南京大学
NANJING UNIVERSITY

数据科学大作业——隐私信息扫描

目录

CONTENTS

整体背景

PART ONE

数据滥用

Facebook-剑桥分析数据丑闻是指英国咨询公司剑桥分析公司在未经Facebook用户同意的情况下获取数百万Facebook用户的个人数据，这些数据主要用于政治广告。该应用获取了多达8700万份Facebook个人用户资料。剑桥分析公司获得这些数据后对此展开分析并根据分析的结果为2016年泰德·克鲁兹和唐纳德·特朗普的总统竞选活动提供帮助。

滴滴

滴滴一下 美好出行

“滴滴出行” App存在严重违法违规收集使用个人信息问题。国家互联网信息办公室依据《中华人民共和国网络安全法》相关规定，通知应用商店下架“滴滴出行” App，要求滴滴出行科技有限公司严格按照法律要求，参照国家有关标准，认真整改存在的问题，切实保障广大用户个人信息安全。

政策法规

欧盟： 2016 年 4 月 27 日出版《通用数据保护条例》(General Data Protection Regulation)，是在欧盟法律中对所有欧盟个人关于数据保护和隐私的规范，涉及了欧洲境外的个人数据出口。GDPR 主要目标为取回个人对于个人数据的控制，以及为了国际商务而简化在欧盟内的统一规范， 2018年5月25日正式生效。

美国：《加州消费者隐私法》（ California Consumer Privacy Act ）是一项旨在加强美国加州居民的隐私权和消费者保护的州级法规。该法案由加利福尼亚州议会通过，并于2018年6月28日由加州州长杰里-布朗签署成为法律，以修订《加州民法典》第3分部第4部分。2019年10月，美国加州州长正式签署五份对CCPA的修正案， CCPA终于完成了漫长而艰难的修正之旅。

中国：2021年8月20日，十三届全国人大常委会第三十次会议表决通过《中华人民共和国个人信息保护法》。个人信息保护法自2021年11月1日起施行。其中明确：①通过自动化决策方式向个人进行信息推送、商业营销，应提供不针对其个人特征的选项或提供便捷的拒绝方式②处理生物识别、医疗健康、金融账户、行踪轨迹等敏感个人信息，应取得个人的单独同意③对违法处理个人信息的应用程序，责令暂停或者终止提供服务。

研究内容

PART TWO

相关条款

第四条　个人信息是以电子或者其他方式记录的与已识别或者可识别的自然人有关的各种信息，不包括匿名化处理后的信息。个人信息的处理包括个人信息的收集、存储、使用、加工、传输、提供、公开、删除等。

第六条　处理个人信息应当具有明确、合理的目的，并应当与处理目的直接相关，采取对个人权益影响最小的方式。收集个人信息，应当限于实现处理目的的最小范围，不得过度收集个人信息。

第九条　个人信息处理者应当对其个人信息处理活动负责，并采取必要措施保障所处理的个人信息的安全。

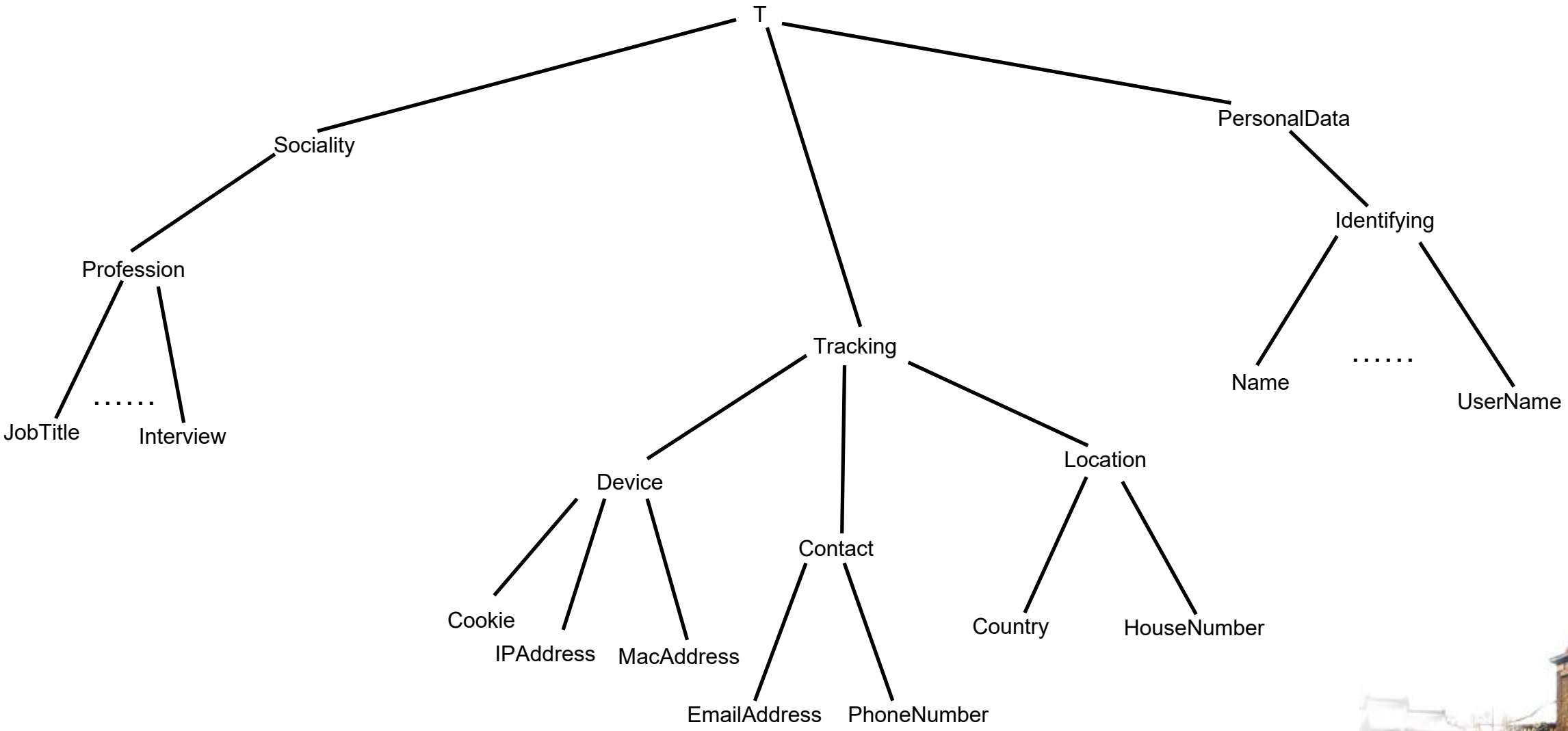
内容说明

第四条　个人信息是以电子或者其他方式记录的与已识别或者可识别的自然人有关的各种信息，不包括匿名化处理后的信息。个人信息的处理包括个人信息的收集、存储、使用、加工、传输、提供、公开、删除等。

