司法大数据

——自动爬取和分词系统

Team: workHard

小组成员

Team: WorkHard

张继华

201250040@smail.nju.edu.cn

邓尤亮

201250035@smail.nju.edu.cn

江　楠

201250033@smail.nju.edu.cn

目 录

CONTENTS

第一部分

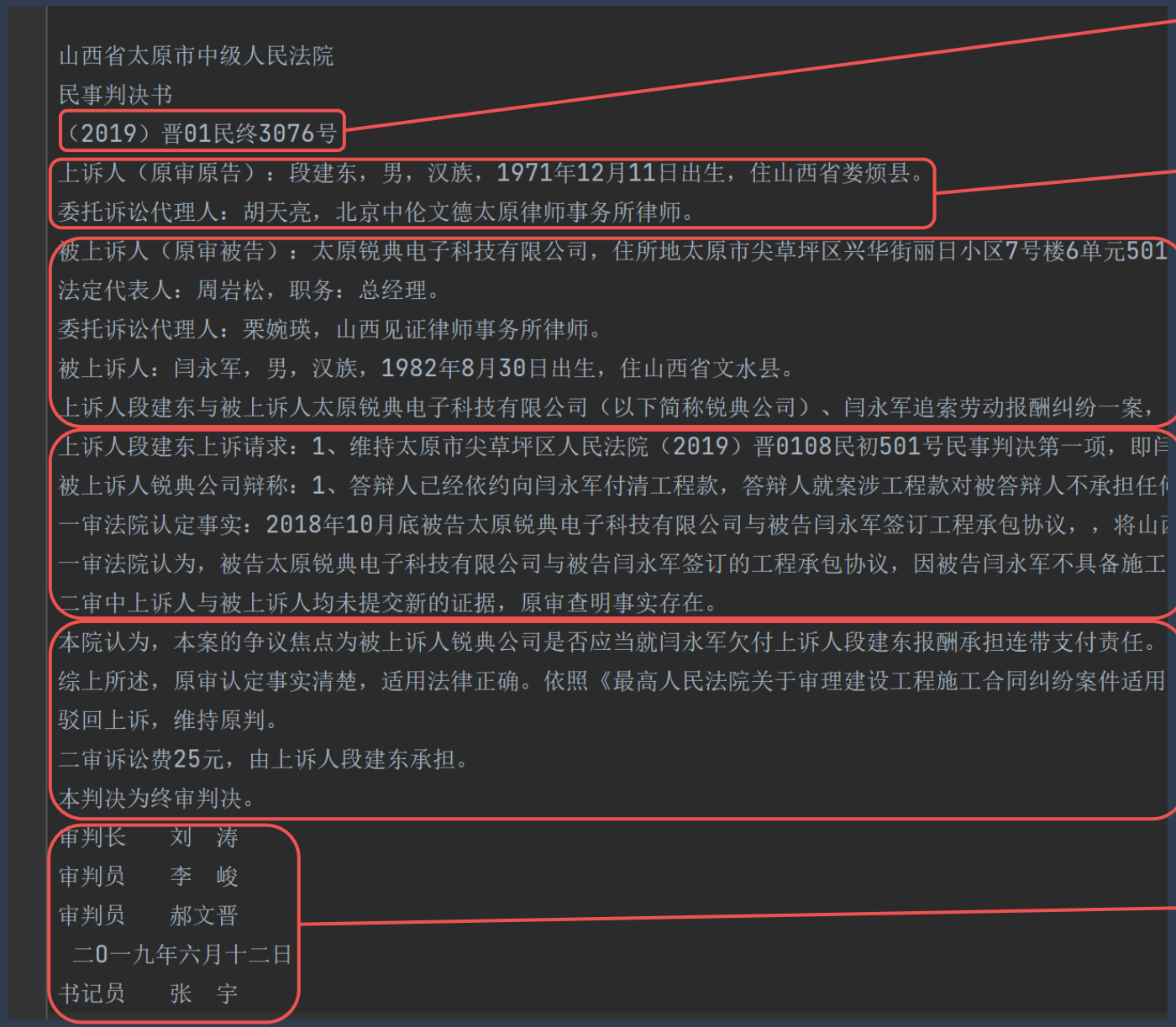
实现意义

实现意义

大数据时代的当下，信息变得越来越重要。不仅是信息本身，对信息的分类、整理、提取、总结等操作的重要性也不言而喻。中国裁判文书网给了我们大量的司法实践的结果，但也缺少对文本的萃取和注解。

