操

作

流

程

|  |  |  |  |
| --- | --- | --- | --- |
| 版本 | 作者 | 版本描述 | 完成日期 |
| 1.0 | Jerry | 基本实现对点评网数据抓取及去重 | 2016.08.25 |

快捷使用介绍

阶段一：前处理

1. 打开dianping/pre/classify.py文件，修改城市cityID（默认北京为2），执行该文件后获得分类对应码文件classify.py
2. 打开dianping/pre/subregion.py文件，修改REGION\_DICT变量为所要爬取的城市下辖各行政区及地区编码，通过执行该文件获得商圈码subregion.py
3. 将上述两文件复制到dianping/catalog文件夹下
4. 在本地搭建名为dianping的数据库，连接文件见dbconnect.py

阶段二：爬虫核心文件修改及启动

1. 修改spiders/spiders.py文件类变量\_city\_id,\_city\_pingyin以及REGION变量
2. 通过在spiders/spiders.py文件的第64行修改if语句来实现对不同频道的店铺信息抓取。
3. 执行Dianping/utils/crawl-free-proxy-from-gatherproxy.py不间断爬取有效代理
4. 调用scrapy crawl dianping启动爬虫

阶段二：数据表合并及去重

1. 将不同频道table数据导出并合并为一个文件再重新导入数据库 (详见下文)
2. 数据库console中执行SELECT DISTINCT `url` FROM `shops` INTO OUTFILE "shops\_url\_no\_repeat.txt"，然后将输出的文件复制到子文件夹dianping/duplicate中
3. Console执行dianping/duplicate文件夹中db\_rid.py文件实现去重处理。

详细使用说明

1. 分类对应码获取

频道分类作为最高级别的分类，其对应码可以从如下页面直接获取：<http://www.dianping.com/search/category/2/0> ，分类表格如表1.1

表1.1 点评网一级分类对应码

|  |  |  |
| --- | --- | --- |
| channel | Englishe | ID |
| 美食 | food | 10 |
| 电影 | movie | 25 |
| 休闲娱乐 | life | 30 |
| 酒店 | hotel | 60 |
| 丽人 | beauty | 50 |
| k歌 | KTV | 15 |
| 运动健身 | sports | 45 |
| 景点（周边游） | view | 35 |
| 亲子 | baby | 70 |
| 结婚 | wedding | 55 |
| 购物 | shoppping | 20 |
| 宠物 | pet | 95 |
| 生活服务 | other | 80 |
| 学习培训 | education | 75 |
| 爱车 | car | 65 |
| 医疗健康 | medical | 85 |
| 家装 | home | 90 |
| 宴会 | hall | 40 |

频道分类又可下分二级分类（图1.1），如美食频道下有火锅、咖啡厅、烧烤、面包甜点、小吃快餐等。



图1-1 频道下二级分类示例

二级分类对应码可以通过pre/classify\_get.py程序得到，输出文件classify.py将这些对应码构建成字典型变量，方便后续程序使用。图1.2为对应码部分数据