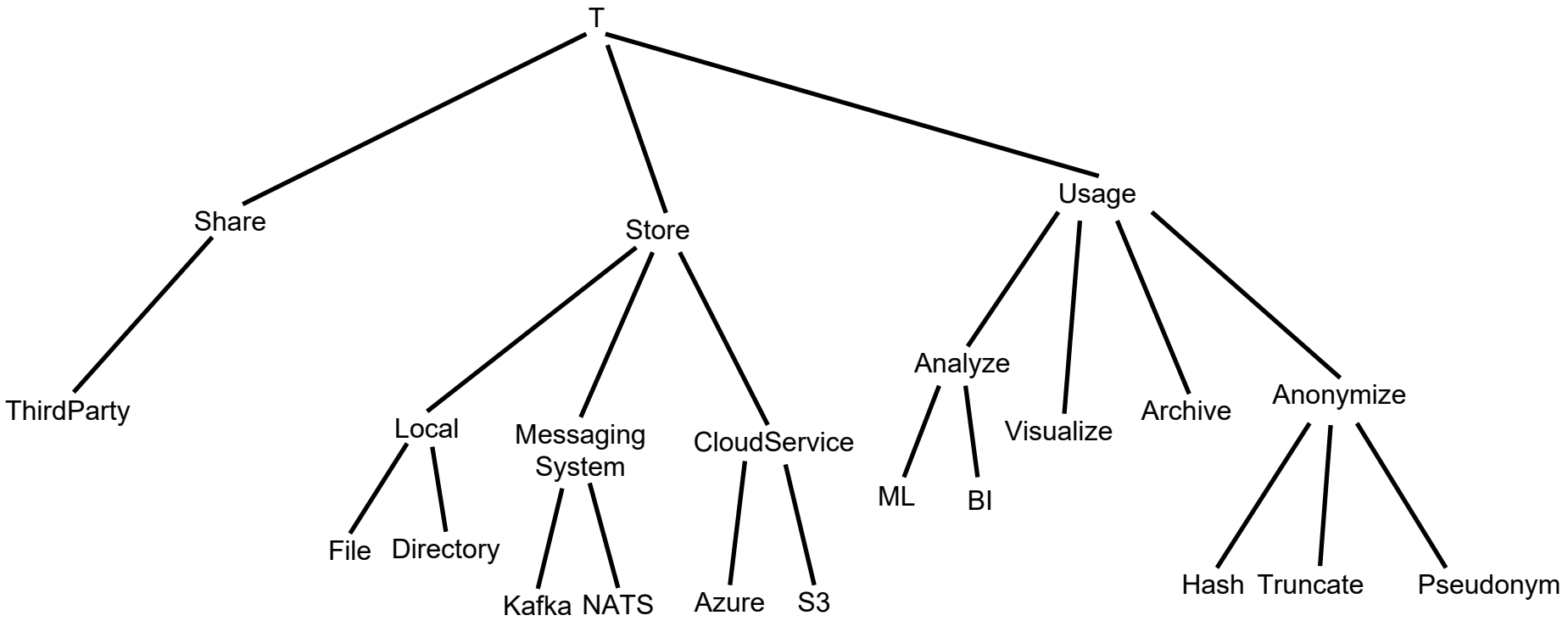
工作内容1：找到代码中对个人信息

工作内容2：找到代码中对个人信息的处理操作

个人信息


T
Sociality
PersonalData
Profession
Identifying
Tracking
JobTitle
.....
Interview
Name
.....
UserName
Device
Location
Contact
Cookie
IPAddress
MacAddress
Country
HouseNumber
EmailAddress
PhoneNumber

処理目的


T
Share
Store
Usage
ThirdParty
Local
Messaging System
CloudService
Analyze
Visualize
Archive
Anonymize
File
Directory
Kafka
NATS
Azure
S3
ML
BI
Hash
Truncate
Pseudonym

AST

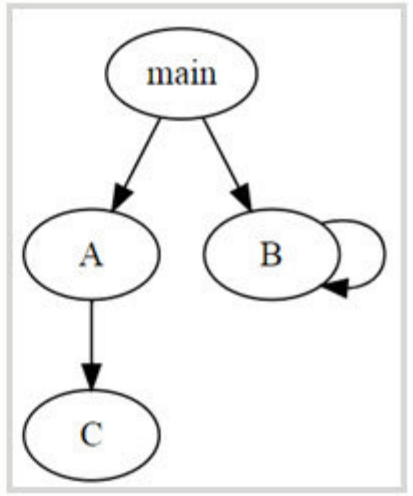
抽象语法树（Abstract Syntax Tree，AST）是源代码语法结构的一种抽象表示。它以树状的形式表现编程语言的语法结构，树上的每个节点都表示源代码中的一种结构。

```
while b ≠ 0
if a > b
a := a − b
else
b := b − a
return a
```

https://zh.wikipedia.org/wiki/%E6%8A%BD%E8%B1%A1%E8%AA%9E%E6%B3%95%E6%A8%B9

Call Graph

```
class Main {
  public static void main(String[] args) {
    A();
    B();
  }
  public static int A(){
    C();
  }
  public static void B(){
      B();
  }
}
```



内容说明

第六条　处理个人信息应当具有明确、合理的目的，并应当与处理目的直接相关，采取对个人权益影响最小的方式。收集个人信息，应当限于实现处理目的的最小范围，不得过度收集个人信息。

工作内容3：对比代码行为是否符合隐私政策

相关条款

当你创建 Apple ID、申请商业信贷、购买和/或激活产品或设备、下载软件更新、报名参加 Apple Store 商店的课程、连接到我们的服务、联系我们(包括通过社交媒体)、参与在线调查或以其他方式与 Apple 互动时，我们可能会收集各种信息，包括：

- **账号信息。**你的 Apple ID 和相关账号详细信息，包括电子邮件地址、注册的设备、账号状态和已使用时间
- **设备信息。**可用于识别你设备的数据 (如设备序列号) 或关于设备的数据 (如浏览器类型)
- **联系信息。**姓名、电子邮件地址、实际地址、电话号码或其他联系信息等数据
- **付款信息。**有关你的账单地址和付款方式的数据，例如银行详细信息、信用卡、借记卡或其他支付卡信息
- **交易信息。**有关 Apple 产品和服务的购买或由 Apple 促成的交易的数据，包括在 Apple 平台上进行购买的相关信息
- **防欺诈信息。**用于帮助识别和防止欺诈的数据，包括设备信用评分
- **使用数据。**有关你在产品上的活动和产品使用的数据，例如服务内的 app 启动数据，包括浏览历史记录、搜索历史记录、产品交互、崩溃数据、性能和其他诊断数据以及其他使用数据
- 位置信息。精确位置 (仅用于支持“查找”功能) 和大致位置
- **健康信息。**与个人健康状况相关的数据，包括与个人身体或心理健康或状况有关的数据。个人健康数据还包括可用于推断或检测个人健康状况的数据。如果你使用 Apple 健康调查研究 App 参与某项研究，可通过阅读 Apple 健康研究 App 隐私政策，了解你的个人数据隐私所适用的政策。
- **健身信息。**有关你选择共享的健身和锻炼数据的详细信息
- **财务信息。**包括工资、收入和资产数据在内的详细信息，以及与 Apple 品牌的金融产品相关的信息
- **政府身份证件数据。**在某些司法辖区，出于提供商业信贷、管理预订或遵守法律规定的目的，我们可能会要求用户提供由政府发放的身份证件，但仅限于为数不多的情形，例如开设无线账号和激活设备
- **你提供给我们的其他信息。**例如你与 Apple 的详细通信内容，包括与客户支持人员的互动内容，以及通过社交媒体渠道进行联系的相关内容

NLP分析

基于词典匹配	基于统计	基于深度学习
优点：速度快、成本低	优点：适应性较强	优点：准确率高、适应性强
缺点：适应性不强，不同领域效果差异大	缺点：成本较高，速度较慢	缺点：成本高，速度慢

NLP分析

中文分词工具
下面排名根据 GitHub 上的 star 数排名：
1.Hanlp
2.Stanford 分词
3.ansj 分词器
4.哈工大 LTP
5.KCWS分词器
6.jieba
7.IK
8.清华大学THULAC
9.ICTCLAS

英文分词工具：

1.Keras
2.Spacy
3.Gensim
4.NLTK

NLP分类

第二十八条　敏感个人信息是一旦泄露或者非法使用，容易导致自然人的人格尊严受到侵害或者人身、财产安全受到危害的个人信息，包括生物识别、宗教信仰、特定身份、医疗健康、金融账户、行踪轨迹等信息，以及不满十四周岁未成年人的个人信息。