借助信息技术，我们可以快速地标记法律文书中的关键信息，生成对应的标记文本；借助标记文本，我们可以实现对大量文书统筹归纳，快速统计等功能；对大量数据的学习，甚至可以实现处罚力度、刑期长短等信息的预测和复核，帮助实施审判和考察法官审判水平（并非判例法模式，有利与审判中对于情节轻重的判断）。可见法律文书的自动化标记拥有者重要的意义和作用。

实现意义



山西省太原市中级人民法院
民事判决书
（2019）晋01民终3076号
上诉人（原审原告）：段建东，男，汉族，1971年12月11日出生，住山西省娄烦县。
委托诉讼代理人：胡天亮，北京中伦文德太原律师事务所律师。
被上诉人（原审被告）：太原锐典电子科技有限公司，住所地太原市尖草坪区兴华街丽日小区7号楼6单元501
法定代表人：周岩松，职务：总经理。
委托诉讼代理人：栗婉瑛，山西见证律师事务所律师。
被上诉人：闫永军，男，汉族，1982年8月30日出生，住山西省文水县。
上诉人段建东与被上诉人太原锐典电子科技有限公司（以下简称锐典公司）、闫永军追索劳动报酬纠纷一案，
上诉人段建东上诉请求：1、维持太原市尖草坪区人民法院（2019）晋0108民初501号民事判决第一项，即
被上诉人锐典公司辩称：1、答辩人已经依约向闫永军付清工程款，答辩人就案涉工程款对被答辩人不承担任
一审法院认定事实：2018年10月底被告太原锐典电子科技有限公司与被告闫永军签订工程承包协议，，将山
一审法院认为，被告太原锐典电子科技有限公司与被告闫永军签订的工程承包协议，因被告闫永军不具备施工
二审中上诉人与被上诉人均未提交新的证据，原审查明事实存在。
本院认为，本案的争议焦点为被上诉人锐典公司是否应当就闫永军欠付上诉人段建东报酬承担连带支付责任。
综上所述，原审认定事实清楚，适用法律正确。依照《最高人民法院关于审理建设工程施工合同纠纷案件适用
驳回上诉，维持原判。
二审诉讼费25元，由上诉人段建东承担。
本判决为终审判决。
审判长　刘　涛
审判员　李　峻
审判员　郝文晋
二0一九年六月十二日
书记员　张　宇

