

# 司法搜索引擎系统的设计与实现

熊天翼	严澜
计 95	建 82
2019011303	2018010005

## 1 需求描述

根据给定的中文法律数据集，实现一个司法搜索引擎系统。

- 核心功能：用户界面，关键词检索，案例详情监视。
- 其他功能：高级检索，关键词高亮，标签显示与搜索，基于法条搜索，类似案例推荐，关键词高级检索，以案搜案。

## 2 框架设计

- 前端：Vue 框架 + Element UI，基于色色引擎 UI 实现。  
<https://github.com/YunYouJun/sese-engine-ui>
- 后端 Django + Whoosh + sqlite3

## 3 核心模块设计

### 3.1 数据的导入与存储

用 `xml.etree.ElementTree` 库对 xml 文件进行解析，提取全文、文首、当事人、法律条文、案件基本情况、裁判分析过程、判决结果、文尾、文书名称、审判程序、法庭、审判理由、省份、年份共十四个字段，使用 `tree.getroot().find('./xxxx').get('value')` 找到相应字段。剔除不满足上述要求的法律案件，总共成功从 68417 个法律案例中导入了 50548 个案件。

在 django 中，针对案件 (Case) 和法条 (Law) 分别建立模型，对应于 sqlite3 数据库中的两张数据表。模型对应的字段如下

```
class Law(models.Model):
    id = models.AutoField(primary_key=True) # 法律id
    name = models.CharField(max_length=50) # 法条名称 (在导入数据时保证不出现重复)
    def __str__(self) -> str:
        return self.name

class Case(models.Model):
    id = models.AutoField(primary_key=True)
    qw_value = models.TextField(max_length=500) # 全文
```

```

head = models.CharField(max_length=200) # 文首
related_people = models.TextField() # 当事人
judicial_record = models.TextField() # 诉讼记录
basic_info = models.TextField() # 案件基本情况
judgement_process = models.TextField() # 判决分析过程
result = models.TextField() # 判决结果
tail = models.TextField() # 文尾
note_name = models.CharField(max_length=30) # 文书名称
judge_prop = models.CharField(max_length=30) # 审判程序（一审/二审）
court = models.CharField(max_length=30) # 经办法院
case_reason = models.CharField(max_length=30) # 案由
province = models.CharField(max_length=200) # 行政区划-省份
year = models.IntegerField(default=2023) # 年份
laws = models.ManyToManyField(to=Law, related_name="cases") # 法律条文

```

由于一个案件中可能与多个法条相关，且同一法条可能

## 3.2 使用 Whoosh 框架建立索引与分词

使用 Whoosh 作为搜索引擎框架，Whoosh 是一个用 Python 编写的全文搜索引擎库，支持倒排索引的自动建立，提供简单易用的 API，支持快速搜索、排序和定制。

使用 Whoosh 进行中文分词时，需导入 Whoosh 源代码 `whoosh_backend_ZW.py`，替换部分不兼容包，并将 `analyzer` 替换为 `ChineseAnalyzer()`，最后在 `settings.py` 中设置对应路径。Whoosh 自动建立分词和倒排索引，存放在 `whoosh_index` 目录下。