工作内容4：对获取隐私政策中的个人信息进行自动分类

参考资料：　https://arxiv.org/pdf/2004.03705.pdf

内容说明

第九条　个人信息处理者应当对其个人信息处理活动负责，并采取必要措施保障所处理的个人信息的安全。

工作内容5：判断用户数据在处理过程中是否进行过安全处理

相关条款

【Android】QQ第三方信息共享清单

功能类型	第三方名称	使用目的	处理方式	个人信息类型	第三方隐私政策或官网链接
推送通知消息（OPPO推送SDK）	广东欢太科技有限公司	用于在OPPO设备上推送消息	通过加密处理的安全处理方式	设备标识信息、网络信息、设备状态信息、应用使用信息	https://open.oppomobile.com/wiki/doc#id=10288
推送通知消息（VIVO推送SDK）	维沃移动通信有限公司及其全球的企业附属公司	用于在VIVO设备上推送消息	通过加密处理的安全处理方式	设备标识信息、网络状态信息	https://www.vivo.com.cn/about-vivo/privacy-policy https://dev.vivo.com.cn/documentCenter/doc/366
推送通知消息（华为推送SDK）	华为终端有限公司及其关联公司	用于在华为设备上推送消息	通过加密处理的安全处理方式	应用标识信息、消息下发记录等	https://developer.huawei.com/consumer/cn/doc/development/HMSCore-Guides/privacy-statement-0000001050042021
信息分享（微博SDK）	北京微梦创科网络技术有限公司	分享内容到新浪微博	通过加密处理的安全处理方式	设备标识信息、网络信息	https://open.weibo.com/wiki/index.php/SDK
广告服务	广告主和/或其委托的代理商、第三方广告监测服务商	帮助广告主投放、评估、监测、提升广告投放效果，具体详见《广告服务第三方信息共享清单》https://privacy.qq.com/document/preview/7c8439a1d632427180da2a497f05dba1	通过去标识化、加密传输和处理的安全处理方式	如IDFA、OAID等设备标识符；如曝光、点击数据等广告数据；如系统语言、屏幕高宽、屏幕方向等设备信息	/
QQ钱包支付	财付通支付科技有限公司	帮助用户使用QQ钱包支付服务	采取去标识、加密等方式进行传输和处理	用户标识信息、网络IP地址、订单金额	https://www.tenpay.com/v3/helpcenter/low/privacy.shtml
微信支付	财付通支付科技有限公司	帮助用户使用微信支付服务	采取去标识、加密等方式进行传输和处理	用户标识信息、网络IP地址、订单金额	https://www.tenpay.com/v3/helpcenter/low/privacy.shtml

程序分类

通过代码的相似度判定进行功能判断

检测分类	中间表现形式	匹配算法	优点	缺点
基于文本	字符	字符匹配	算法实现简单，几乎可以检测所有编程语言的源代码	不能识别程序的语法、语义等信息，检测准确率较低
基于词法	Token序列[1,10-11]	LCS、后缀树[24]、语义索引、Karp-Rabin指纹算法	使用轻量级工具，可扩展到对多种编程语言的代码和纯文本的检测，同时相对于复杂算法具有更低的时空复杂度	不能识别程序的语法、语义等逻辑信息，检测准确率较低
基于语法	抽象语法树[12-14]	子树匹配[12]	可识别程序的语法信息，检测准确率较高	构造AST的代价较大，子树匹配算法的复杂度较高
基于语义	程序依赖图[15-20]	子图匹配[22]、程序切片[17,29]	可识别程序的语义逻辑信息，检测准确率较高	构造PDG和子图同构的PDG的代价较大；随着程序规模的扩大，时间复杂度和空间复杂度也提高
基于度量值	程序属性[30-33]	直接比较[30]、欧氏距离[31]、度量值比较[30,32]	检测准确率高，便于代码重构	局限于固定粒度的检测，如果粒度太大，则漏检率会很高

http://www.jsjkx.com/CN/article/openArticlePDF.jsp?id=18909

研究数据

PART THREE

隐私政策

腾讯隐私政策：https://privacy.qq.com/policy/apps-privacypolicy

华为隐私政策：https://www.huawei.com/cn/privacy-policy

苹果隐私政策：https://www.apple.com.cn/legal/privacy/szh/?cid=CDM-CN-DM-c00540-M00675

字节隐私政策：http://privacy.bytedance.com/zh/policy

阿里隐私政策：https://privacy.alibabagroup.com/#/home

……

项目代码

ERP类型、购物类型等信息相关的WEB项目代码都可以作为示例，可以自选：

https://github.com/wendaojidian/cmdb-python-master

https://github.com/Dolibarr/dolibarr

……

工具推荐

Java：wala: https://github.com/wala/WALA

java-callgraph: https://github.com/gousiosg/java-callgraph

soot: http://soot-oss.github.io/soot/

Python：python自带AST

pyan: https://github.com/davidfraser/pyan

pycallgraph: https://github.com/gak/pycallgraph

作业提交

PART FOUR

评分标准

方案一：工作内容1+工作内容2+工作内容5

方案二：工作内容1+工作内容3+工作内容4

……

基本工作：完成三个工作内容，构成一套完整流程，能够自动化进行隐私扫描工作

加分制：　完成文档；　（上限30分）

工作内容一；　（上限20分）

工作内容2/3/4/5；　（上限30分）

提交要求

1. 小组人数：1-3人自由组队。
2. 提交时间：期末考试后两周提交
3. 提交格式：研究报告

任课教师邮箱：

陈振宇　　zychen@njuedu.cn

助教联系邮箱：

隐私扫描选题-赵源　zhaoyuan@smail.nju.edu.cn

司法大数据选题-顾明政　　839306681@qq.com

格式要求

小组信息：人数 + 学号_姓名_邮箱+ 组员分工职责
研究问题：你选择研究的问题，你对该问题的认识，你的研究出发点。
代码开源地址：给出分析代码的地址，并且解释代码与研究问题的对应关系，代码的实现逻辑。
研究方法（重点）：包括你用到的数据分析方法、所使用的数据集，及其他你所用到的一切方法。这部分需要你的详细说明，要做到逻辑清晰且易理解。
案例分析：针对研究问题的分析，格式不作限制。
对这门课的意见和想对老师说的话，也欢迎感兴趣的同学们申请加入实验室的研究。
附录：你认为需要补充到研究报告中帮助读者理解的数据、图表等。
注：以上为研究报告的标准格式，如果你有特别的思路，也可以适当调整报告的结构。
提交PDF格式，宋体小四，1.5倍行距。

答疑

软件定义世界

质量保障未来

司法大数据自动化标注与分析

数据科学大作业选题-司法大数据分析篇

CONTENTS

研究背景

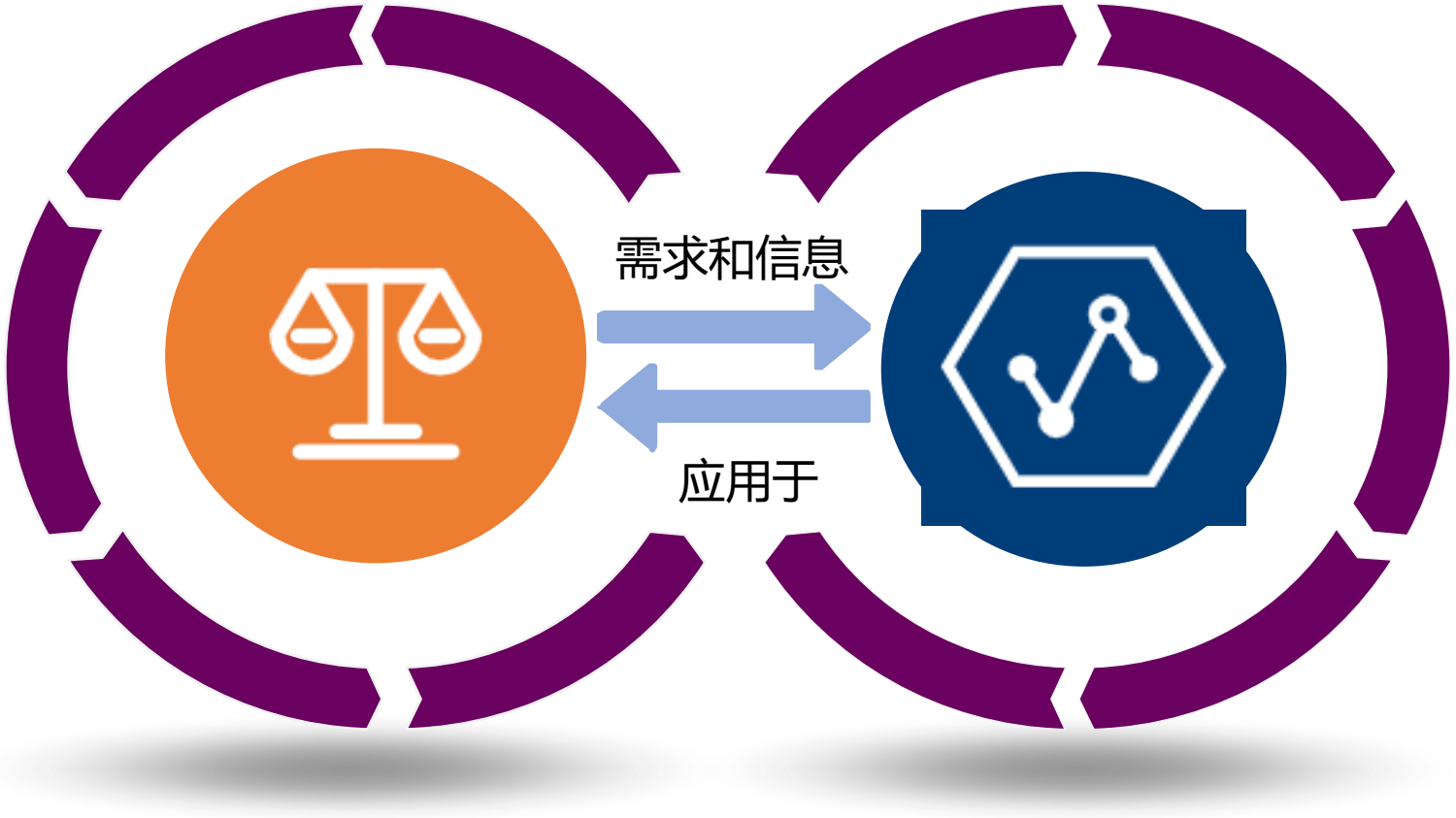
智慧法院是依托现代人工智能，围绕司法为民、公正司法，坚持司法规律、体制改革与技术变革相融合，以高度信息化方式支持司法审判、诉讼服务和司法管理，实现全业务网上办理、全流程依法公开、全方位智能服务的人民法院组织、建设、运行和管理形态。

