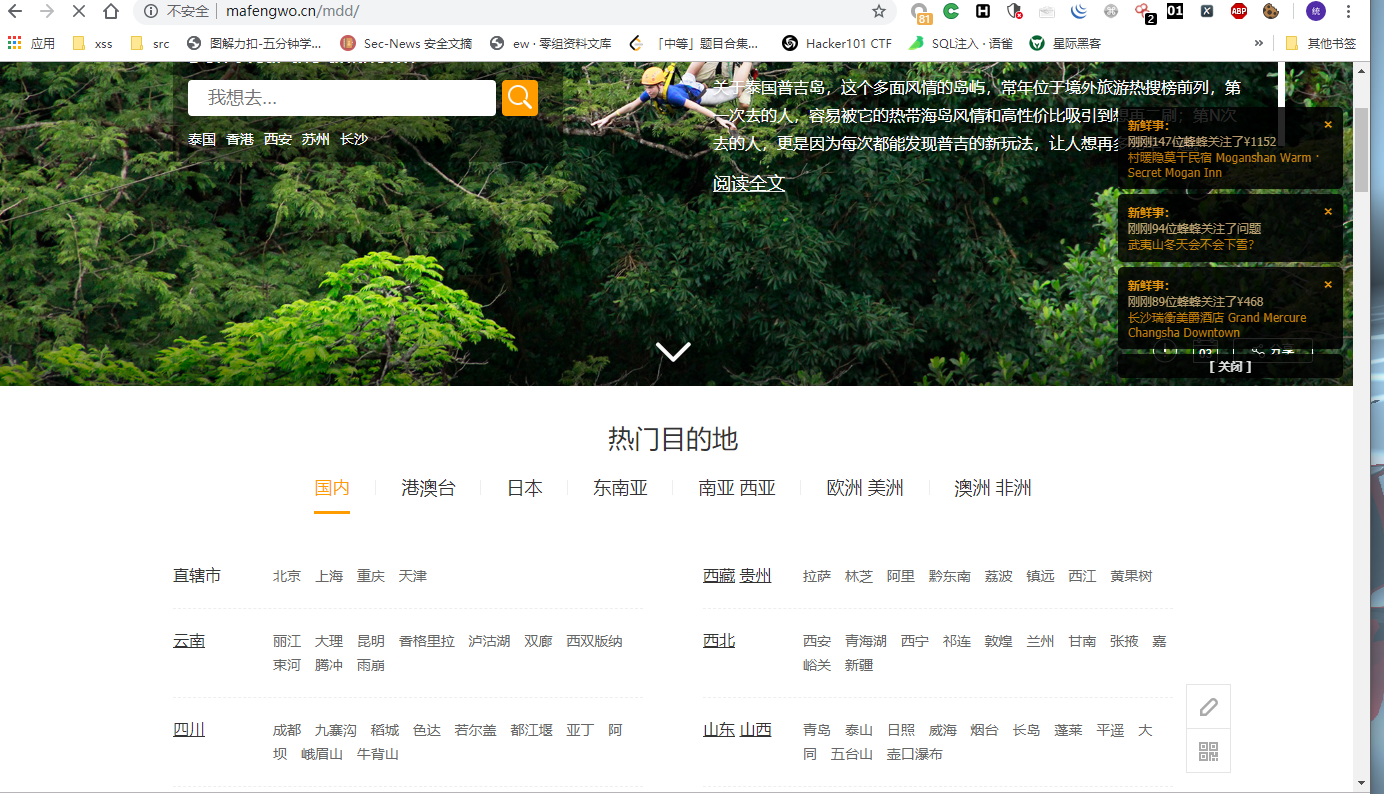
实现爬取马蜂窝网旅游景点和美食资料，并生成热力图、饼图、柱状图形象直观地体现热门景点和美食，该工具还可根据用户输入的城市进行搜索，且也可以生成数据图。

