

# 司法搜索引擎 - 实验报告

## # 问题描述

使用助教提供的 `Legal_data.zip` 数据集（XML格式），实现一个Demo级别的司法搜索系统。需要实现前端界面、关键词检索、类案检索等必做功能，以及精细化检索、高亮、标签显示、查询推荐、相似案例推荐等选做功能。

## # 实现模块

### 预处理

`./process/` 为预处理的文件夹，用于离线地得到各种后端需要的数据。

各模块如下：

`extract_xml.py`：

预处理XML文件中的内容，生成容易访问的json文件，提取了正文、标题、文书制作单位、文书种类、经办法院、文书名称、案号、案件类型、省份、城市（最后没有使用到）、年份、案由、案件类别（最后处理成几类）、涉及法条。

最后得到结果的一个例子如下：

```
{
```

"text": "浙江省东阳市人民法院 民事判决书 （2016）浙0783民初17571号  
原告：韦斌姬，女，1972年9月22日出生，汉族，住东阳市。 被告：韦斌强，男，1969年6月17日出生，汉族，住东阳市。 被告：杜满萍，女，1968年11月25日出生，汉族，住东阳市。 委托代理人：陈菊华、贾凌珂。 原告韦斌姬为与被告韦斌强、杜满萍民间借贷纠纷一案，于2016年12月1日向本院提起诉讼，请求判令两被告归还借款10万元，并支付利息（自起诉之日起按中国人民银行同期同档次贷款基准利率计算至实际履行之日止）。本院受理后，依法由审判员甘震适用简易程序独任审判。被告杜满萍在提交答辩状期间对管辖权提出异议，本院裁定予以驳回。杜满萍不服该裁定，上诉至金华市中级人民法院。后金华市中级人民法院驳回上诉，维持原裁定。2017年4月20日，被告杜满萍申请对借条中“杜满萍”的签名是否系其本人书写进行鉴定，本院依法委托金华天鉴司法鉴定所进行鉴定。本院于2017年7月6日公开开庭进行了审理。原告韦斌姬、被告韦斌强及被告杜满萍的委托代理人贾凌珂到庭参加诉讼。本案现已审理终结。 本院经审理查明：2011年4月25日，被告韦斌强、杜满萍向原告韦斌姬借款10万元，并共同出具了借条一份，内容为：“今向韦斌姬借人民币拾万元正。”庭审中，原告韦斌姬自认被告杜满萍已归还款项2.4万元，其余款项未归还。被告韦斌强陈述其分文未还。被告杜满萍陈述其对案涉借款不知情，故分文未还。在本案审理过程中，被告杜满萍申请对借条中“杜满萍”的签名是否系其本人书写进行鉴定，本院依法委托金华天鉴司法鉴定所进行鉴定，鉴定意见为借条落款处的“杜满萍”的签名字迹与杜满萍样本字迹系同一人书写形成。 本院依照《中华人民共和国合同法》第二百零六条、第二百零七条之规定，判决如下： 一、被告韦斌强、杜满萍于本判决生效之日起十日内归还原告韦斌姬借款7.6万元，并支付利息（自2016年12月1日起按中国人民银行公布同期同档次贷款基准利率计算至实际归还之日止）。 二、驳回原告韦斌姬的其他诉讼请求。 如果被告未按本判决指定的期间履行给付金钱义务，应当依照《中华人民共和国民事诉讼法》第二百五十三条之规定，加倍支付迟延履行期间的债务利息。 案件受理费2300元，减半收取1150元，由原告韦斌姬负担276元，由被告韦斌强、杜满萍负担874元；鉴定费3200元（已由被告杜满萍预交），由被告杜满萍负担。 如不服本判决，可在本判决书送达之日起十五日内向本院递交上诉状，并按对方当事人的人数提出副本，上诉于浙江省金华市中级人民法院。 审判员甘震 二〇一七年七月七日 代书记员许天瑶 ”，

"title": "浙江省东阳市人民法院 民事判决书 （2016）浙0783民初17571号",

"wszzdw": "法院",

"fywszl-wszl": "裁判文书",

"jbfy-cbjg": "浙江省东阳市人民法院",

"wsmc": "判决书",

"ah": "（2016）浙0783民初17571号",

"ajlx-ajlb": "民事一审案件",

"province": "浙江",

"city": "金华市",

"year": "2017",

"reason": [

"民间借贷纠纷"

],

"ajlbsimple": "民事",

```
    "ft": [  
        "中华人民共和国合同法"  
    ]  
}
```

#### `get_index.py` :

用whoosh库建立索引，需要建立索引的字段：`text`、`wszzdw`、`wsmc`、`year`、`reason`、`province`、`ajlbsimple`、`fywszl-wszl`，其中`year`是INT类型，其他都是TEXT类型。`text`和`reason`是检索字段，其他都是筛选字段。在构建索引前会去除停用词。