## 3.3 前端页面

前端页面共有三个，入口首页面、查询页面和详情页。

入口首页和查询页面均支持基础搜索、高级检索和类案检索三种搜索功能。详情页面展示案件具体内容，以及相关法律条文和案例推荐。

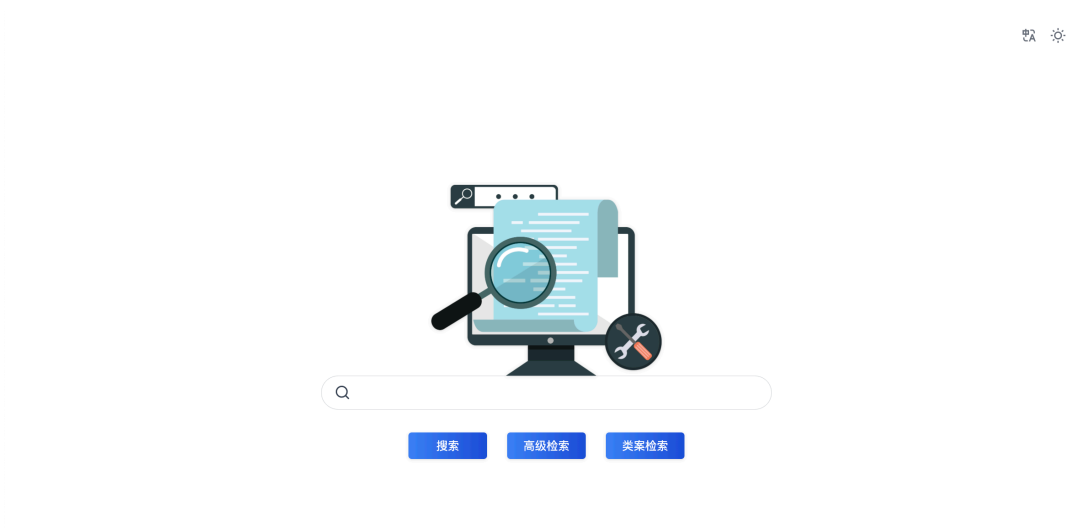


图 1: 搜索引擎首页面

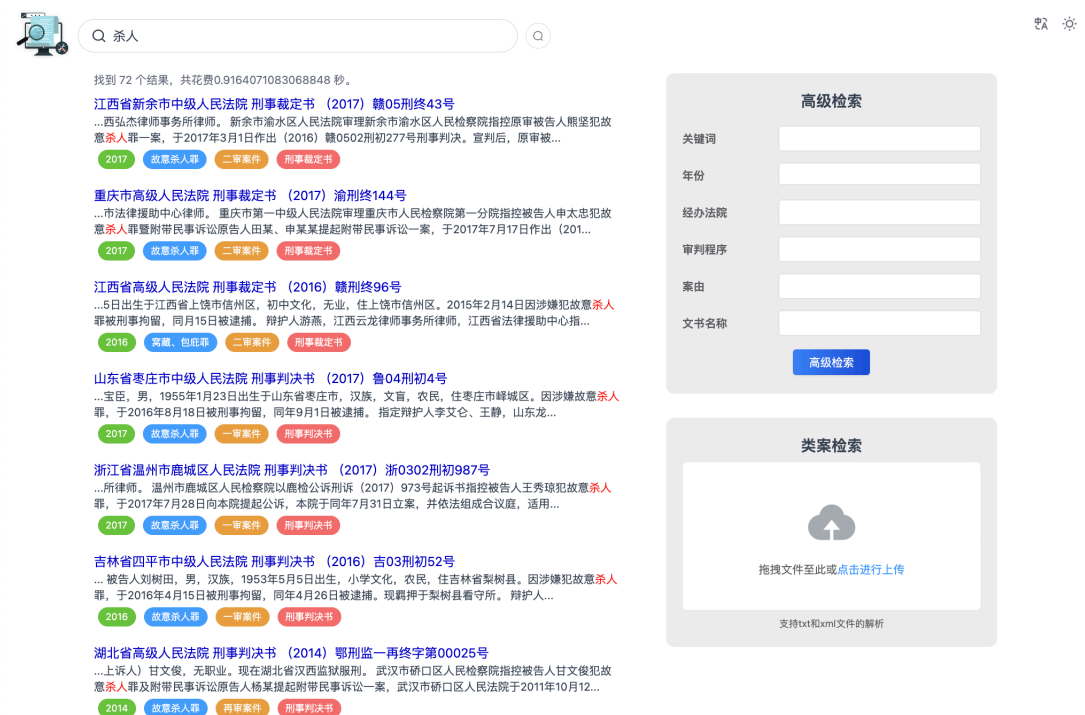


图 2: 搜索引擎查询页面

[返回](#)[搜索](#)

## 浙江省温州市鹿城区人民法院 刑事判决书（2017）浙0302刑初987号

年份：2017  
经办法院：浙江省温州市鹿城区人民法院  
审判程序：一审案件  
文书名称：刑事判决书  
案由：故意杀人罪

当事人：公诉机关浙江省温州市鹿城区人民检察院。被告人王秀琼，女，1966年11月28日出生，汉族，四川省南充市人，文化程度初中，捕前无业，家住四川省南充市高坪区。因本案于2017年2月12日被抓获，次日被刑事拘留，同年3月9日被逮捕。现羁押于温州市鹿城区看守所。