1. **实验原理**

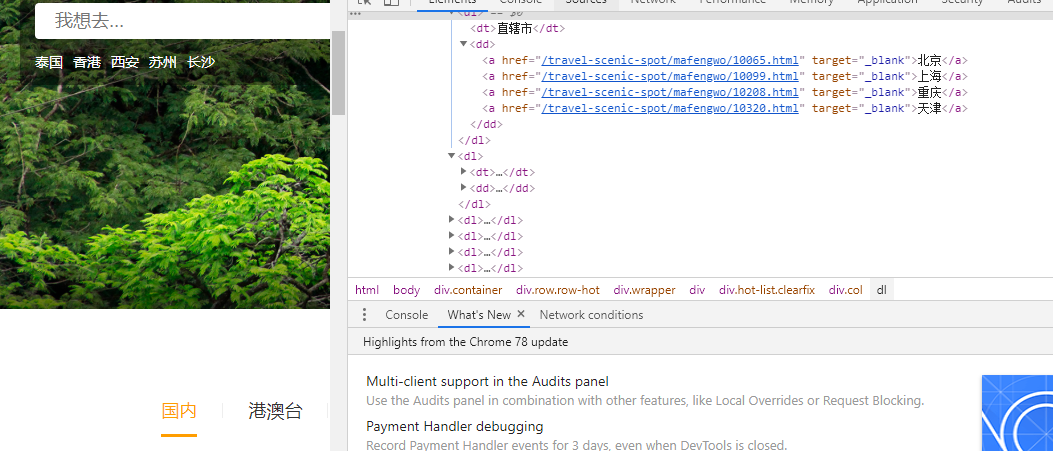
通过requests、selenium、beautifulsoup对网页进行分析以及对元素定位爬取，抓取所有热门城市的url然后逐个分析其景点和美食。通过<http://www.mafengwo.cn/search/q.php?q=>提供的搜索接口，进行指定城市的搜索和数据爬取。

最后利用pandas、matplotlib以及百度地图api进行柱状图、饼图、热力图的生成。

1. **实验过程和分析**



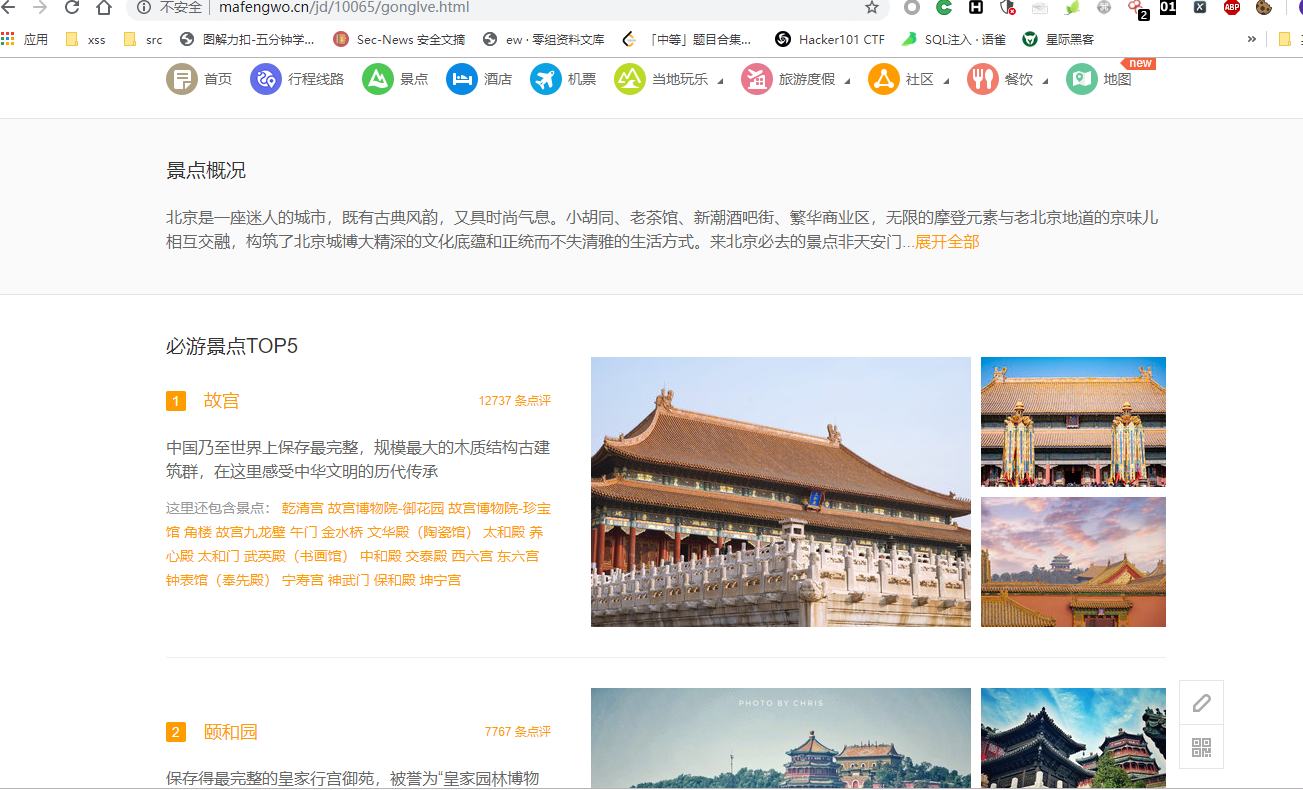
在目的地处找到国内热门目的地的url如下：



利用beautifulsoup对url进行定位批量爬取，会遇到封代理ip的问题。由于这个网站使用requests库访问难以爬取到信息所以采用的是selenium，但是selenium切换代理ip不像requests一样在每次请求时就可以换，它要在打开一个浏览器之前就换。网上没有找到这种需求的selenium该怎么写，于是我用goto完成了这个需求。