——最高人民法院

何为司法大数据?

司法领域

大数据和人工智能技术



需求:
罪名预测、
法条推荐、
相似案例匹配……

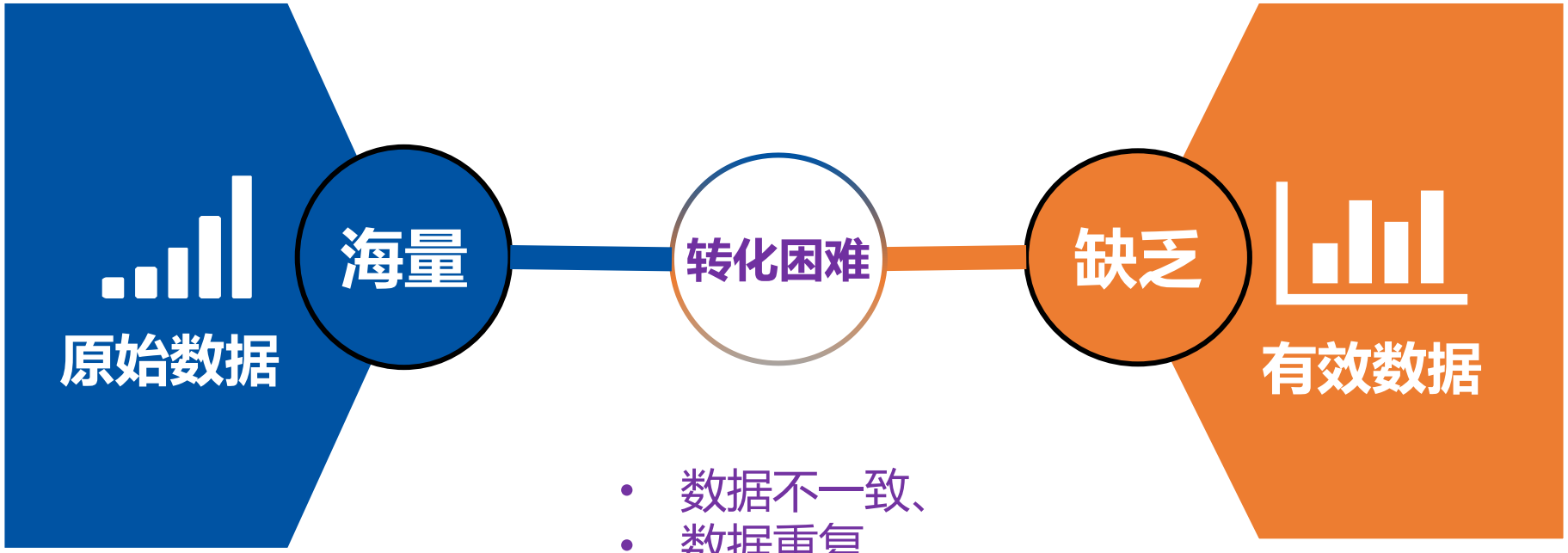
信息:
判决文书、
法律材料……

技术:
NLP、
OCR、
深度学习……

司法大数据有何用?

- 辅助判决
- 辅助各类法律工作
- 辅助法律材料管理
- 提高判决公平性
- 提高司法程序效率
- ……

司法大数据难点?