诉讼记录：援助辩护人周春琴，浙江建桥律师事务所律师。温州市鹿城区人民检察院以鹿检公诉刑诉（2017）973号起诉书指控被告人王秀琼犯故意杀人罪，于2017年7月28日向本院提起公诉，本院于同年7月31日立案，并依法组成合议庭，适用简易程序公开开庭审理了本案。温州市鹿城区人民检察院指派检察员周某、助理检察员陈某出庭支持公诉，被告人王秀琼及其援助辩护人周春琴到庭参加诉讼。被告人王秀琼自愿认罪。现已审理终结。

案件基本情况：经审理查明：被告人王秀琼系本区仰义街道蓝某立洗染厂员工。2017年2月12日18时许，被告人王秀琼在蓝某立洗染厂看到工友王某2的儿媳唐某与她的女儿即被害人袁某1（2017年1月9日出生）正在王某2的暂住处内，被告人王秀琼认为自己与王某、王某2夫妇俩有矛盾，遂产生报复念头，即以要抱孩子为由将袁某1抱在手上，然后快步跑到厂外面的桥上，将袁某1扔到河里，紧追出来的唐某及其闻讯赶来的丈夫袁某2跳到河里将袁某1救起。随后，袁某2的哥哥袁某3报警，公安人员赶到现场将被告人王秀琼抓获。经现场勘验检查：桥面距离河面的高度为250CM，小河水深19CM。同年2月14日，袁某1因“溺水后40小时”入院（温州医科大学附属第二医院）治疗8天，有明显的溺水病史，临床诊断“溺水后，吸入性肺炎”。经鉴定：被害人袁某1损伤是否客观存在无法认定；被告人王秀琼伴有精神病性症状的抑郁症，刑事责任能力为限制刑事责任能力。在审理期间，被告人王秀琼家属已代为赔偿被害人袁某1经济损失人民币10000元，双方就此达成和解协议，被害人袁某1的法定代理人袁某2、唐某对被告人王秀琼表示谅解。上述事实，被告人王秀琼在开庭审理过程中亦无异议，且有证人唐某、袁某2、袁某3、季某、王某1的证言、调取证据清单、视频光盘及视频说明、病历证明、出生医学证明、制作说明、情况说明、相关辨认笔录、检查笔录、现场勘验检查笔录及照片、法医学人体损伤程度鉴定书、温州律证司法鉴定所的司法鉴定意见书、和解协议书、谅解书、归案经过、户籍信息及被告人王秀琼在侦查机关的供述等证据证实，足以认定。关于本案情节是否属于较轻的问题。本院认为：判断故意杀人罪中的“情节较轻”，主要是考虑行为入实施的杀人方式是否恶劣、行为是否具有可宽恕的杀人动机、行为入是否有再次实施犯罪的可能性及民众所认同的道理与情感。在司法实践中，故意杀人罪“情节较轻”的情形包括义愤杀人、大义灭亲、帮助自杀、受被害人长期迫害而杀人等。结合本案，被告人王秀琼出于报复目的，站在桥面上将出生只有30余天的婴儿即被害人袁某1扔到桥下的小河里，虽然小河水深只有19CM，但桥面距离河面的高度有250CM，而且根据照片显示，小河河底还有石块、木条等杂物，足以威胁到袁某1的生命安全。因此，被告人王秀

## 相关法律条文

《中华人民共和国刑法》第六十七条  
《中华人民共和国刑法》第二百三十二条  
《中华人民共和国刑法》第二十三条  
《中华人民共和国刑法》第十八条

## 案例推荐

安徽省六安市中级人民法院 刑事裁定书（2017）皖15刑终205号  
...机关六安市金安区人民检察院。上诉人（原审被告）马永乐，男，汉族，1966年12月24日出生，安徽省霍邱县人，大专文化，原霍邱县地税局征管分局局长，副科级，住安徽省霍邱县。被告人马永乐因...

广东省高级人民法院 刑事裁定书（2017）粤刑终707号  
...机关广东省汕头市人民检察院。上诉人（原审被告）吕敬胜，男，1995年9月2日出生于江西省九江市，汉族，初中文化，职业打工，户籍地江西省九江市都昌县。因本案于2016年4月4日被羁押，同...