文书编号

原告信息

被告信息

文书细节

最终判决

其他信息

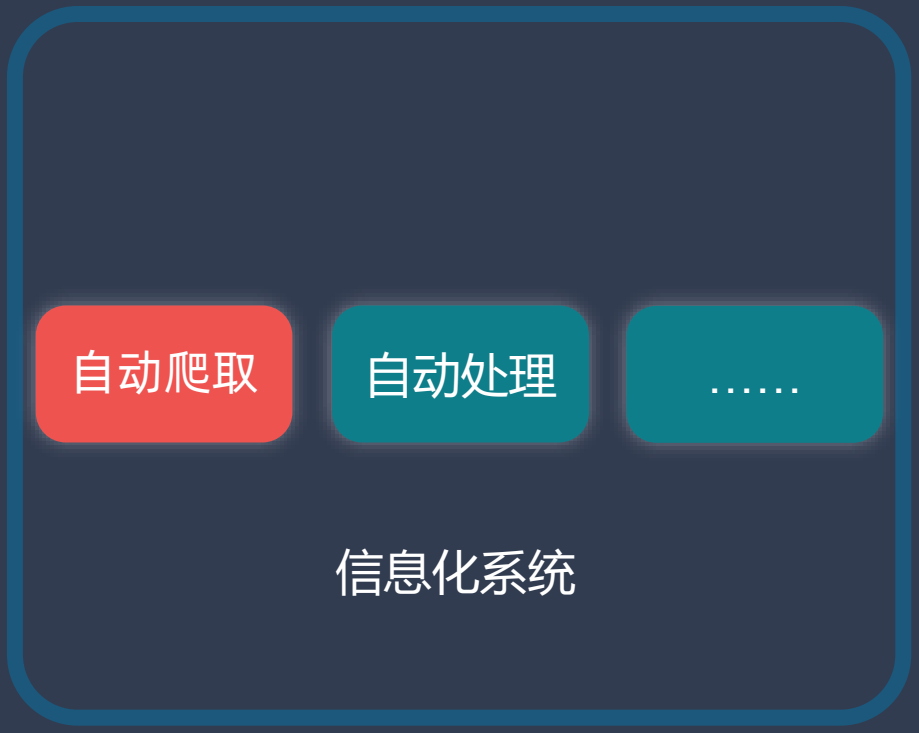
Information & Data

实现意义

司法领域中在处理文档内容
时会遇到各种**难题**

实现意义

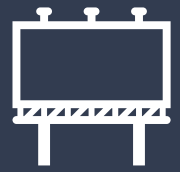
法律文书



标记文档

机器学习

审判参考



信息公开

…

第二部分

实现目标

实现目标

收集海量原始文书数据，对其进行关键信息的半自动化标注。包括各个当事人信息、案由和法院的信息提取。

为用户提供美观易用的前端交互界面。用户可以输入文书信息，或使用实时爬虫系统爬取文书。原始数据经预处理后返回关键词信息，以供进一步手动标注。

后端对文书内容进行爬取；对文书进行分词与预处理并辅助生成 JSON 文件。

使用框架构建前后端交互的一整套系统，并尝试部署在云服务器上，方便用户快速地访问与使用。

第三部分

具体细节

具体细节

JSON 标注

标注基本的当事人信息、案由和法院。

我们希望不局限于刑事案件，而是扩大范围，对于任意的文书都能够做到信息的标注。在其他文书中可能会涉及到**多个当事人**，以及存在**自然人**和**法人**的区别。为此，我们将对其进行分类，并获取不同的信息以供标注。例如，对于自然人，标注其姓名、性别、出生日期和民族；对于法人则标注其名称。