- 数据不一致、
- 数据重复、
- 数据损坏、
- 缺少标注
- ……

CONTENTS

研究内容与方向

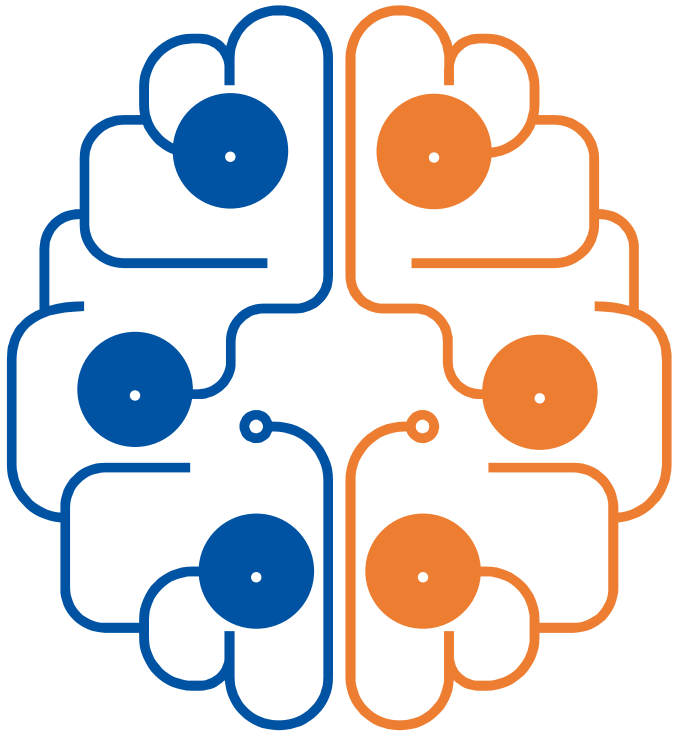
性别分布

时间范围

案由比例

地区人口

根据数据分布筛选
与预处理
对应的刑事案件
数据集



对获取的案件数据集
进行分析
对犯罪基本信息进行
自动化标注

非线性回归

假设检验

随机采样

相关性分析

CONTENTS

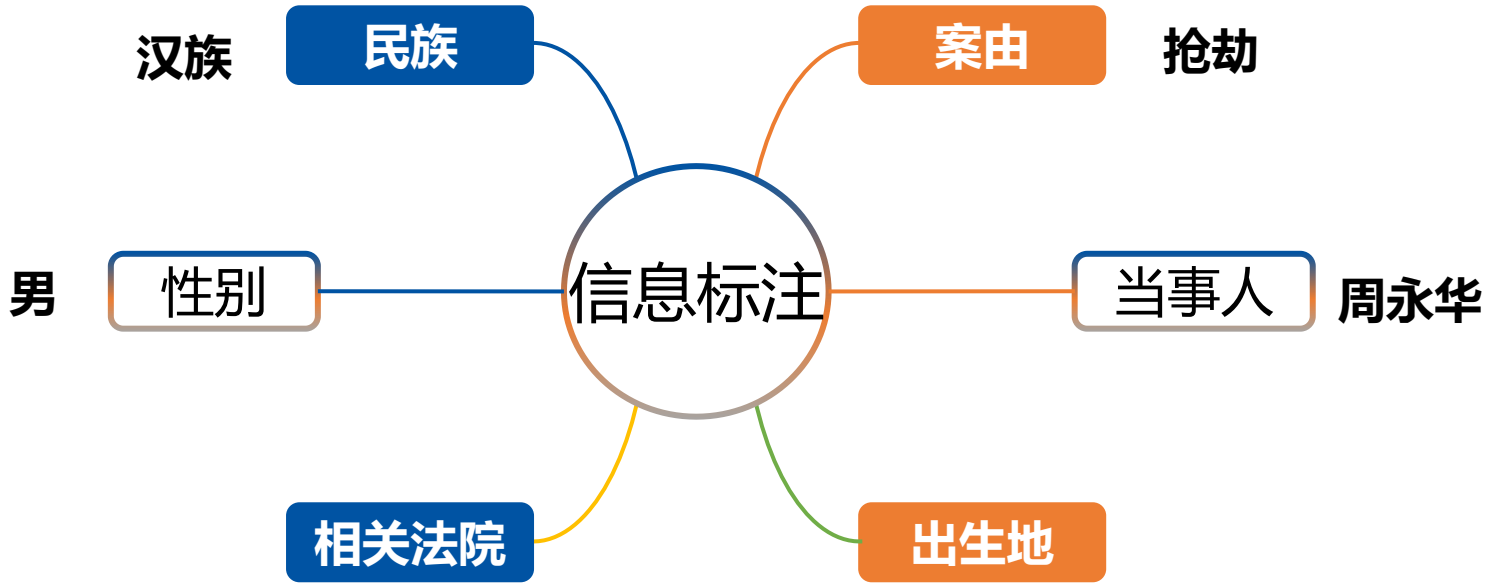
数据来源　中国裁判文书网：https://wenshu.court.gov.cn/

犯罪基本信息标注

中华人民共和国最高人民法院
刑 事 裁 定 书
被告人周永华，小名周玉，男，汉族，1975年1月20日出生，贵州省威宁彝族回族苗族自治县人，文盲，农民，户籍地威宁彝族回族苗族自治县××镇××村××组，住所地云南省个旧市××镇××街×××旁出租房。2019年8月6日被逮捕。现在押。
云南省红河哈尼族彝族自治州中级人民法院审理红河哈尼族彝族自治州人民检察院指控被告人周永华犯抢劫罪一案，于2019年12月30日以（2019）云25刑初160号刑事附带民事判决，认定被告人周永华犯抢劫罪，判处死刑，剥夺政治权利终身，并处没收个人全部财产。宣判后，周永华提出上诉。云南省高级人民法院经依法开庭审理，于2020年5月26日以（2020）云刑终218号刑事裁定，驳回上诉，维持原判，并依法报请本院核准。本院依法组成合议庭，对本案进行了复核，依法讯问了被告人。现已复核终结。
经复核确认：
1.2019年4月10日凌晨，被告人周永华携带尖刀到云南省个旧市××镇××××蔬菜批发市场，准备偷抢物品变卖用以购买毒品吸食
……
上述事实，有第一审、第二审开庭审理中经质证确认的受案登记表、立案决定书、手机购买发票复印件等书证，证人孙某的证言，被害人辛某的陈述，价格鉴定意见，现场勘验、检查、辨认笔录，监控视频等证据证实。被告人周永华亦供认。足以认定。
本院认为，被告人周永华以非法占有为目的，使用暴力手段强行劫取他人财物，其行为已构成抢劫罪。周永华多次持刀劫取他人财物，致一人死亡，犯罪情节极其恶劣，社会危害极大，罪行极其严重，应依法惩处。第一审判决、第二审裁定认定的事实清楚，证据确实、充分，定罪准确，量刑适当。审判程序合法。依照《中华人民共和国刑事诉讼法》第二百四十六条、第二百五十条和《最高人民法院关于适用〈中华人民共和国刑事诉讼法〉的解释》第三百五十条第（一）项的规定，裁定如下：
核准云南省高级人民法院（2020）云刑终218号维持第一审以抢劫罪判处被告人周永华死刑，剥夺政治权利终身，并处没收个人全部财产的刑事裁定。
本裁定自宣告之日起发生法律效力。

犯罪基本信息标注

标注结果需要以json
文件的格式保存到本地



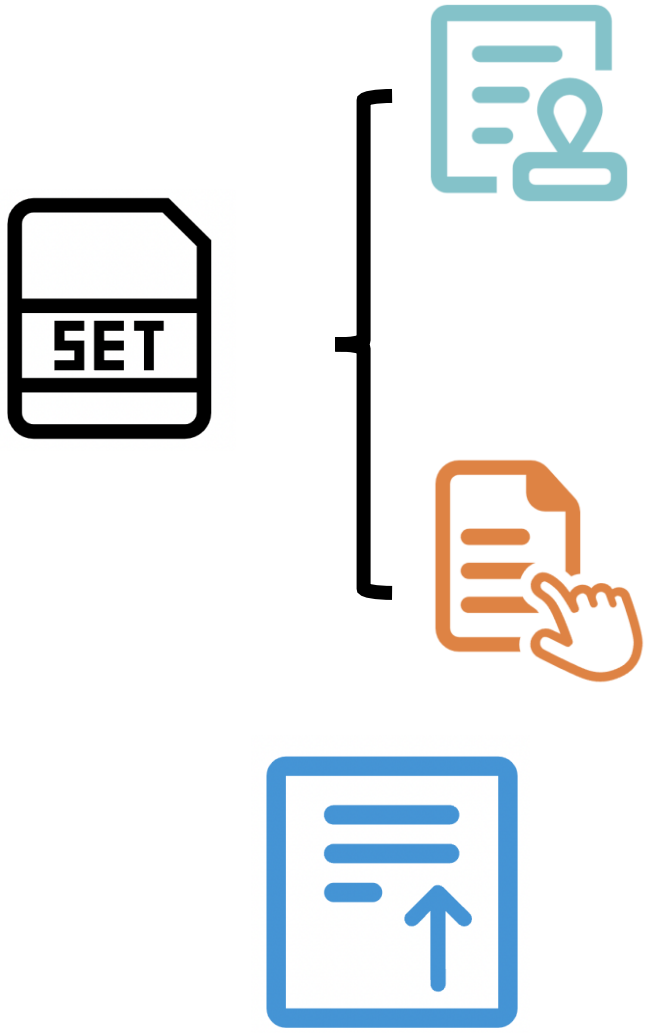
云南省红河哈尼族彝族
自治州中级人民法院、
云南省高级人民法院、
中华人民共和国最高人民法院

贵州省威宁彝族
回族苗族自治县

{
"Criminals": "周永华",
"Gender": "男",
"Ethnicity": "汉族",
"Birthplace": "贵州省威宁彝族回族苗族自治县",
"Accusation": "抢劫",
"Courts": "云南省红河哈尼族彝族自治州中级人民法院, 云南省高级人民法院, 中华人民共和国最高人民法"
}