#### `get_es_index.py` :

在发现whoosh性能过于差之后，又实现了一个elastic search的版本，其他同上。使用时需要开启elastic search的服务。

#### `get_vector_index.py` :

使用faiss库建立稠密向量索引，向量的生成使用Lawformer，具体见 `/search/match.py` 的描述。

#### `get_database.py` :

建立一个sqlite3数据库，用于相似案例推荐的查询。

共两个表：表一储存法条字段和对应的docid，在法条字段建立索引；表二，存储docid到内容相似案例的映射。内容相似案例使用上文提到的LawFormer得到的向量结合faiss库进行查询，调用了 `./search/match.py`。

#### `lawwords.py` :

处理法律词汇，以用于构建字典树，用于查询补全。处理流程如下：先将THUNLP的OCL词库的法律部分作为基础词库，对于助教提供的法律文书XML文件，提取其中各节点的value值：按中文标点和空格分割后，选取列表长度小于等于3的结果，以字典的形式保存下来，字典的值是出现次数，用于补全时的排序。最后得到一个200w大小的法律词汇集合。

#### `utils.py` :

将停用词表读入内存。

#### `stat.py` :

统计XML文件处理后的各个字段，主要用于预处理探索阶段。

## 检索

位于 `./search/` 文件夹下，是搜索、匹配等算法的核心模块。

### `trie_tree.py` :

实现了Trie树，使用上文得到的法律词汇（多了一步操作：将长度大于25的词语，以及包含数字的词语去掉；最后得到130w个法律词汇）来构建Trie树（在每次后端跑起来的时候会构建一次写入内存）。

提供了 `get_recommend_words(word, num)` 函数，传入搜索词汇，在Trie树中获得以该词汇为前缀的所有词汇，按出现次数排序，返回num个。

### `match.py` :

类案检索算法的实现。提供如下几个接口：

`text2vector(text, max_length=4096)`：使用LawFormer模型将文本转换为稠密向量。具体是获得[CLS]对应的last\_hidden\_state作为文本的向量。

`match(text, max_length=500, limit=10, bm25=False, vector_limit=None, bm25_p=0, search_func=None)`：类案检索算法，其中search\_func是外部传入的，可以用whoosh或es实现的search函数。limit表示返回的结果数，bm25表示是否在稠密检索后加入稀疏检索，vector\_limit表示稠密检索返回的结果数，bm25\_p用于限制最后返回的分数阈值，如果bm25\_p = 0表示不采取分数截断。

另外，稠密检索使用的是faiss库提供的搜索算法。

`get_match_ft(ft, num)`：从sqlite3数据库中获得该案件法条匹配类案，随机返回num条。

`get_match_text(docid, num)`：从sqlite3数据库中获得该案件文本匹配类案，随机返回num条。

### `search.py` :

`search(search_dict, limit=None, p=0)`：使用whoosh库的搜索算法实现，评分函数是 **BM25F**。search\_dict传入各个字段的搜索词，p是指分数小于top2\_score \* p的结果不返回。搜索关键词、案由关键词采用 **Or** 策略，年份字段采用 **NumericRange** 策略，而其他字段均是精确匹配。

**es\_search.py**：

`search(search_dict, limit=None, p=0)`：也提供了whoosh算法类似的接口，这也是本代码最终的默认实现。

## 后端

位于 `./flask/` 文件夹下。

使用flask搭建后端，共提供了以下API接口（都需要调用 `./search/` 文件夹下的内容）：

### search

关键词搜索的函数，从前端获得关键词，及（可选的）精细化搜索的条件，返回搜索结果及搜索时间。

### match

类案匹配的函数，从前端获得上传的文件，返回匹配结果及匹配时间。

### show

文章展示页面获取文章内容的函数，从前端获得docid，返回文章内容，包括标签、正文、相似案例的匹配等结果。

### recommend

获得查询词的推荐词的接口，从前端获得查询词，返回推荐补全的词汇。

`config.py` 文件夹用于配置搜索算法，可以选择关键词检索是用whoosh还是elastic search、搜索返回个数（默认为40）、匹配返回个数（默认为40）、案件匹配的算法（可选稠密检索、稀疏检索或二者的组合）。