对于具体的刑事案件，由于刑事文书格式较为工整，后端可以较为准确地对当事人和案由进行直接标注，从而可以简化用户手动标注的过程。

具体细节

项目框架

预计使用 **Spring boot + Vue** 作为前后端的整体框架选择。

由于本项目的数据格式都为纯文本与 json 标记，所以数据库为可选系统。

具体细节

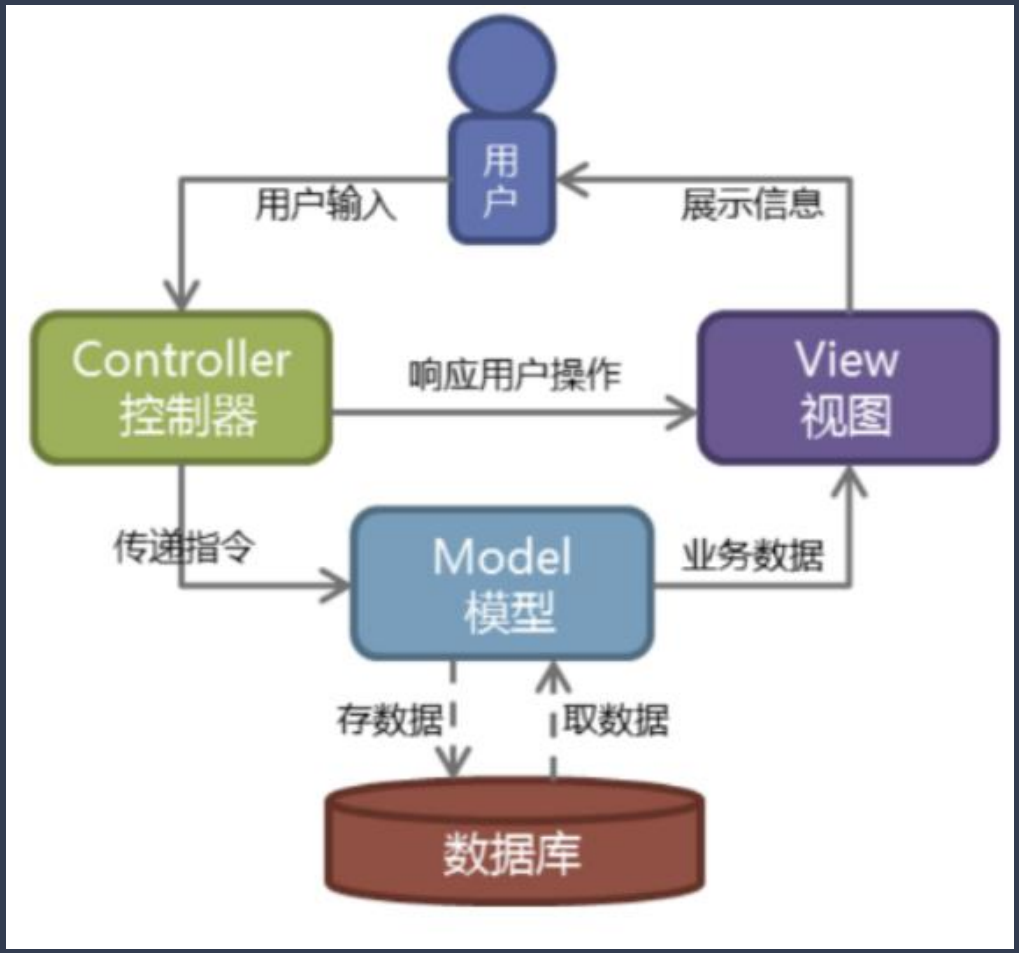
前端
后端

用户提供已经下载的文书(以 .txt 格式)将文书发送到前端，前端按照用户需求自动化标记，并生成对应的 .json 文件；用户在前端输入自己需要检索的文书的条件(时间、案件类型、地区、当事人姓名...)，系统根据用户输入的条件，自动化地从文书网上爬取若干份符合条件的文书，按照用户的要求标记文本，将文书(以 .txt 格式) 和标记文本(以 .json 格式) 打包下载保存。

后端使用 Java 开发，使用 Selenium 进行文书内容爬取；使用 HanLP 对文书进行分词与预处理；使用 FastJson 辅助完成 JSON 文件的生成。

