## 快递包裹数-数据抓取需求

### 一、背景

由于报告中添加了新的维度和指标，需要计算各季度的快递包裹数，所以需要数源部门提供支撑，抓取邮政管理局快递包裹数相关数据，保证季报数据的按时生产。

抓取范围：各地区邮政管理局从2015年1月到2019年1月邮政业发展情况报告中的数据

第一步抓取的地域：所有合作地域（详见附件或第五章）

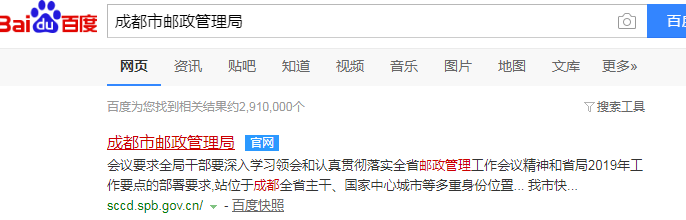
第二步抓取的地域：所有省、市（地域表中ID长度为2和4的）

当然，能一次性全抓下来最好！

### 二、具体抓取需求

本次抓取的对象是各省市邮政管理局网站的各月份邮政行业运行情况报告中的邮政相关数据。由于不同网站的步骤基本相同，因此以下以成都市为例进行解释，特殊情况在第三部分注明。