湖北省高级人民法院 刑事判决书（2014）鄂刑监一再终字第00025号  
...机关武汉市硚口区人民检察院。申诉人（一审被告人）吕敬胜，男，1995年9月2日出生于湖北省汉西监狱服刑。武汉市硚口区人民检察院指控被告人甘文俊犯故意杀人罪及附带民事诉讼原告人杨某提起附带民...

河北省张家口市桥东区人民法院 刑事附带民事判决书（2018）冀0702刑...  
...关张家口桥东区人民检察院。被告人崔某某，男，1968年12月24日出生于张家口市桥西区，汉族，初中文化，群众，原系张家口市泰达保安服务有限公司保安，捕前住张家口桥东区，2015年10...

河南省新乡市牧野区人民法院 刑事判决书（2018）豫0711刑初194号  
...关新乡市牧野区人民法院。被告人王伟涛，男，1993年3月20日出生于河南省原阳县，汉族，初中文化，农民，住河南省原阳县。因涉嫌危险驾驶罪，于2018年2月23日被刑事拘留，于2018年...

图 3: 搜索引擎详情页面

## 3.4 基础搜索

基础搜索支持对于给定搜索词句，返回相关的案例。用户可在入口首页面或者查询页面输入需要搜索的词句，随后搜索引擎将采用 Whoosh 框架自动对搜索词句进行分词去除停用词等处理，对于符合要求的返回结果进行随机排序返回，返回的结果采用分页器支持翻页查看。

对于每一条返回结果，支持案件标题的有限长度显示，搜索关键词的高亮显示（红色），法律案件关键标签的显示（彩色 tag），以及对应标签类别的法案的扩展搜索（点击彩色 tag 可自动返回同类案件）。

用户点击感兴趣的法律案件标题即可进入详情页，详情页分为左右两部分，左侧包含案件全文和案件关键标签的汇总展示（包含年份、经办法院、审判程序、文书名称和案由），右侧为案件涉及的法律条文一览，以及案例推荐。

## 3.5 高级检索

高级检索支持关键词、年份、经办法院、审判程序、案由、文书名称六个字断的共同匹配。用户可以在查询页面使用高级搜索，选取六个字断中的任意一到六个进行安检搜索，搜索引擎将返回各个字段搜索结果的交集。

## 3.6 类案检索

类案检索支持上传案件并检索相似案件。

用户可以任意文件格式上传法律案件，如果文件为 xml 文件并包含文首、当事人、法律条文、案件基本情况、裁判分析过程、判决结果、文尾、文书名称、审判程序、法庭、审判理由、省份、年份十四个字段，搜索引擎采用基于案由和文书名称的匹配方式返回相似案件。

对于其他格式的文件或者不满足上述十四个字段的 xml 格式案件，我们收集到了案件关键词，搜索引擎将基于给定的案件关键词匹配相似案件。

//加人工放入的关键词

## 4 测试结果及样例分析

### 4.1 关键词测试

//2020 北京盗窃

//强奸

### 4.2 案例测试

//xml

//江哥

//版权

## 5 实现功能明细

### 5.1 搜索

### 5.2 筛选

### 5.3 推荐

### 5.4 界面

## 6 参考资料

在作业的实现过程中，我们主要参考了 django、Vue、Whoosh、ElementUI 的说明文档，并对于具体问题在参考了 CSDN 和 stackoverflow 上的解决方案。在作业的开始阶段，我们学习了网络上的一份 Django + Haystack + Whoosh 实现的搜索引擎代码，了解了搜索引擎的大致实现方式。

## 7 总结

感谢刘老师和艾老师的辛勤付出和专业教导，以及两位助教学长的耐心答疑解惑，在本学期的搜索引擎课程中，我们对搜索引擎有了全面的认识，不但了解到了各个子系统的原理和技术，也在业界的专家讲座中管窥了推广搜的实战。

在本次法律搜索引擎的大作业中，我们共同研究和实践了不同搜索框架的使用，我们一起分析需求、制定计划，并协作完成了法律搜索系统的搭建和调优。通过大作业，我们深化了索引构建、查询解析、性能优化等方面的实际操作技能，同时也锻炼了团队合作和沟通能力。

## 7.1 小组分工

- **熊天翼**：前端及部分后端，基础搜索与高级搜索，共同完成代码调试和文档撰写。
- **严澜**：后端及部分前端，类案检索，共同完成代码调试和文档撰写。