具体细节

总体结构


用户
用户输入
展示信息
Controller
控制器
响应用户操作
View
视图
传递指令
Model
模型
业务数据
存数据
取数据
数据库

代码托管

代码管理和多人合作平台：

GitLab

第四部分

当前进度

已经实现

使用 Selenium 模拟网页行为，爬取文书

使用 HanLP 对文书的日期、地点、组织和人名进行单独提取；对动词和形容词进行提取。

使用 FastJson 辅助实现 JSON 标注

已经实现

模拟
爬虫

README.md | Crawler.java | 施丽琴、缙云县人民政府五云街道办事处行政监察(监察)二审行政裁定书 | FluentWait.class | Marker.java

```java
    }

    private boolean login(String account, String pswd, String preIndex) {
        WebElement loginButton = driver.findElement(By.id("loginLi"));
        if (!loginButton.getText().contains("登录")) return true;

        try {
            Thread.sleep( millis: 1000);
        } catch (InterruptedException e) {
            e.printStackTrace();
        }

        loginButton.click();

        while (driver.findElements(By.cssSelector("iframe")).size() == 0) {
            try {
                Thread.sleep( millis: 1000);
                driver.navigate().refresh();
            } catch (InterruptedException e) {
                e.printStackTrace();
            }
        }
        // 输入框被一个 iframe 包裹，先切换
        driver.switchTo().frame( nameOrId: "contentIframe");
        // 填写用户名和密码
        WebElement login = new WebDriverWait(driver, Duration.ofSeconds(WAIT_SECONDS))
                .until(driver -> driver.findElement(By.cssSelector(".phone-number-input")));
        login.sendKeys(account);
        driver.findElement(By.cssSelector(".password")).sendKeys(pswd);
```

README.md | Crawler.java | 施丽琴、缙云县人民政府五云街道办事处行政监察(监察)二审行政裁定书 | FluentWait.class | Marker.java

行政监察（监察）
浙江省丽水市中级人民法院
行 政 裁 定 书
（2020）浙11行终122号
上诉人（原审原告）施丽琴，女，1973年6月2日出生，汉族，住缙云县。
委托代理人成孟，男，1967年12月15日出生，汉族，住缙云县，系原告配偶。
被上诉人（原审被告）缙云县人民政府五云街道办事处。住所地缙云县五云街道复兴街154号。统一社会信用代码11332526002660694N。
法定代表人张伟峰，主任。
应诉负责人陈伟秋，办事处副书记。
委托代理人蒋土福，浙江民辉律师事务所律师。
被上诉人（原审被告）缙云县人民政府。住所地缙云县五云街道普龙路38号。统一社会信用代码：113325266671424888。
法定代表人王正飞，县长。
委托代理人施丽昌。
委托代理人樊理洪，浙江博翔（缙云）律师事务所律师。
上诉人施丽琴因与缙云县人民政府五云街道办事处（以下简称五云街道办事处）、缙云县人民政府其他行政管理行政复议一案，不服莲都区人民法院（2020）浙1102行初43号行政裁定，向本院提起上诉。本院依法组成合议庭审理了本案，现已审理终结。
原审法院经审理认定，成孟系原告施丽琴丈夫。2019年1月28日，经五云街道人民调解委员会调解，签订了以成孟为甲方，林岩富、周岩松、林维彬、林维光为乙方的《人民调解协议书》，约定成孟一次性补偿生产队用地费用2700元，生产队将200平米左右的土地交
原审法院认为，《中华人民共和国农村土地承包法》第十三条第一款规定，农村集体所有的土地依法属于村农民集体所有的，由村集体经济组织或者村民委员会发包，已经分别属于村内两个以上农村集体经济组织的农民集体所有的，由村内各该农村集体经济组织或者村
上诉人施丽琴上诉称，上诉人在（2019）浙1102行初120号的诉讼请求是要求履行，而本案是要求督促履行，诉讼请求有质的区别。且被诉行政机关不能委托下属单位参加诉讼，上诉人在一审时当庭不认可，但一审法院没有采纳。综上，请求撤销一审裁定，并重新审
被上诉人五云街道办事处答辩称，上诉人先后以同一事实理由和同一诉请申请行政复议并提起行政诉讼，明显不符合行政复议及行政诉讼的法定受理条件。从上诉人先后提起的两次行政诉讼起诉状可知，其诉讼请求均是以2019年1月28日调解协议为依据，要求五云街
被上诉人缙云县人民政府答辩称，被答辩人属于重复起诉事实清楚，前后两诉的诉请没有本质区别。答辩人一审出庭人员符合法律规定，缙云县司法局行政复议科具体负责缙云县人民政府的行政复议工作，一审出庭工作人员系缙云县司法局行政复议科科员，出庭资格符
本院经审理查明的事实与一审认定的事实一致。
本院认为，本案的争议焦点是上诉人提起本诉是否属于重复起诉。上诉人以五云街道办事处未履行责任田分包到户的法定职责为由，曾提起诉讼，莲都区人民法院作出（2019）浙1102行初120号行政裁定驳回其起诉，该裁定已生效。本诉被告虽包括缙云县人民政府，
驳回上诉，维持原裁定。
本裁定为终审裁定。
审 判 长 吴林雄
审 判 员 邹一峻
审 判 员 吴金花
二〇二〇年十一月五日

