**抓取数据覆盖率总结报告**

**一、抓取数据覆盖率总结**

成都抓取数据分别以公示系统和信用平台作为数据池，总计在公示系统抓取工商数据92万条，在信用平台抓取数据178.7万条。

利用搜索人工匹配（即根据绘图店名直接在公示系统搜索）时得到的工商数据作为样本，根据工商注册号（共计2363个）分别与公示系统抓取数据和信用平台抓取数据进行精确匹配，统计公示系统抓取数据和信息平台抓取数据的覆盖率，统计结果如下：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 数据来源 | 抓取数据量 | 搜索数据 | 匹配量 | 覆盖率 |
| 公示系统 | 920654 | 2363 | 1747 | 73.93% |
| 信用平台 | 1786738 | 2363 | 1432 | 60.60% |

*备注：搜索数据即搜索人工匹配、第三方网站搜索时得到的工商数据，该数据主要为双流区、武侯区数据；*

由于搜索数据主要为双流区、武侯区的数据，且绘图商户主要以个体工商户为主，故特统计了两种抓取数据在上述条件下的分配，统计结果如下：

|  |  |  |  |
| --- | --- | --- | --- |
| 数据来源 | 抓取数据量 | 武侯/双流数据量 | 个体工商户量 |
| 公示系统 | 920654 | 129914 | 123923 |
| 信用平台 | 1786738 | 275818 | 131602 |

总体而言，虽然信用平台抓取数据量较大，但个体工商户的样本量仅占总数据的50%左右。

对两方数据匹配上的结果合并处理，用工商注册号去重后，匹配结果为2337个，故总的覆盖率为98.90%。

**二、后续工作**

为验证上述结果的正确性，接下来准备对海霸王商圈做数据集成工作，由于数据量较大，主要选择登记机关为郫县、四川工商注册、以及郫县周围区县的抓取数据作为匹配数据库。