CONTENTS

系统功能——60分

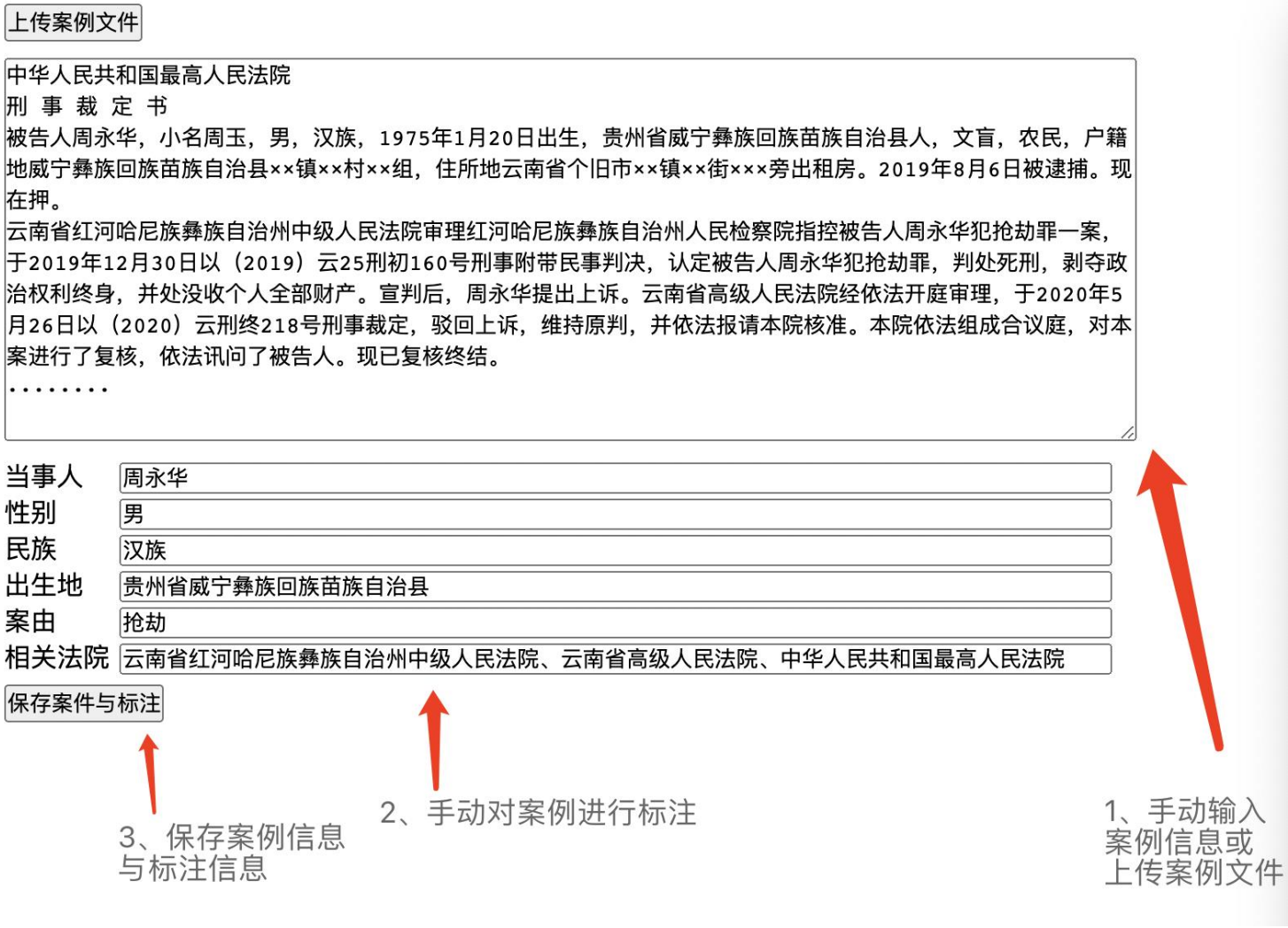


- 获取并标注100份数据集
内容包括：案件文本.txt、标注.json

- 提交这两个文件和研究报告

系统功能——70分

- 实现自动化标注工具
  - 功能点：
    - 用户可以在前端页面上传本地的文本或者直接输入案件信息
    - 用户可以在页面上手工输入需要提取的基本信息，如输入当事人，然后输入张三
    - 用户可以点击保存按钮将案件和基本信息保存到本地，文件分别为：案件文本.txt、标注.json
  - 提交内容：
    - 源码
    - 说明文档
    - 演示视频
    - 研究报告



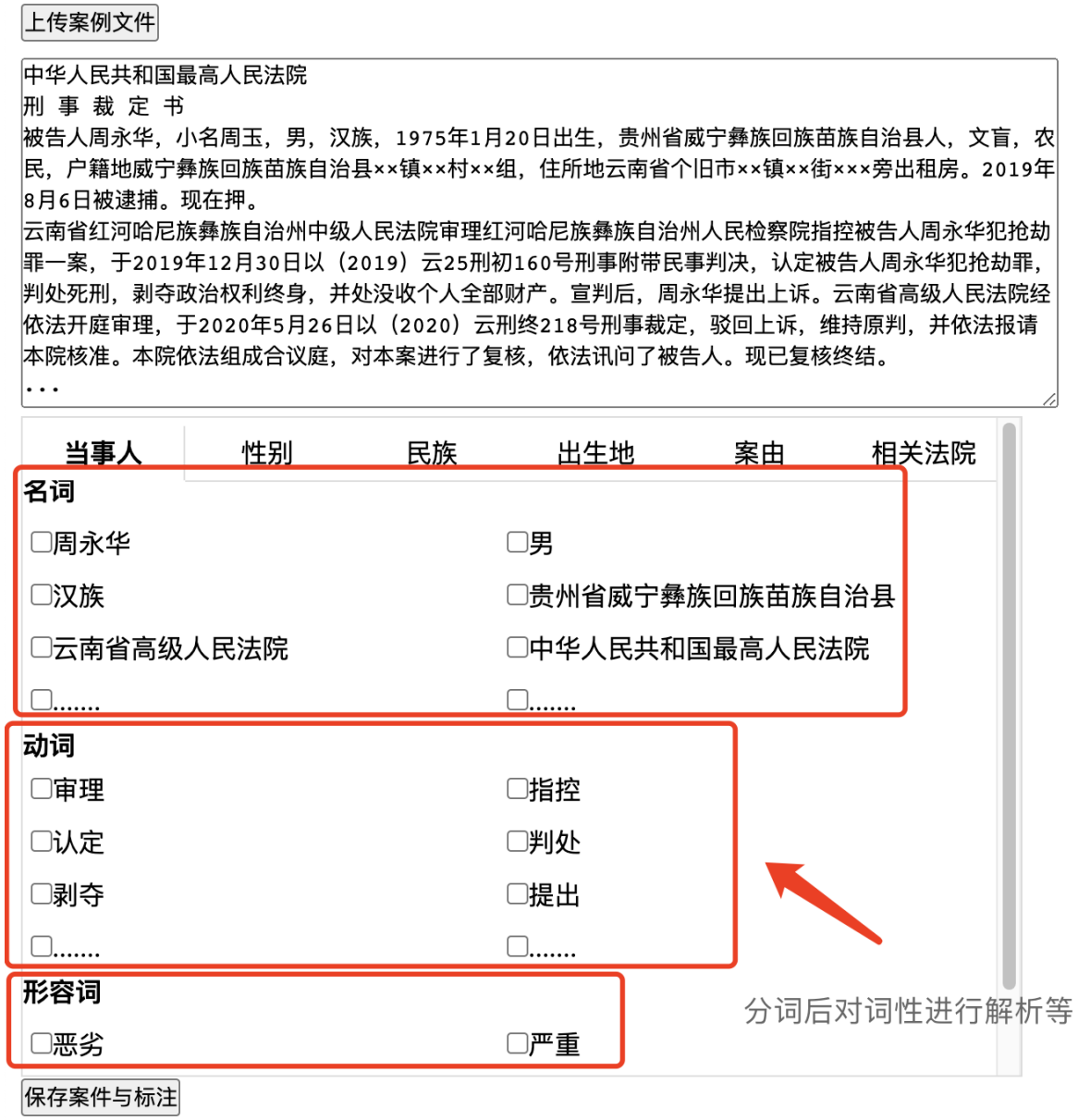
系统功能——80分

- 实现自动化标注工具
  - 功能点：
    - 用户可以在前端页面上传本地的文本或者直接输入案件信息，系统对文本内容进行自动化分词，并将分词结果展示为可选项形式
    - 用户可以根据需要的基本信息如当事人等，点击对应的可选项，以此对基本信息进行标注
    - 用户可以点击保存按钮将案件和基本信息保存到本地，文件分别为：案件文本.txt、标注.json
      - 提交内容：
        - 源码
        - 说明文档
        - 演示视频
        - 研究报告