已经实现

HanLP
分词

```java
@Test
public void test1() {
    try {
        participle.process( text: "委托合同纠纷\n" +
                "中华人民共和国最高人民法院\n" +
                "民 事 裁 定 书\n" +
                "（2018）最高法民申3249号\n" +
                "再审申请人（一审原告、二审上诉人）：张军，男，汉族，1965年10月8日出生，住安徽省阜阳市颍州区。\n" +
                "委托诉讼代理人：胡泽晶，河南亚太人律师事务所律师。\n" +
                "委托诉讼代理人：牛晓婷，河南亚太人律师事务所律师。\n" +
                "被申请人（一审被告、二审被上诉人）：阜阳安厦建设（集团）有限公司。住所地：安徽省阜阳市颍州区顾西镇临泉路411号。\n" +
                "法定代表人：申志文，该公司总经理。\n" +
                "委托诉讼代理人：冯继辉，安徽蓝邦律师事务所律师。\n" +
                "委托诉讼代理人：耿建生，安徽承义律师事务所律师。\n" +
                "再审申请人张军因与被申请人阜阳安厦建设（集团）有限公司（以下简称安厦公司）委托合同纠纷一案，不服安徽省高级人民法院（2017）皖民终664号民事判决，向本院申请再审。本院依法组成合议庭对本案
                "张军申请再审称，（一）原审判决认定事实错误。1.张军依协议约定完成了自己的委托义务。张军向安厦公司提供了安徽继华置业有限公司（以下简称继华公司）另外享有69%股权的证据，该份股权名义上由合
                "安厦公司提交书面意见称，张军的再审申请不能成立，应予驳回。（一）关于双方签订的《债权转让协议》，名为债权转让，实为委托协议，按照合同法规定，随时成立也可以随时解除，况且协议中也约定安厦
                "本院经审查认为，根据张军的再审请求及理由与安厦公司的答辩意见，本案的争议焦点在于张军主张安厦公司取得的保利汉铭公司16%股权中的5.76%应归其所有的请求权基础是否存在。\n" +
                "张军主张其与安厦公司签订了《债权转让协议》与《协议书》等协议，双方已经就张军委托安厦公司代为主张张军对张继华、继华公司享有的1800万元债权达成一致意见。因此，对本案争议焦点的审查，必须以
                "其一，张军主张其已经按照与安厦公司约定全面履行自己合同义务的依据不足。虽然张军于2014年6月3日和安厦公司签订《债权转让协议》，约定将张军对继华公司、张继华享有的全部债权转让给安厦公司，但
                "其二，安厦公司取得保利汉铭公司16%股权有事实依据。安厦公司于2014年11月6日向张继华、继华公司账户支付保证金3200万元，各方对此均无异议。根据安厦公司与张继华、继华公司签订的《工程总承包协
                "其三，张军主张安厦公司取得的16%股权中包含其1800万元债权的事实依据不足。1.从2014年6月25日《协议》内容来看，仅涉及张继华、继华公司欠安厦公司、张中良款项的问题，未提及所欠张军款项。对此
                "因此，张军主张安厦公司取得的保利汉铭公司16%股权中的5.76%应归其所有的请求缺乏事实和法律依据，原审法院对其诉讼请求未予支持并无不当。至于张军与继华公司、张继华之间的债权债务纠纷可另行处理
                "综上，张军的再审申请不符合《中华人民共和国民事诉讼法》第二百条第二项、第六项规定的情形。依照《中华人民共和国民事诉讼法》第二百零四条第一款，《最高人民法院关于适用<中华人民共和国民事诉讼
                "驳回张军的再审申请。\n" +
                "审判长刘京川\n" +
                "审判员杨立初\n" +
                "审判员刘慧卓\n" +
                "二〇一八年九月十七日\n" +
                "法官助理王戈\n" +
                "书记员叶和申");
    } catch (IOException e) {
        e.printStackTrace();
    }
    System.out.println(participle.getDateSet().toString());
    System.out.println(participle.getLocationSet());
    System.out.println(participle.getOrgSet());
    System.out.println(participle.getPersonSet());
    System.out.println(participle.getVerbSet());
    System.out.println(participle.getAdjSet());
}
```

```
"C:\Program Files\Java\jdk1.8.0_291\bin\java.exe" ...
[1965年10月8日, 2017, 2014年6月3日, 2014年6月11日, 2014年7月3日, 2014年11月6日, 2014年6月25日, 2014年, 2016年, 2015年1月, 当日, 二〇一八年九月十七日]
[安徽省阜阳市颍州区, 安徽省阜阳市颍州区顾西镇临泉路411号, 皖, 安徽, 中华人民共和国]
[中华人民共和国最高人民法院, 最高, 河南亚太人律师事务所, 阜阳安厦建设（集团）有限公司, 安徽蓝邦律师事务所, 安徽承义律师事务所, 安厦公司, 安徽省高级人民法院, 安徽继华置业有限公司, 继华公司, 合肥三易投资管理有
[张军, 胡泽晶, 牛晓婷, 申志文, 冯继辉, 耿建生, 张继华, 张中良, 刘京川, 杨立初, 刘慧卓, 王戈, 叶和申]
[申, 出生, 住, 简称, 委托, 不服, 申请, 再审, 组成, 进行, 审查, 终结, 称, 认定, 完成, 提供, 享有, 持有, 是, 出面, 办理, 缴纳, 约定, 有, 协商, 利用, 应, 视为, 能, 违反, 提交, 收到, 交, 无, 获得, 履行
[最高, 有限, 高级, 充分, 错误, 一样, 一致, 真实, 不足, 合理, 矛盾, 相同, 实际, 不当]

Process finished with exit code 0
```

