大众点评店铺信息爬虫

作者：Tooi6
原文链接：https://ld246.com/article/1608435763808
来源网站：链滴

大众点评店铺信息爬虫

花式反反爬之抓取大众点评店铺信息。

项目地址

Gitee：https://gitee.com/Tooi/dianping

项目目录

```
│  config.py
│  dianping.py
│  parse.py
│  proxy.py
│  README.md
│  requirements.txt
│
├─utils
│      common.py
│      ua.log
│      __init__.py
│
└─view
       analysis*.png
       db.png
```

环境依赖

```
pip3 install -r requirements.txt
```

抓取流程

 美食店铺首页开始，遍历抓取每页内容。

```
    for i in range(50):
        print("第%d页：" %(i+1))
        response = self.get_store_list_page(INIT_URL.format(str(i+1)))
        self.parse_data(response)
        time.sleep(eval('%.1f'%random.random()))
        # 测试仅抓取第一页
        break
```

花式反爬

 大众点评的反爬措施很强，可以ban掉大部分爬虫。当然，误伤率也比较高。测试期间发现的反爬施有：

- 常规链接404（店铺详情页链接404页面）
- 请求头校验

- 多类型字体反爬（偏移量，自定义css）
- 验证码（常规四字中英文混合）
- cookies
- ban ip

 例如，当我用 Postman 测试链接时发现，若不用UA则返回 403，要求输入验证码方可正常访问。大众点评总的来说还是基于IP检测，所以，爬虫的重点在于：**代理IP的质量。**

反反爬

- 挂代理（加强型爬虫代理）
- Headers 添加随机 UA 和 Refer 参数
- 随机抓取时延

 注意：

- 加强型爬虫代理非一般隧道或API类型代理IP，成活率较高
- 原始headers中没有Refer参数，测试发现，添加Refer参数可提高请求头伪装效率

解释说明

1. 为测试方便，未实现自动获取css、svg_font、svg_num等链接，请自行复制（测试期间每间隔一变化一次）
2. 未处理数字、文字拼接逻辑。例如：页面要处理的数字为 1081 中的 1、0、1，实际结果可能为 81 1（日后再更）
3. config.py 中的所有代理均已失效或涂改，需要自己更换

运行

 命令行切换至根目录：

```
>>> python dianping.py
```

抓取结果

公告

 本代码仅作学习交流，切勿用于商业用途，否则后果自负。若涉及点评网侵权，请邮箱联系，会尽处理。