系统能够完成分词功能，通过手动选取每个标注信息所对应的词条

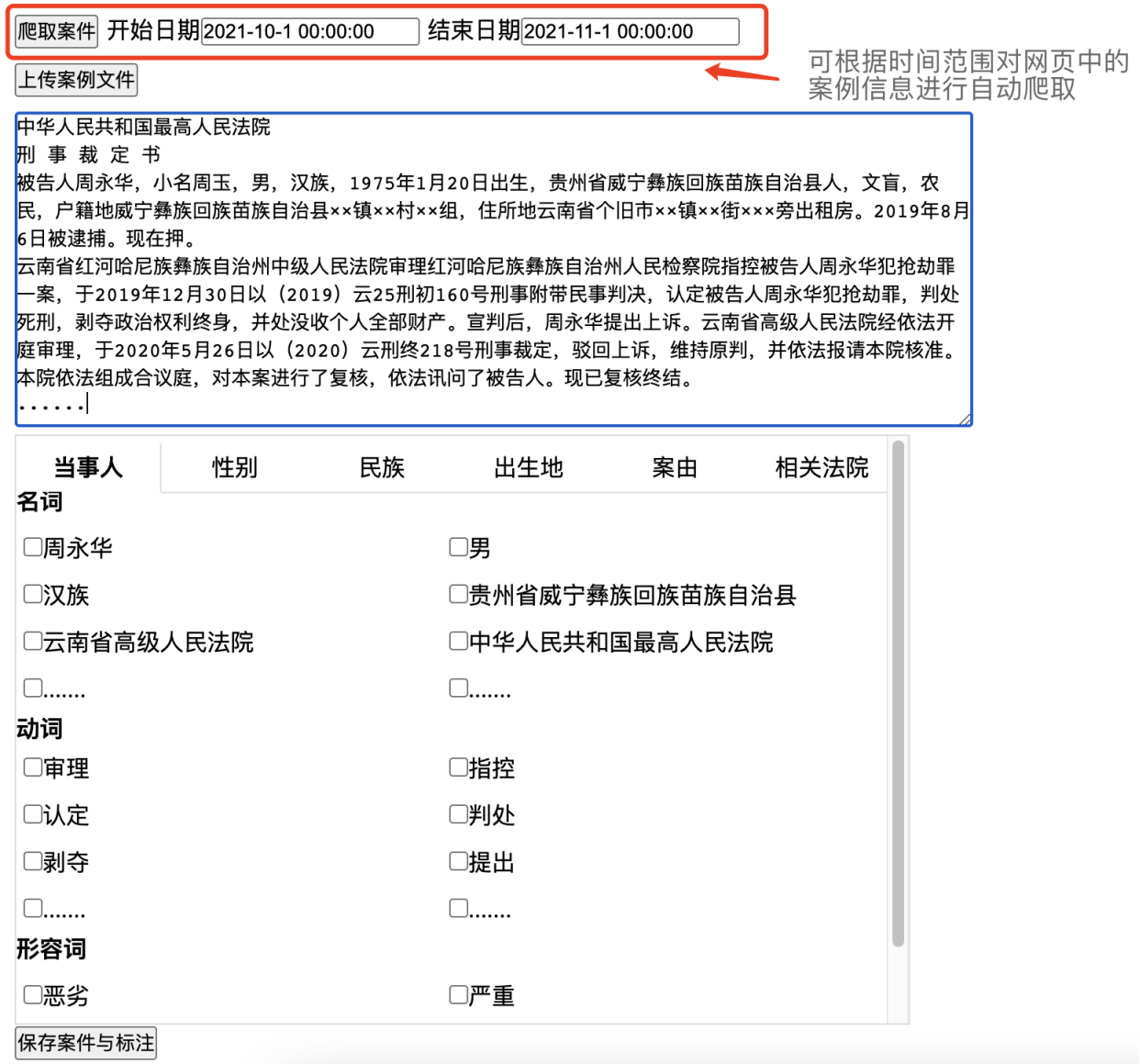
系统功能——90分

- 实现自动化标注工具
  - 功能点：
    - 用户可以在前端页面上传本地的文本或者直接输入案件信息，系统对文本内容进行自动化分词，通过文本命名实体、词性分析等方法获取基本信息可能对应的实体，并将其展示为可选项形式
    - 用户可以根据需要的基本信息如当事人等，点击对应的可选项，以此对基本信息进行标注
    - 用户可以点击保存按钮将案件和基本信息保存到本地，文件分别为：案件文本.txt、标注.json
  - 提交内容：
    - 源码
    - 说明文档
    - 演示视频



系统功能——95分

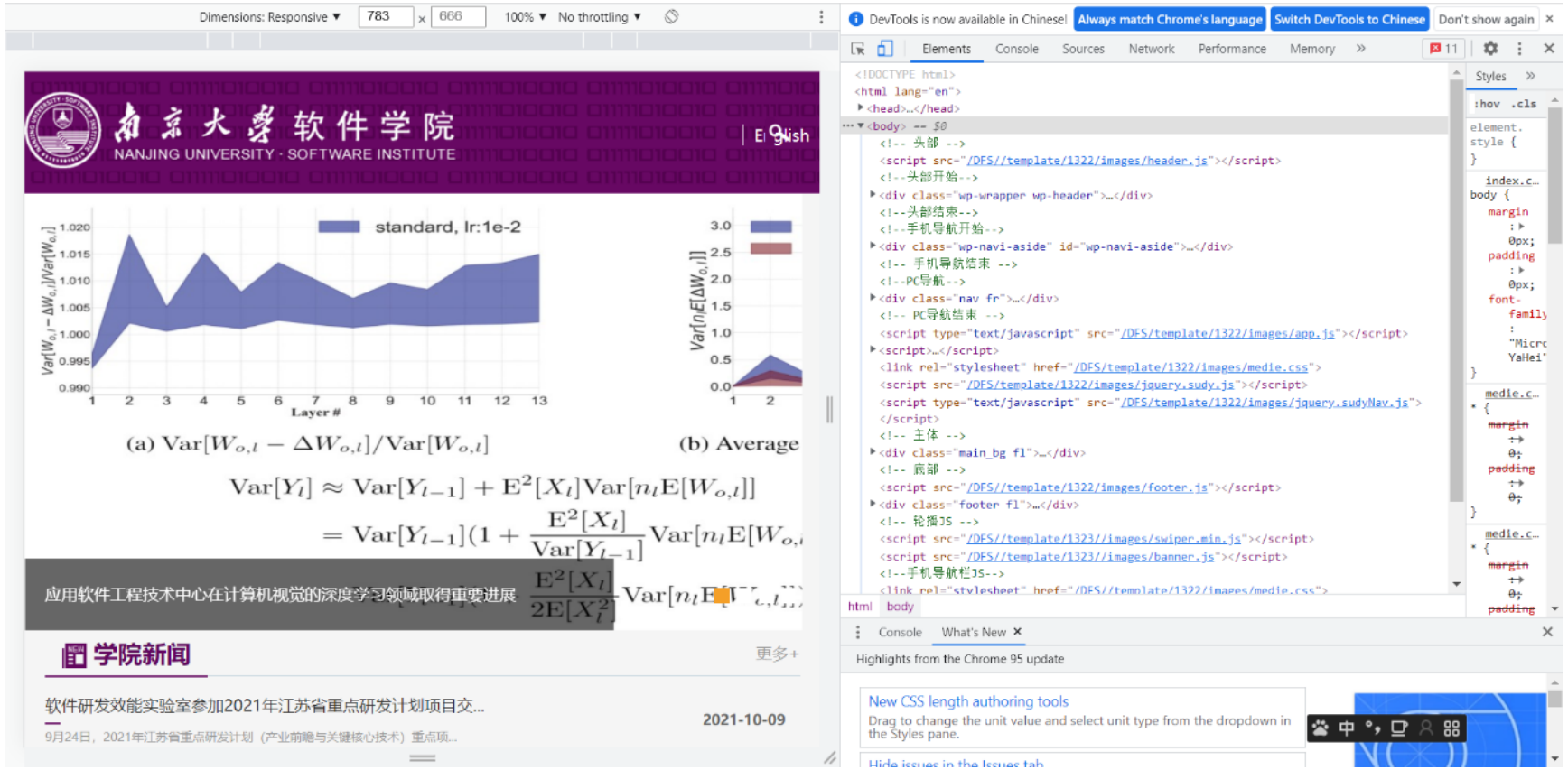
- 实现自动化爬虫和标注工具
  - 功能点：
    - 用户可以在前端页面上选择选择年份以及案件数量，系统根据用户的选择自动爬取对应的数据集并保存到本地
    - 用户可以在前端页面上传本地的文本或者直接输入案件信息，系统对文本内容进行自动化分词，通过文本命名实体、词性分析等方法获取基本信息可能对应的实体，并将其展示为可选项形式
    - 用户可以根据需要的基本信息如当事人等，点击对应的可选项，以此对基本信息进行标注
    - 用户可以点击保存按钮将案件和基本信息保存到本地，文件分别为：案件文本.txt、标注.json
      - 提交内容：
        - 源码
        - 说明文档
        - 演示视频
        - 研究报告