已经实现

JSON 标注

```java
@Test
public void test() {
    ArrayList<SubjectInfo> subjectInfos = new ArrayList<>();
    ArrayList<String> courts = new ArrayList<>();

    SubjectInfo subjectInfo1 = new SubjectInfo( name: "大连红枫房地产发展有限公司",  partiesType: "被执行人",  isNatual: false);
    subjectInfos.add(subjectInfo1);
    SubjectInfo subjectInfo2 = null;
        subjectInfo2 = new SubjectInfo( name: "吴丽红",
                partiesType: "复议申请人（案外人）",
                isNatual: true,
                gender: "女",
                birthPlace: null,
                birthDate: "1973年2月25日",
                ethnicity: "汉族");
    subjectInfos.add(subjectInfo2);
    courts.add("辽宁省高级人民法院");
    Marker marker = new Marker(subjectInfos,  accusation: "企业借贷纠纷", courts);
    System.out.println(marker.toJson());

}
```

```json
{
  "主体": [
    {
      "名称": "大连红枫房地产发展有限公司",
      "身份": "被执行人",
      "自然人": false
    },
    {
      "名称": "吴丽红",
      "身份": "复议申请人（案外人）",
      "自然人": true,
      "性别": "女",
      "出生日期": "1973年2月25日",
      "民族": "汉族"
    }
  ],
  "案由": "企业借贷纠纷",
  "法院": [
    "辽宁省高级人民法院"
  ]
}
```

尚未实现

爬取部分需要丰富检索条件

分词部分需要添加特殊情况的处理。
例如部分案件的人物信息较为集中且规范，可以特殊处理，提高精准度。

前端页面的构建

前后端框架和服务器的搭建、前后端交互

云服务器部署

谢谢