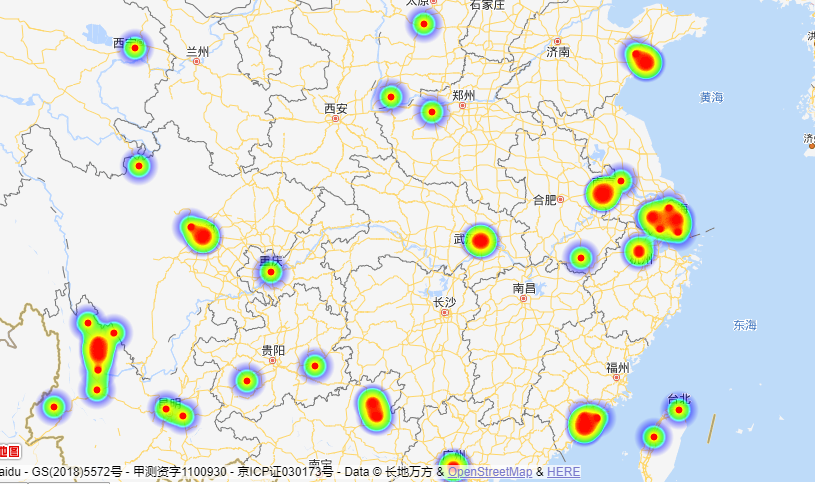


之后就是进入各城市主页进行景点和美食的爬取

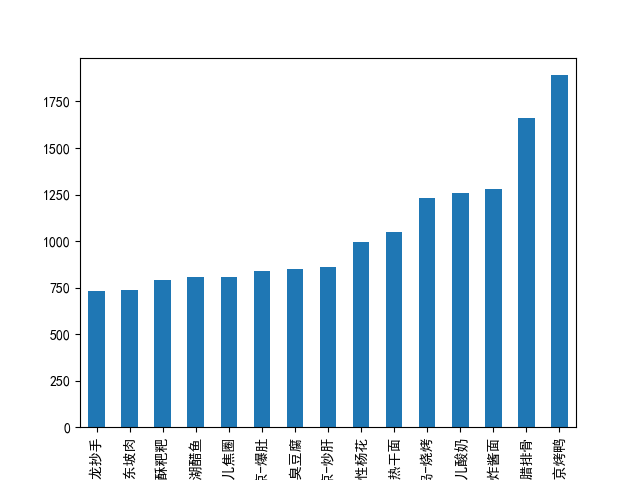


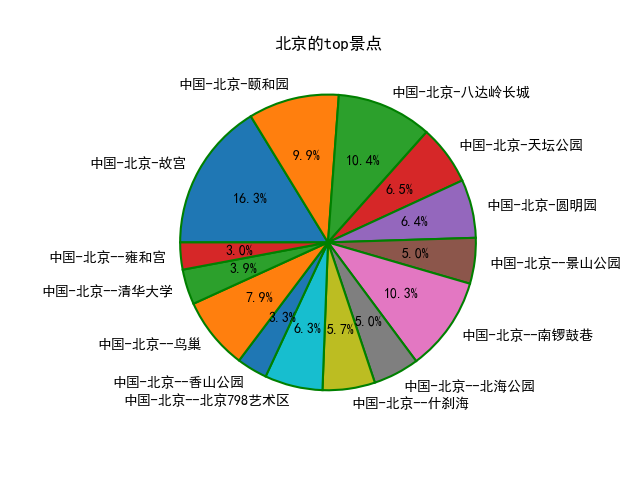


生成热力图之前需要在config.ini中配置网站根目录地址，之后使用<http://127.0.0.1/hot_map.html>进行查看热力图。



下图为美食柱状图和景点饼图：





代码分为attacker.py、analyse.py、spider.py。

Spider.py包含了所有数据爬取部分的代码，analyse.py包含了分析数据生成数据图的代码。attacker.py整合了两部分代码，实现参数化调用方法进行使用工具代码。

最终调用attacker.py即可实现所用功能。