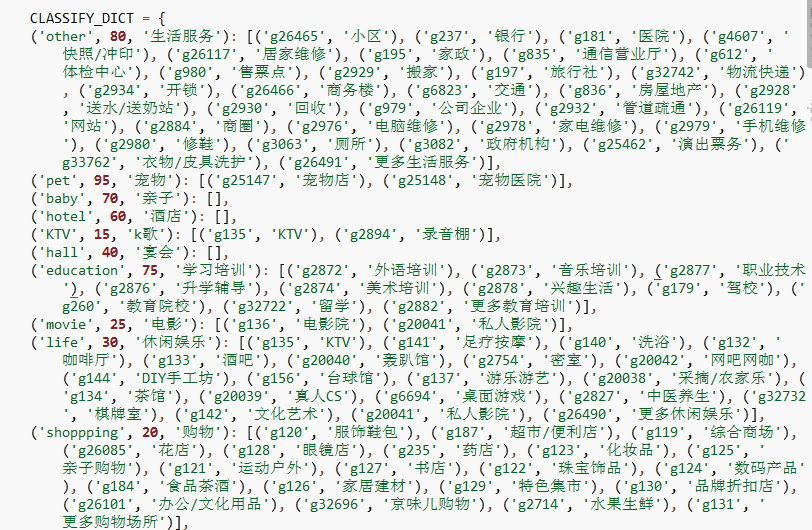


图1-2 点评网二级分类对应码

1. 地区商圈对应码获取

行政地区以及下辖商圈的对应码同样可以通过简单的脚本程序获得，程序：pre/subregion\_get.py,输出：subregion.py

表1.1 点评网一级分类对应码

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 对应码 | 朝阳区 | 对应码 | 海淀区 | 对应码 | 东城区 |
| r14 | 不限 | r17 | 不限 | r15 | 不限 |
| r2580 | 三里屯 | r1489 | 五道口 | r1475 | 王府井/东单 |
| r1471 | 望京 | r1488 | 中关村 | r1504 | 崇文门 |
| r2578 | 国贸 | r1493 | 公主坟/万寿路 | r2066 | 东直门 |
| r1466 | 朝外大街 | r2588 | 五棵松 | r1503 | 前门 |
| r1470 | 亚运村 | r1490 | 北太平庄 | r1486 | 地安门 |
| r1469 | 亮马桥/三元桥 | r1494 | 紫竹桥 | r23023 | 南锣鼓巷/鼓楼东大街 |
| r2078 | 大望路 | r1996 | 魏公村 | r23024 | 北新桥/簋街 |
| r2579 | 双井 | r1491 | 苏州桥 | r1478 | 安定门 |
| r2871 | 十里堡 | r1495 | 航天桥 | r1477 | 东四 |
| r1467 | 朝阳公园/团结湖 | r2872 | 远大路 | r2591 | 和平里 |
| r1465 | 建外大街 | r2589 | 清河 | r1479 | 朝阳门 |
| r2583 | 酒仙桥 | r2587 | 双榆树 | r1476 | 建国门/北京站 |
| r22998 | 青年路 | r1492 | 北下关 | r2590 | 广渠门 |
| r1472 | 劲松/潘家园 | r1496 | 上地 | r1505 | 天坛 |
| r2584 | 管庄 | r1497 | 颐和园 | r23022 | 雍和宫/地坛 |
| r1473 | 安贞 | r1498 | 海淀其它 | r2875 | 沙子口 |
| r22997 | 太阳宫 | r23034 | 万柳 | r23021 | 东四十条 |
| r1474 | 朝阳其它 | r7510 | 香山 | r2874 | 左安门 |
| r1468 | 左家庄 | r23033 | 人民大学 | r23026 | 沙滩/美术馆灯市口 |
| r23019 | 常营 | r23032 | 四季青 | r23025 | 光明楼/龙潭湖 |

1. 本地mysql数据库配置

Mysql数据库的创建以及table的创建详见相关软件文档，数据库与scrapy框架的连接文件见dianping/dbconnect.py,这里需要注意的是，如果使用的是python2,需要导入mysqldb模块，如果使用python3,建议pip安装pymysql模块。

1. 爬虫文件修改与执行

修改spiders/spiders.py文件类变量\_city\_id,\_city\_pingyin以及REGION变量，这里，REGION变量的设定可以参考pre/subregion.py文件，此外，该变量元素中第三项用以处理酒店频道下url的构建。

修改spiders/spiders.py文件的第64行if语句来实现对不同频道的店铺信息抓取，爬虫文件执行代码：scrapy3 crawl dianping –logfile=channel\_name.py. 每次执行后将在mysql数据库中生成名为shops的数据表，及时将其更改为对应的频道名，以免下一次操作将本次操作数据冲掉。

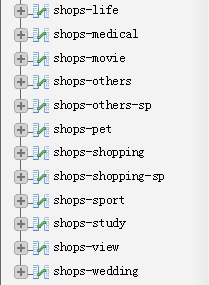
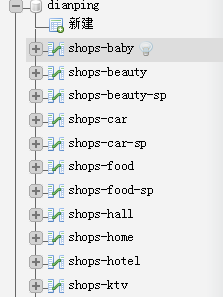


图1-3 点评网频道抓取数据表

1. 数据合并及去重

1）数据库中导出无重复数据txt

将18个数据表table导出成sql文件，使用notepad++打开各个文件，然后除去reconds中id，并将其合并成一个sql文件，重新导入成数据库，这样就形成一个合并后的数据表shops。

在数据表shops中执行如下语句：

SELECT DISTINCT `url` FROM `shops-baby` INTO OUTFILE "shops\_url\_no\_repeat.txt"

这样，生成的文件shops\_url\_no\_repeat.txt就包含了数据表shops所有无重复的url纪录，将这一文本文件复制到dianping/duplicate下。

2）反查数据库并删除重复数据

在console执行dianping/db\_rid.py文件，这一程序实现从数据库中查找出重复（根据店铺url）纪录，然后删除掉这些重复纪录。