#### 1.登陆成都市邮政管理局官网。



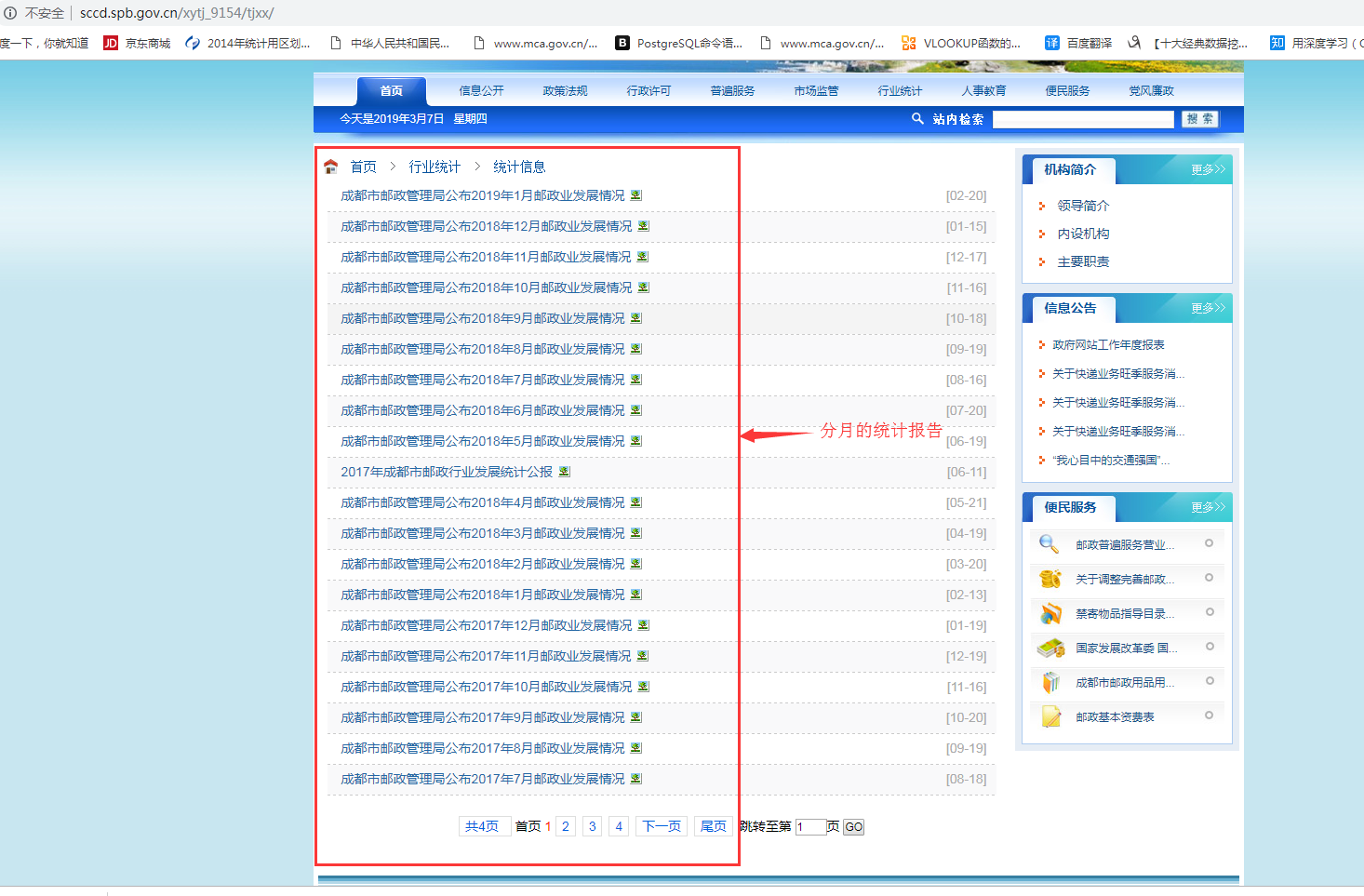
网站界面



#### 2.在导航栏上选择 行业统计🡪统计信息 进入。



#### 3.统计信息下能看到成都市各月份邮政业发展情况报告，其中包含各月份邮政相关数据，每篇报告格式较为统一。需要遍历抓取2015年1月到2019年1月的所有报告中快递业务量的数据



#### 4.从每份报告中提取数据

以2019年1月报告为例，打开报告，提取报告文字中关于快递服务企业业务量的数据，例如成都市报告中需要提取：（1）快递服务企业业务量完成：1.02亿件

（2）业务收入完成：11.72亿元

红字部分为需要提取的内容。注意：业务量和业务收入单位为**亿元或亿件**的，需要转换为**万元或万件**



### 三、保存格式

数据表保存格式参照以下格式（id递增）。需要抓取所有省、市从2015年1月到2019年1月的数据。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| id递增即可 | 地区编号 | 年月 | 业务量（万件） | 业务收入（万元） |
| id | areaid | date | amount | money |
| 1 | 5101 | 2015-01 | 10200 | 117200 |
| …… | …… | …… | …… | …… |
| …… | …… | …… | …… | …… |
| …… | …… | 2019-01 |  |  |

### 四、特殊情况说明

在此告知部分已发现的特殊情况，如还有例外，再讨论和对接。

##### 1.报告格式

有些省市的报告是以附件形式挂在网上的，如江苏省是word附件，苏州市为PDF文件。



##### 2.报告位置

（1）有些网站 **行业统计** 下无内容，如济南市邮政管理局网站，直接进入行业统计即可；有些不仅有统计信息，还有统计制度，仍然进入统计信息。



（2）有些网站 没有行业统计栏目，如拉萨市，查看报告的路径为：信息公开🡪统计及数据分析





##### 3. 12月份数据位置

有些地区12月份数据在年度报告中，而不在12月份报告中，如青岛市。



### 五、时间

由于3月20日就要生产季报，时间紧迫，因此数据抓取分为两个部分完成：

（1）抓取所有合作地域的邮政管理局的快递业务量数据，抓取时间希望控制在3月18日或之前。附一张需要抓取的地域表（同邮件附件）

|  |  |
| --- | --- |
| **地域** | **地域ID** |
| 成都 | 5101 |
| 达州 | 5117 |
| 大理 | 5329 |
| 广安 | 5116 |
| 海口 | 4601 |
| 海南 | 46 |
| 河北 | 13 |
| 湖北 | 42 |
| 吉林 | 22 |
| 济南 | 3701 |
| 昆明 | 5301 |
| 乐山 | 5111 |
| 凉山 | 5134 |
| 泸州 | 5105 |
| 绵阳 | 5107 |
| 南充 | 5113 |
| 内江 | 5110 |
| 宁夏 | 64 |
| 四川 | 51 |
| 雅安 | 5118 |
| 宜宾 | 5115 |
| 云南 | 53 |
| 昭通 | 5306 |
| 重庆 | 50 |
| 自贡 | 5103 |
| 宝鸡 | 6103 |
| 巴彦淖尔 | 1508 |
| 内蒙古 | 15 |

（2）抓取所有省、市（地域表中ID长度为2和4的）的邮政管理局的快递业务量数据，时间可在季报时间之后进一步商榷。

时间紧迫，请数源部的同事尽快评估可行性及时间，并及时给出回复（邮件）。

感谢！