爬虫测试

任务

- 利用 Python/BeautifulSoup，建立一个法律诉讼结果的数据库 (https://legalref.judiciary.hk/lrs/common/ju/judgment.jsp?L1=FA#H1)
- 如图一所示，每个法院分类都有子分类。
- 如图二所示，每个子分类都有不同的法律诉讼结果，每一宗案件都有日期和编号
- 所有分类和内容必须记录
- 输出结果为 CSV 或 sqlite

图 1

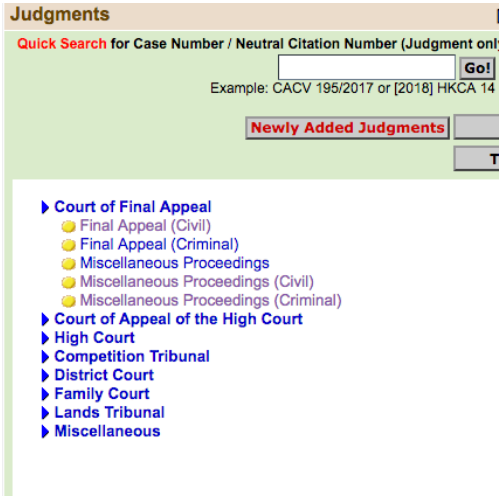

图 2

加分项目

- 利用 NER (实体抽取) 把每篇文档的人名和公司名都提取出来。
- 总结所有诉讼结果文档，发现什么有趣的分析?
- 在不考虑去重名的情况下，谁或那家公司涉及最多的诉讼案件?

时限

- 6 天