数据科学大作业汇报初稿

小组成员：
张继华 201250040
邓尤亮 201250035
江楠 201250033

理解

司法领域中在处理文档内容时会遇到各种问题，如：文档录入繁琐，文书检索困难，文书阅读效率低，文书内容不便归纳统计等。

借助信息技术，我们可以快速的标记法律文书中的关键信息，生成对应的标记文本这样，我们可以在短时间内迅速了解文书中的关键信息；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）。可见对于法律文书的自动化标记的重要意义和作用。

利用信息技术，实现智慧法院也是最高法指导各级法院建设的重要内容；在信息技术的加持下，智慧法院可以实现，网上办理、信息、流程公开，智能服务等特性，提高法院办事效率，提升法院服务水平；

基本思路

我们的目标：

1. 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**法院信息**的提取。
2. 为用户提供美观且易用的前端交互界面。用户可以选择输入文书信息，或者使用我们的实时爬虫系统爬取文书。原始文书数据经过后端程序预处理后返回分类别的关键词信息，以供用户进一步手动进行精确标注。
3. 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完成 JSON 文件的生成。
4. 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。

实现细节

JSON 信息标注格式说明

1. 基本的当事人信息、案由和法院。
2. 对于当事人信息，我们希望不局限于刑事案件，而是将范围扩大化，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。
3. 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，从而简化了用户手动标注的过程。

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，故不考虑数据库。

项目管理

使用 GitLab 进行代码管理和多人合作。

后端

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。

前端

前端提供两种文书获取方式：

1. 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；
2. 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。

工作进度

完成的内容

- 使用 Selenium 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。

```
@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
```

- 使用 FastJson 辅助实现 JSON 标注

```
@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());

}
```

```
1 {
```

```
  "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"
  ]
}
```


未完成的内容

- 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
- 云服务器部署项目