## 前端

位于 `./bebr2-legal/` 文件夹下

前端使用Vue搭建，主要包括以下模块：

### SearchForm

是一个组件，实现了精细化搜索的选择对话框。

### UploadForm

是一个组件，实现了类案检索上传文件的对话框。

### HomeView

搜索的主页，实现了搜索、结果展示以及上述两个组件的实例化，以及实现了几个与后端接口交互的函数。另外还实现了高亮等选做功能对应的函数，详细见代码注释。

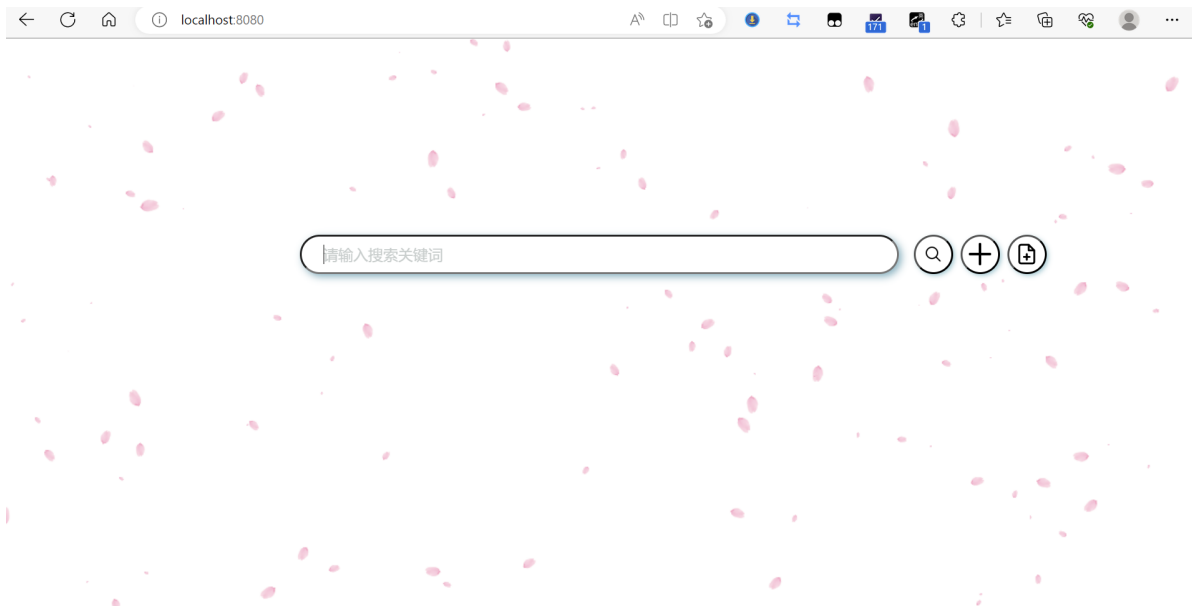
### Article

文章展示的页面，通过docid向后端获取信息并展示。

## # 关键功能

---

首页展示：其他细节见下。



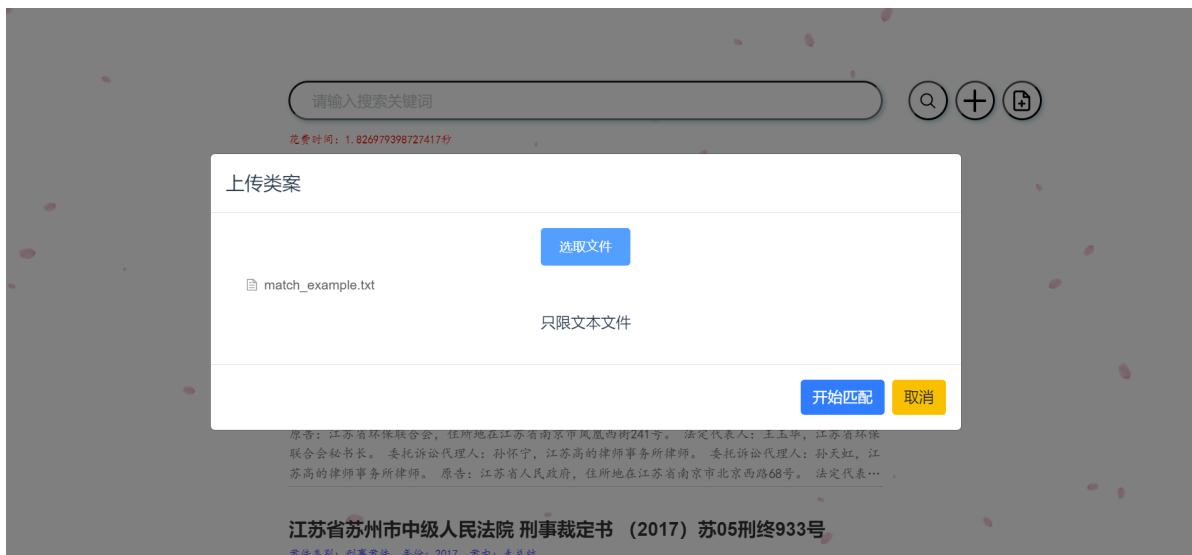
## 关键词检索

用户可以输入关键词进行检索，搜索结果的展示见下文“测试结果”的分析。

关键词检索实现了whoosh的BM25F和elastic search的“practical scoring function”，默认使用更快的elastic search。

## 类案检索

用户可以上传一个文本文件进行类案检索，其中上传的界面如图，搜索结果的展示见下文“测试结果”的分析。



另外，如上文提到，类案检索实现的算法可以在 `config.py` 中进行修改，可以自行选择三种匹配方式：稠密检索、稀疏检索、稠密检索 + 稀疏检索的精排。加上调用不同的稀疏检索方式（whoosh的BM25F和elastic search的“practical scoring function”），最多有五种组合。下文没有特殊说明，默认使用elastic search的稀疏检索，因为最省时。

## 精细化检索

用户可以点击“+”按钮添加精细搜索条件，这些条件从XML的各字段提取，可以作为关键词搜索的限制。

展示如下：



使用上述条件，加上关键词“婚姻家庭纠纷”，搜索结果如下：



婚姻家庭纠纷

花费时间: 0.036281585693359375秒

### 河北省沧州市中级人民法院 民事调解书 (2015) 沧民终字第2209号

案件类别: 民事案件 年份: 2015 案由: 离婚纠纷

上诉人(原审被告): 王某甲, 委托代理人: 宋盛莹, 河北铭鉴律师事务所律师。 被上诉人(原审原告): 王某乙。 委托代理人: 田增明, 沧州市新华区建北法律事务所法律工作者。 上诉人王某甲因离婚纠纷一案, 不服新华区人民法院(2014)新民初字第781号民事判决, 向本院提出上诉, 请求依法...

