# 大众点评网维护更新

|  |  |  |  |
| --- | --- | --- | --- |
|  | 作者 | 版本描述 | 完成日期 |
| 1.1 |  | 基本实现对点评网数据抓取及去重 | 2017.3.1 |

## 改动介绍

因为时间过了很久，网页的结构也与之前不同了，有些数据也取不到了。所以对照之前的代码对网页又重新分析了一下，并稍微改动了一下代码，确保数据的完整性。大的方向有两条：一个是爬取代理网站的时候有改动，另一个就是在把所有数据爬取下来之后存入数据库时做了细微改动。

首先在爬取代理网站的时候，因为之前<http://gatherproxy.com/proxylist/anonymity/?t=Elite>这个网站是HTML，如今加入了JS，导致以前的提取IP的xpath规则取不到值了，所以得重新定义数据提取规则。在dianping-for-github/utils/crawl-proxy-nonblock.py文件中把之前的sels.xpath("//div[contains(@class,'proxy-list')]/table/tr")这句xpath规则改成sels.xpath("//div[contains(@class,'proxy-list')]/table/script/text()")。不再按之前的规则直接提取IP地址和端口了，而是找到网页中script内容，再从这些内容当中利用正则匹配提取IP地址和端口，获取端口的代码：int(sel.re(r'"PROXY\_PORT":"(\w+)"')[0],16)。

其次就是在spiders/spider.py文件中返回的items数据在存储到数据库的时候总有语法错误，经过分析网页得之，有的店铺已经没有了POI字段，导致空的数据在插入数据的时候报错。所以就得在取不到的字段时给一个默认值，分别在字段poi、original\_latitude、original\_longitude、info（商铺信息）、addr（商铺地址）和hour（营业时间）等这几个字段设置了默认值，确保在存储的时候不为空。把数据类型是列表格式的在存入数据库的时候提出来转换成字符串。

其他的没有什么改动，主要就完善了这两个地方。