# 法律参考资料系统资料爬取与分析

1. 数据爬取

使用selenium+Beautiful Soup对网站进行数据抓取。

通过分析网页数据的构成，在网页加载的过程中涉及到js代码生成HTML，因此，使用selenium+chromedriver进行动态数据抓取。抓取到HTML文件后，使用Beautiful Soup进行结构解析以及提取数据。

所提取的数据字段为：

COURT 法院

CASE\_TYPE 案件类型

CASE\_YEAR 案件年份

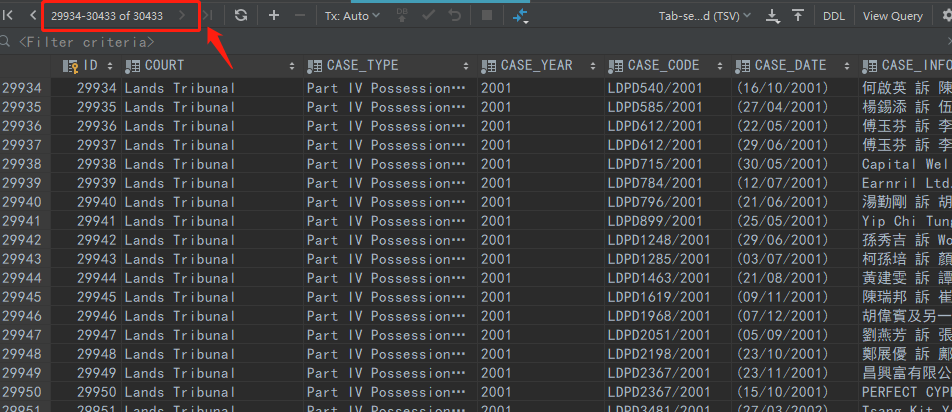
CASE\_CODE 案件代码

CASE\_DATE 案件具体日期

CASE\_INFO 案件信息（诉讼人）

CASE\_FILE 案件诉讼文件链接

最后，将数据存储到sqlite中，共抓取数据30433条：



具体代码见：<main.py>

1. 数据分析
   1. 案件诉讼趋势分析

下图是1945年到2018年的诉讼案件数量变化情况。

从该图中我们可以发现，1945到2018年，总体诉讼量呈上升趋势。

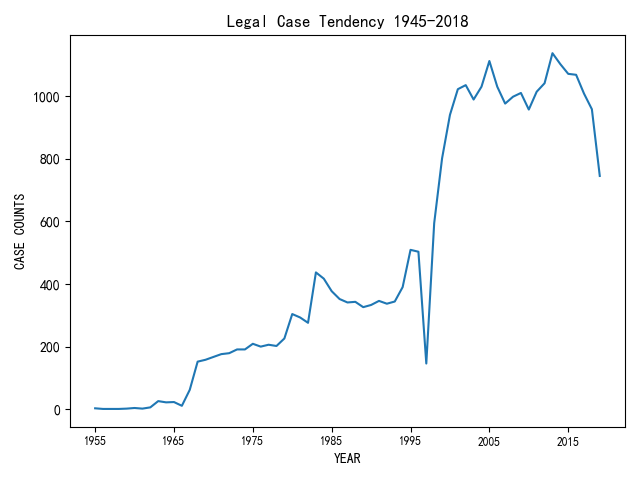
1945-1995年间，香港当时处于殖民地阶段，在英国的管理下，案件的处理数量处于较低的水平。

1995-1998年间，香港正式回归，从图中可以看是这段时间的诉讼量锐减，原因应该是殖民政府忙于撤离香港事宜，而无法处理太多的诉讼案件。

1997年，香港回归中国之后，诉讼案件数量激增，应该是前期尚未处理的案件堆积在一起，造成数据波动如此之大。

此后，2000-2011十多年时间里，诉讼案件的数量一直维持在一个较高的水平，猜测应该和回归中国后经济的发展有关，经济发展过快造成了一定的财产纠纷。

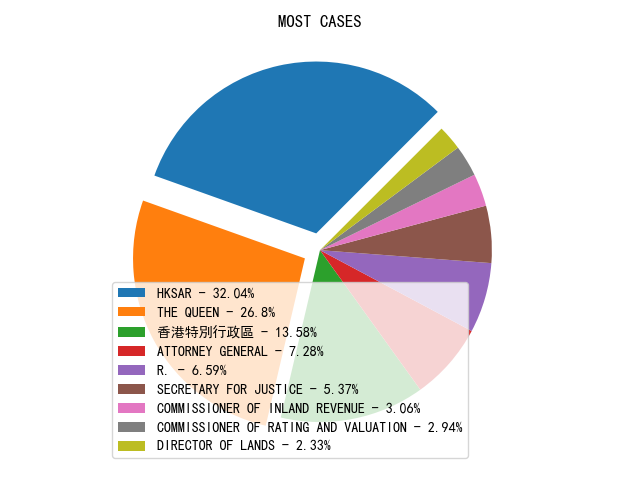
而从2011年到2018年，案件诉讼量锐减，体现出香港地区得到有效的治理。



图一 1945年到2018年的诉讼案件趋势

* 1. 诉讼案例相关者分析

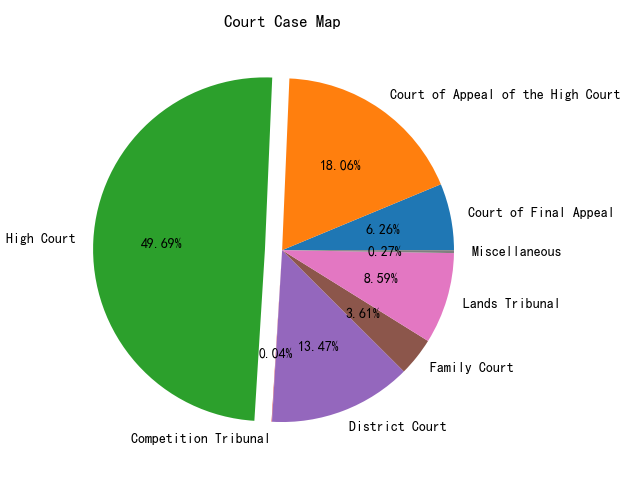
从下图可以看到，诉讼案例最多的前9个相关组织/个人中，HKSAR香港特区政府占了最高比例，达到32.04%。而在相关者中，大都是一些组织。因此，1994年至今，组织之间的纠纷占了相当大一部分。



图二 前9个诉讼案例最多的相关者

* 1. 诉讼法院分析

从下图可以看到，1945年至2018年，处理诉讼案件最多的法院为High Court，案件比例高达49.49%，其次是Court of Appeal of the High Court，最高法院上述法庭，比例达18.06%。



图三 诉讼法院占比

分析具体代码见：<analyze.py>