### 湖南省永兴县人民法院 民事调解书 (2012) 永民初字第647号 (2012) 永民初字第647号

案件类别: 民事案件 年份: 2012 案由: 离婚纠纷

原告: 马雪芬, 女, 1987年11月14日出生, 汉族, 永兴县人, 无业, 住永兴县城关镇内环路。 被告: 李维, 男, 1984年10月23日出生, 汉族, 永兴县人, 无业, 住永兴县城关镇内环路。 原告马雪芬诉被告刘竹环离婚纠纷一案, 本院于立案受理后, 依法由审判员李文斌适用简易程序于2012年7月4日公开开庭...

### 河北省保定市莲池区人民法院 民事调解书 (2016) 冀0606民初991号

案件类别: 民事案件 年份: 2016 案由: 遗嘱继承纠纷

原告王雪丰, 男, 汉族, 1990年10月生, 户籍地天津市武清区, 住保定市莲池区。 被告张春华, 女, 汉族, 1968年3月生, 中国康婷生物工程股份有限公司的全球代理, 住保定市莲池区, 系原告之母。 案由:

可以看到结果确实和婚姻、家庭以及离婚有关(注意“案由”字段是模糊检索, 而不是精确匹配), 时间也被限制为2017年以前, 且都是法院的民事调解书。

## 高亮(关键词飘红)

实现了关键词飘红(注意只飘红搜索框的字段, 精细搜索和类案检索没有进行飘红处理)。

实现时没有简单地按字来飘红, 因为这样可能会有一些不相关的字被飘红, 采用了jieba分词后的结果来飘红, 例如下面的例子, 虽然“中”字出现在搜索词中, 但由于不是一个独立的词, 搜索结果中“中级法院”“延中民再字”等的“中”字没有被飘红。

中国人民保险公司

花费时间: 0.047261714935302734秒

### 福建省闽侯县人民法院 民事判决书 (2017) 闽0121民初1373号

案件类别: 民事案件 年份: 2017 案由: 机动车交通事故责任纠纷

原告: 林佩华, 女, 1958年5月21日出生, 汉族, 住福建省仙游县。 委托诉讼代理人: 郭赛娟, 福建纵然律师事务所律师。 被告: 肖扬弘, 男, 1989年12月7日出生, 汉族, 住福建省福州市鼓楼区。 被告: 中国人民财产保险股份有限公司福州市广达支公司, 住所地福建省福州市台江区五一南路20号...

### 河南省郑州市金水区人民法院 民事判决书 (2014) 金民一初字第1590号

案件类别: 民事案件 年份: 2014 案由: 生命权、健康权、身体权纠纷

原告梁某甲。 委托代理人张育铭, 郑州市二七区淮河路法律服务所法律工作者。 被告河南百事通旅游有限公司, 住所地郑州市金水区。 法定代表人张力, 总经理。 委托代理人葛宗辉。 被告单位员工。 被告中国平安财产保险股份有限公司重庆分公司, 住所地重庆市渝中区。 负责人陈晓牧, 总经理。 ...

### 吉林省延边朝鲜族自治州中级人民法院 民事判决书 (2014) 延中民再字第14号

案件类别: 民事案件 年份: 2014 案由: 劳动合同纠纷

申请再审人(一审原告、二审上诉人): 田志余, 男, 住汪清县汪清镇。 委托代理人: 王淑华(田志余之妻)。 被申请人(一审被告、二审被上诉人): 汪清县庙岭白灰厂。 法定代表人: 魏长海, 该厂厂长。 委托代理人: 王昊, 吉林王昊律师事务所律师。 申请再审人田志余因与被申请人汪清县庙岭...

## 标签显示

标签也是提取于XML文件，展示了案件类型、文书名称、经办机构、年份（以上在标题下展示）、案号、文书制作单位、案由（以上在左侧展示）。用户可以清晰地看到案件的各种属性，便于快速筛选有用信息。



另外，案件类别、年份、案由等标签也在搜索展示页面直接展示，具体可以见上一个部分的展示图。

## 相似案例推荐

相似案例包括“内容相似推荐”和“法条一致案例”进行推荐。内容相似推荐采用 LawFormer 离线计算各篇文章的向量，然后对每篇数据库中的文章进行余弦相似度的计算，获得最相似的8篇案例，在展示该文章时随机抽取三篇展示；法条一致案例则是通过“法条”标签，离线计算各法条对应的文书，然后在展示文章时随机抽取展示。具体展示见上图和下图：

