# **面试题报告**

1. 爬虫为spider.py
2. 爬取的原数据存储为legal\_proceedings\_data.csv
3. 经过将数据提取，最终提取的人名或公司名存储为name\_no\_and.txt
4. 总结所有的诉讼结果文档，我将1945-2018这个时间段的诉讼结果文档每年的数目用pyecharts做了分析（文件为：year\_analysis\_line.html），结果如图4.1：



图4.1

从折线图反映出在1945-2001年诉讼结果文档都在增加（我的猜测：在1997香港回归，那时事故比较多，因此处于上涨状态），在2001年时达到高峰，2001-2018整体处于下降趋势（我的猜测：在2001年以后，香港社会秩序相对稳定，法律体系越来越完善，因此诉讼结果文档数目处于下降状态）。

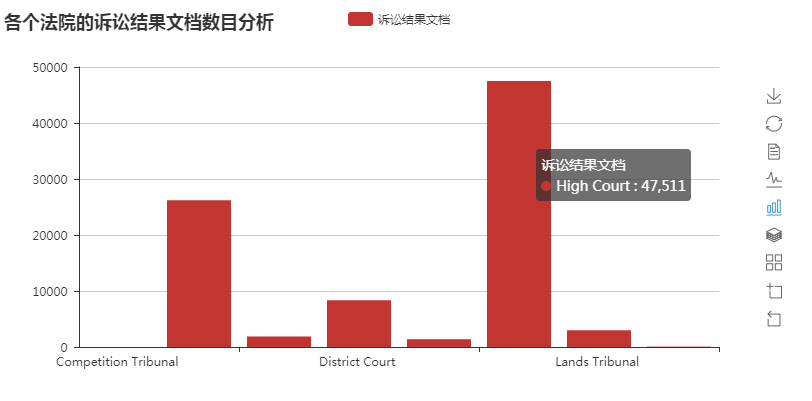
接着总结了各个法院诉讼文件的数据，做了柱状图如图4.2：

图4.2

从图中看出High Court诉讼文档最多，第二多的为[Court of Appeal of the High Court](https://legalref.judiciary.hk/lrs/common/ju/judgment.jsp?L1=CA#H2) ，也就是说一搬的诉讼结果文件都存于高等法院。

5.在不考虑去重的情况下，HKSAR(香港特别行政区（政府）)涉及最多的诉讼案件。

# **总结**

在爬取数据的时候，面临的问题是：目录层级多，分页嵌套较多，数据动态加载且数据直接解析很繁琐。于是我建立了爬取方案，使用selenium+phantom模拟浏览器加载动态数据，然后制定相应的规则来实现分页爬取，接着使用BeautifulSoup进行数据解析，最后将数据存储到CSV文件中。虽然效率慢一些，但是爬取的准确性很好。

解决了数据爬取之后，紧接着就是数据分析，提取了人名和公司名，然后通过分析得到：在不考虑重复的情况下“HKSAR“涉及诉讼案件最多，这也是让我觉得比较有趣的地方。另外，也将这些数据做了可视化，发现了最高法院（包含High Court，[Court of Appeal of the High Court](https://legalref.judiciary.hk/lrs/common/ju/judgment.jsp?L1=CA#H2) ）的诉讼结果文件是最多的，以及2001年诉讼结果文件最多，在2001年之前总体呈上升趋势，2001年之后总体呈下降趋势,这样的趋势和香港的回归有关，在香港回归的那几年比较动乱，因此诉讼结果文档会暴增，后面随着香港社会的安定诉讼结果文件呈下降趋势。

此道面试题让我学习到了新的知识点也巩固了以前学习的知识，但还有不足，感谢贵公司的考核，望加入贵公司团队！

姓名：吴志强

申请职位：大数据爬虫工程师