(2014) 常商初字第342号	
法条一致案例	
中华人民共和国侵权责任法	
(2013) 静民初字第5873号	
(2015) 平刑重字第1号	
(2015) 遵民初字第06051号	
最高人民法院关于审理环境民事公益诉讼案件适用法律若干问题的解释	
(2016) 苏01民初1203号	
(2015) 民申字第3380号	
(2015) 民申字第3381号	
关于审理环境民事公益诉讼案件适用法律若干问题的解释	
(2015) 鲁海法立字第00001号	
	诉讼人的诉讼代理人吴小红、张场，被告许建惠、许玉仙的委托代理人宗龙喜、蒋顺到庭参加诉讼。本案现已审理终结。
	公益诉讼人诉称，常州市天宁区人民检察院在履行职责中发现：2010年上半年至2014年9月，许建惠、许玉仙在常州市武进区遥观镇东方村民委员会东方村租用他人厂房，在无营业执照、无危险废物经营许可证情况下，擅自从事废树脂桶和废油桶清洗业务，违法处置清洗废桶过程中产生的废水、废渣，造成了环境严重污染，两被告行为已经构成污染环境罪被依法追究刑事责任。但现场留存的废桶、残渣与污水池里的污水、污泥尚未清除，污水池里的污水仍在不断渗透，对土壤和地下水造成污染。经相关部门检测认定，现场留存的废桶、残渣均属于《国家危险废物名录》中的危险废物，应依法及时处置，消除对环境的危险；经鉴定机构评估调查，污水池渗漏造成了污水池下方土壤的污染，对被污染的土壤应依法进行修复；两被告长期非法处置和排放至少500吨的废水，对地下水 and 外环境造成了严重影响，根据《环境污染损害数额计算推荐方法》的确定原则，可采用虚拟治理成本法来计算修复费用。
	公益诉讼人认为，被告许建惠、许玉仙实施了污染环境的行为，造成了环境的严重污染，损害了社会公共利益，根据《中华人民共和国环境保护法》第六条、《中华人民共和国民事诉讼法》第一百二十四条、《中华人民共和国民事诉讼法》第五十五条、《全国人民代表大会常务委员会关于授权最高人民检察院在部分地区开展公益诉讼试点工作的决定》规定，向本院提起诉讼，请求判令被告许建惠、许玉仙赔偿污染环境修复费用356.2万元，请求判令被告消除危险，对于场地内遗留废物应当及时合法处置，并承担本案的鉴定评估费用。诉讼过程中，公益诉讼人将诉讼请求变更为：1、判令两被告依法及时处置场地内遗留的危险废物，消除危险；2、判令两被告依法及时修复被污染的土壤，恢复原状；3、判令两被告依法赔偿场地排污对环境影响的修复费用，以虚拟治理成本30万元为基数，根据该区域环境敏感程度以4.5-6倍计算赔偿数额，该款项支付至常州市环境公益基金专用账户。
	被告许建惠、许玉仙辩称，一、本案不属于环境公益诉讼，公益诉讼人就本案也不具有诉讼主体资格，其

## 查询补全

查询补全会动态地根据用户输入提示查询词，用户可以方便地点击补全的词直接搜索，具体算法在“实现模块”中描述，不再赘述。

展示如下：

广东

广东

广东省高级人民法院

广东省广州市中级人民法院

广东省深圳市中级人民法院

广东省佛山市中级人民法院

## # 测试结果和样例分析

### 测试关键词

使用关键词“**清华大学玉泉医院**”进行查询，部分结果展示如下：

清华大学玉泉医院

花费时间: 0.06437969207763672秒

Q

+

+

北京市第三中级人民法院 民事判决书 （2016）京03民终2447号

案件类别: 民事案件 年份: 2016 案由: 劳动争议

上诉人（原审原告）袁萍，女，1964年8月25日出生。 被上诉人（原审被告）三星（中国）投资有限公司，住所地北京市朝阳区东三环中路5号楼财富金融中心4-10层01-08单元。 法定代表人裴敬泰，总裁。 委托代理人李桃山，男，1982年2月15日出生。 委托代理人李炫学，男，1988年4月4日出生。 ...

北京市高级人民法院 民事裁定书 （2015）高民申字第03300号

案件类别: 民事案件 年份: 2015 案由: 房屋买卖合同纠纷

再审申请人（一审原告、二审上诉人）：李方，女，1947年10月27日出生。 委托代理人：刘鸿武（李方之夫），男，1943年12月14日出生。 被申请人（一审被告、二审被上诉人）：王艳琴，女，1965年11月22日出生，清华大学玉泉医院医生。 再审申请人李方因与被申请人王艳琴房屋买卖合同纠纷一...

北京市海淀区人民法院 民事判决书 （2017）京0108民初54184号

案件类别: 民事案件 年份: 2017 案由: 生命权、健康权、身体权纠纷

原告：金鑫，男，1988年5月10日出生，汉族，无业，住河北省廊坊市固安县。 被告：新升物业管理有限公司，住所地北京市海淀区。 法定代表人：张苏飞，执行董事。 委托诉讼代理人：李昂，北京市京徽律师事务所律师。 委托诉讼代理人：赵雅琳，女，北京市京徽律师事务所实习律师，住北京市东...

陕西省咸阳市中级人民法院 刑事判决书 （2016）陕04刑初49号

案件类别: 刑事案件 年份: 2016 案由: 未总结

公诉机关咸阳市人民检察院。 被告人张某某，女，汉族，1970年8月30日出生于山东省昌邑县，身份证号码，户籍所在地北京市石景山区石景山路玉泉小区5号楼1层西户（居民身份证住址：北京市石景山区石景山路5号院5号楼4门103号），现住西安市西医大家属院48号楼5层西户，系青海鸿泰慧聚医疗科技...

呼和浩特市玉泉区人民法院 民事判决书 （2015）玉民一初字第00510号

案件类别: 民事案件 年份: 2015 案由: 侵权责任纠纷

原告王某某，男，2004年1月21日出生，汉族，呼和浩特市玉泉区通顺街小学学生，住呼和浩特市玉泉区。 法定代理人王斌（系王某某的父亲），男，1979年9月30日出生，汉族，无固定职业，住呼和浩特市玉泉区西。 委托代理人刘福成，内蒙古信真律师事务所律师。 被告刘某，男，2003年4月27日出...

呼和浩特市人民检察院 起诉书 呼检公一刑诉（2017）3号

案件类别: 刑事案件 年份: 未知 案由: 未总结

被告人张辉，男，1979年\*\*月\*\*日出生，身份证号码为6224251979\*\*/\*\*\*/\*\*，汉族，初中文化，无业，户籍地甘肃省陇西县\*\*镇\*\*村\*\*社\*\*号，捕前暂住呼和浩特市新城区\*\*小区。 2016年12月7日因涉嫌故意伤害罪被呼和浩特市公安局玉泉区分局刑事拘留，同年12月19日经呼和浩...

北京市第一中级人民法院 民事判决书 （2017）京01民终7167号

案件类别: 民事案件 年份: 2017 案由: 生命权、健康权、身体权纠纷

上诉人（原审原告）：苏夫泉，男，1968年10月10日出生，住江苏省苏州市。 上诉人（原审被告）：陆明珍，女，1969年5月10日出生，住江苏省苏州市。 二上诉人共同委托诉讼代理人：王秀全，北京济和律师事务所律师。 被上诉人（原审被告）：清华大学，住所地北京市海淀区清华园。 法定代表人...

对于前十个结果（第一页的结果），标注如下（相关性的判断如下：点进去看全文，如果有涉及清华大学的玉泉医院的内容，判断为相关）：

排名	标题	相关性
1	北京市第三中级人民法院 民事判决书 （2016）京03民终2447号	相关
2	北京市高级人民法院 民事裁定书 （2015）高民申字第03300号	相关
3	北京市海淀区人民法院 民事判决书 （2017）京0108民初54184号	相关
4	陕西省咸阳市中级人民法院 刑事判决书 （2016）陕04刑初49号	相关
5	呼和浩特市玉泉区人民法院 民事判决书 （2015）玉民一初字第00510号	不相关
6	呼和浩特市人民检察院 起诉书 呼检公一刑诉（2017）3号	不相关
7	北京市第一中级人民法院 民事判决书 （2017）京01民终7167号	不相关
8	内蒙古自治区高级人民法院 刑事裁定书 （2016）内刑终108号	不相关
9	呼和浩特市玉泉区人民法院 民事判决书 （2014）玉民一初字第00378号	不相关
10	甘肃省天水市秦州区人民法院 民事判决书 （2011）秦郡民初字第153号	不相关

## 构建Pooling池

前文提到本仓库还实现了其他关键词检索算法，这里采用Whoosh提供的BM25F方法作为第二个搜索算法，以构建Pooling池。检索结果如下，可以看到明显whoosh的速度不如es：

北京市第一中级人民法院 民事判决书 （2017）京01民终7167号

案件类别: 民事案件 年份: 2017 案由: 生命权、健康权、身体权纠纷  
上诉人（原审原告）：苏夫泉，男，1968年10月10日出生，住江苏省苏州市。 上诉人（原审被告）：陆明珍，女，1969年5月10日出生，住江苏省苏州市。 二上诉人共同委托诉讼代理人：王秀全，北京济和律师事务所律师。 被上诉人（原审被告）：清华大学，住所地北京市海淀区清华园。 法定代表人…

北京市第三中级人民法院 民事判决书 （2016）京03民终2447号

案件类别: 民事案件 年份: 2016 案由: 劳动争议  
上诉人（原审原告）袁萍，女，1964年8月25日出生。 被上诉人（原审被告）三星（中国）投资有限公司，住所地北京市朝阳区东三环中路5号楼财富金融中心4-10层01-08单元。 法定代表人裴敬泰，总裁。 委托代理人李桃山，男，1982年2月15日出生。 委托代理人李炫学，男，1988年4月4日出生。 …

北京市高级人民法院 民事裁定书 （2015）高民申字第03300号

案件类别: 民事案件 年份: 2015 案由: 房屋买卖合同纠纷  
再审申请人（一审原告、二审上诉人）：李方，女，1947年10月27日出生。 委托代理人：刘鸿武（李方之夫），男，1943年12月14日出生。 被申请人（一审被告、二审被上诉人）：王艳琴，女，1965年11月22日出生，清华大学玉泉医院医生。 再审申请人李方因与被申请人王艳琴房屋买卖合同纠纷一…

排名	标题	相关性
1	北京市第一中级人民法院 民事判决书 （2017）京01民终7167号	不相关
2	北京市第三中级人民法院 民事判决书 （2016）京03民终2447号	相关
3	北京市高级人民法院 民事裁定书 （2015）高民申字第03300号	相关
4	北京市第二中级人民法院 民事判决书 （2004）二中民初字第7055号	不相关
5	广东省广州市天河区人民法院 民事判决书 （2016）粤0106民初17562号	不相关
6	北京市高级人民法院 民事判决书 （2005）高民终字第923号	不相关
7	北京市海淀区人民法院 民事判决书 （2016）京0108民初33437号	不相关
8	重庆市渝北区人民法院 民事判决书 （2017）渝0112民初20204号	不相关
9	北京市海淀区人民法院 民事判决书 （2017）京0108民初54184号	相关
10	北京市丰台区人民法院 民事判决书 （2016）京0106民初13836号	不相关

两种方法共得到16个案例，其中4个相关案例。

### 评分计算

比较es默认算法（本代码默认使用）和whoosh的bm25f算法：

在查询词为“清华大学玉泉医院”条件下，两种方法评分如下：

搜索方法	AP	P@10	RR	Success@10
es默认算法（本代码默认使用）	1.0	0.4	1.0	1
whoosh的bm25f算法	0.375	0.3	0.5	1

可以看到第二种方法的效果也不如第一种好。

### 测试案例

使用了在“北大法宝”下载的[北京市人民检察院第四分院诉朱清良、朱清涛环境污染民事公益诉讼案](#)进行案件匹配，该案例的链接如下[指导案例206号：北京市人民检察院第四分院诉朱清良、朱清涛环境污染民事公益诉讼案-北大法宝V6官网\(pkulaw.com\)](#)。将其复制为txt文件后进行匹配。

部分结果展示如下：

<p><b>江苏省常州市中级人民法院 民事判决书 （2015）常环公民初字第1号 公益诉讼人江苏省常州市人民检察院，住所地江苏省常州市永宁北路10号。</b></p> <p>案件类别：民事案件 年份：2015 案由：环境污染责任纠纷</p> <p>法定代表人葛志军，江苏省常州市人民检察院检察长。 诉讼代理人吴小红，江苏省常州市人民检察院检察员。 诉讼代理人张扬，江苏省常州市人民检察院助理检察员。 被告许建忠。 被告许玉仙（许建忠之妻）。 两被告共同委托代理人宗龙喜，江苏龙成律师事务所律师。 两被告共同委托代理人蒋顺...</p>
<p><b>江苏省南京市中级人民法院 民事判决书 （2016）苏01民初1203号</b></p> <p>案件类别：民事案件 年份：2016 案由：环境污染责任纠纷</p> <p>原告：江苏省环保联合会，住所地在江苏省南京市凤凰西街241号。 法定代表人：王玉华，江苏省环保联合会秘书长。 委托诉讼代理人：孙怀宁，江苏高的律师事务所律师。 委托诉讼代理人：孙天虹，江苏高的律师事务所律师。 原告：江苏省人民政府，住所地在江苏省南京市北京西路68号。 法定代表...</p>
<p><b>江苏省苏州市中级人民法院 刑事裁定书 （2017）苏05刑终933号</b></p> <p>案件类别：刑事案件 年份：2017 案由：未总结</p> <p>原公诉机关苏州市姑苏区人民检察院。 上诉人（原审被告）王菊明，男，1963年9月25日出生，汉族，小学文化，系昆山锦鹿建筑工程有限公司负责人，户籍地江苏省昆山市）於家湾98号，住江苏省昆山市。因涉嫌污染环境罪，于2016年7月6日被刑事拘留（同年7月5日被羁押），同年8月11日被逮...</p>
<p><b>贵州省仁怀市人民法院 行政判决书 （2017）黔0382行初4号 公益诉讼人纳雍县人民检察院，住所地：贵州省毕节市纳雍县新城区政法园区。</b></p> <p>案件类别：行政案件 年份：2017 案由：未总结</p> <p>法定代表人郑雄 该院检察长。 委托诉讼代理人周志雄 该院副检察长。 委托诉讼代理人李黔 该院...</p>

**浙江省湖州市中级人民法院 行政赔偿判决书 （2014）浙湖行赔初字第5号**

案件类别: 行政案件 年份: 2014 案由: 未总结

原告浙江永立石材综合开发有限公司,住所地安吉县递铺镇天荒坪南路48号。法定代表人许群显,该公司董事长。委托代理人陆云良,浙江浙联律师事务所律师。被告安吉县人民政府,住所地安吉县递铺镇行政大楼。法定代表人沈铭权,该县县长。委托代理人黄立科,浙江振源律师事务所律师。 …

**安徽省芜湖市镜湖区人民法院 行政判决书 （2016）皖0202行初66号 公益诉讼人芜湖市镜湖区人民检察院，住所地安徽省芜湖市镜湖区银湖南路\* \*号。**

案件类别: 行政案件 年份: 2016 案由: 诉不履行法定职责

被告无为县国土资源局,住所地安徽省无为县无城镇二坝路2号,统一社会信用代码11340217003274330E。法定代表人仇后亮,局长。出庭负责人耿治明,副局长。委托代理人朱长华,安徽全人律师事务所律师。委托代理人莫宏明,安徽铭盛律师事务所律师。公益诉讼人芜湖市…

**贵州省遵义市中级人民法院 行政判决书 （2017）黔03行终291号**

案件类别: 行政案件 年份: 2017 案由: 诉不履行法定职责

上诉人(一审第三人): 贵州铜仁紫玉旅游工艺品有限公司。住所地: 贵州省铜仁市碧江区共青团路2号,统一社会信用代码: 91520600741143340N。法定代表人: 田仁贵,公司执行董事。委托诉讼代理人: 卓宝松,公司经理。委托诉讼代理人: 邓建国,北京子悦律师事务所律师。被上诉人(一审公…

**西安铁路运输法院 行政判决书 （2017）陕7102行初60号 公益诉讼人西**

对于前十个结果（第一页的结果），标注如下（相关性的判断如下：由于该案例涉及“生态环境功能损失赔偿”“环境污染与修复”，所以正文涉及其中某一类的案例标注为相关）：



排名	标题	相关性
1	江苏省常州市中级人民法院 民事判决书（2015）常环公民初字第1号 公益诉讼人江苏省常州市人民检察院，住所地江苏省常州市永宁北路10号。	相关
2	江苏省南京市中级人民法院 民事判决书（2016）苏01民初1203号	相关
3	江苏省苏州市中级人民法院 刑事裁定书（2017）苏05刑终933号	相关
4	贵州省仁怀市人民法院 行政判决书（2017）黔0382行初4号 公益诉讼人纳雍县人民检察院，住所地：贵州省毕节市纳雍县新城区政法园区。	相关
5	浙江省湖州市中级人民法院 行政赔偿判决书（2014）浙湖行赔初字第5号	相关
6	安徽省芜湖市镜湖区人民法院 行政判决书（2016）皖0202行初66号 公益诉讼人芜湖市镜湖区人民检察院，住所地安徽省芜湖市镜湖区银湖南路**号。	相关
7	贵州省遵义市中级人民法院 行政判决书（2017）黔03行终291号	相关
8	西安铁路运输法院 行政判决书（2017）陕7102行初60号 公益诉讼人西安市新城区人民检察院。	相关
9	广东省珠海横琴新区人民法院 刑事判决书（2017）粤0491刑初43号	相关
10	云南省昆明市中级人民法院 民事判决书（2012）昆环保民初字第6号 公益诉讼人云南省宜良县国土资源局	相关

## 构建Pooling池

前文提到本仓库还实现了其他类案检索算法，这里采用LawFormer向量检索的方法作为第二个搜索算法，以构建Pooling池。LawFormer的匹配结果如下：

排名	标题	相关性
1	中华人民共和国最高人民法院 民事裁定书 （2015）民申字第3377号	相关
2	中华人民共和国最高人民法院 民事裁定书 （2015）民申字第3381号	相关
3	中华人民共和国最高人民法院 民事裁定书 （2015）民申字第3381号	相关
4	北京市第一中级人民法院 行政判决书 （2018）京01行初24号	不相关
5	杭州市拱墅区人民法院 民事裁定书 （2017）浙0105民初8114号	不相关
6	中华人民共和国最高人民法院 民事裁定书 （2012）民监字第539号	不相关
7	中华人民共和国最高人民法院 民事裁定书 （2017）最高法民再121号	不相关
8	海南省高级人民法院 民事裁定书 （2013）琼立一终字第155号	不相关
9	北京市朝阳区人民法院 民事判决书 （2016）京0105民初71177号	不相关
10	北京市第一中级人民法院 行政判决书 （2017）京01行初1134号	不相关

两种方法共得到20个案例，其中13个相关案例。

## 评分计算

搜索方法	AP	P@10	RR	Success@10
es默认算法（本代码默认使用）	1.0	1.0	1.0	1
lawformer向量检索	0.3	0.3	1.0	1

可以看到未微调过的Lawformer生成的向量直接检索，效果是比较差的。

## # 开源资料的使用

---

Elastic Search搜索引擎及其python包

Whoosh搜索引擎

Lawformer语言模型: [thunlp/Lawformer · Hugging Face](#)

faiss稠密向量检索

THUNLP提供的法律词汇: [THUOCL: 清华大学开放中文词库-法律 \(thunlp.org\)](#)

前端一些组件使用了bootstrap等库

## # 参考他人代码

---

后端没有参考他人代码。

前端的樱花飘落背景使用了他人代码: [使用纯JavaScript实现全网页动态樱花飘落特效 樱花飘落js 繁华尽头满是殇的博客-CSDN博客](#)