CONTENTS

相关工具——爬虫

- 什么是爬虫？
  - 基于预定义行为的作用于网页的信息探测器和信息采集器
  - “爬虫”
- 爬虫的原理是什么？
  - 模拟用户
  - 发送HTTP请求
  - 获取HTTP响应
  - 解析响应内容
  - ……



相关工具——爬虫

- 常用框架：
  - request　　基础的HTTP请求GET/POST/OPTIONS…发送
    - 参数设置
    - 请求头设置
    - ……

```python
import requests

r = requests.get('http://cuiqingcai.com')
print type(r)
print r.status_code
print r.encoding
#print r.text
print r.cookies
```

```python
import requests

payload = {'key1': 'value1', 'key2': 'value2'}
headers = {'content-type': 'application/json'}
r = requests.get("http://httpbin.org/get", params=payload, headers=headers)
print r.url
```

相关工具——爬虫

- 常用框架：
  - XPath　　一门用于在xml文件中查找信息的语言
    - https://www.w3school.com.cn/xpath/index.asp

路径表达式	结果
bookstore	选取 bookstore 元素的所有子节点。
/bookstore	选取根元素 bookstore。 注释：假如路径起始于正斜杠(/)，则此路径始终代表到某元素的绝对路径！
bookstore/book	选取属于 bookstore 的子元素的所有 book 元素。
//book	选取所有 book 子元素，而不管它们在文档中的位置。
bookstore//book	选择属于 bookstore 元素的后代的所有 book 元素，而不管它们位于 bookstore 之下的什么位置。
//@lang	选取名为 lang 的所有属性。

路径表达式	结果
/bookstore/book[1]	选取属于 bookstore 子元素的第一个 book 元素。
/bookstore/book[last()]	选取属于 bookstore 子元素的最后一个 book 元素。
/bookstore/book[last()-1]	选取属于 bookstore 子元素的倒数第二个 book 元素。
/bookstore/book[position()<3]	选取最前面的两个属于 bookstore 元素的子元素的 book 元素。
//title[@lang]	选取所有拥有名为 lang 的属性的 title 元素。
//title[@lang='eng']	选取所有 title 元素，且这些元素拥有值为 eng 的 lang 属性。
/bookstore/book[price>35.00]	选取 bookstore 元素的所有 book 元素，且其中的 price 元素的值须大于 35.00。
/bookstore/book[price>35.00]/title	选取 bookstore 元素中的 book 元素的所有 title 元素，且其中的 price 元素的值须大于 35.00。

相关工具——爬虫

- 常用框架：
  - lxml　　用于处理xml的分析库，支持xpath

```
from lxml import etree
html = etree.parse('hello.html')
print type(html)
result = html.xpath('//li')
print result
print len(result)
print type(result)
print type(result[0])
```

运行结果

```
<type 'lxml.etree._ElementTree'>
[<Element li at 0x1014e0e18>, <Element li at 0x1014e0ef0>, <Element li at 0x1014e0f38>, <Element li
5
<type 'list'>
<type 'lxml.etree._Element'>
```

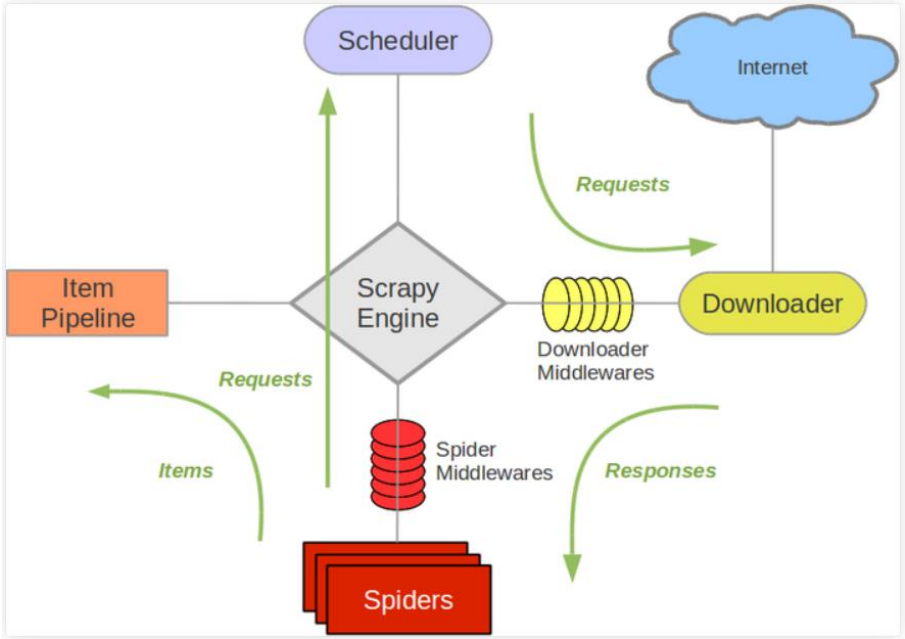
相关工具——爬虫

- 常用框架：
  - BeautifulSoup4　　用于处理xml的分析库，将xml处理成树形结构

    - Tag　　标签
    - NavigableString　　文字
    - BeautifulSoup　　特殊的Tag，表示整个文档的全部内容
    - Comment　　特殊的NavigableString，表示注释文字

相关工具——爬虫

- 常用框架：
  - Selenium　　自动化测试框架，模拟用户行为，跳脱解析思路
    - 页面切换、页面上下拖动、登录验证、cookies
  - Scrapy　　爬虫框架



- Scrapy Engine(引擎)：负责Spider、ItemPipeline、Downloader、Scheduler中间的通讯，信号、数据传递等。
- Scheduler(调度器)：它负责接受引擎发送过来的Request请求，并按照一定的方式进行整理排列，入队，当引擎需要时，交还给引擎。
- Downloader（下载器）：负责下载Scrapy Engine(引擎)发送的所有Requests请求，并将其获取到的Responses交还给Scrapy Engine(引擎)，由引擎交给Spider来处理，
- Spider（爬虫）：它负责处理所有Responses,从中分析提取数据，获取Item字段需要的数据，并将需要跟进的URL提交给引擎，再次进入Scheduler(调度器).
- Item Pipeline(管道)：它负责处理Spider中获取到的Item，并进行进行后期处理（详细分析、过滤、存储等）的地方。
- Downloader Middlewares（下载中间件）：你可以当作是一个可以自定义扩展下载功能的组件。
- Spider Middlewares（Spider中间件）：你可以理解为是一个可以自定扩展和操作引擎和Spider中间通信的功能组件（比如进入Spider的Responses;和从Spider出去的Requests）

相关工具——爬虫

- 其他框架:
  - PyQuery
  - PhantomJS
  - ……

相关工具——NLP

- NLTK：https://www.nltk.org/
- HanLP：https://github.com/hankcs/HanLP/tree/1.x（包含Java和pyhon版本，可直接调用接口）
- CoreNLP：https://stanfordnlp.github.io/

- 哈工大的语言技术平台 （LTP）
- jieba

相关工具——LTP

- LTP安装： pip install ltp
- LTP初始化： ltp = LTP()，也可以下载ltp预训练模型
- LTP分词：使用LTP分句只需要调用ltp.sent_split函数
- LTP自定义词典： ltp.init_dict
- LTP词性标注： ltp.seg
- LTP命名实体识别： seg, hidden = ltp.seg([str])　　ner = ltp.ner(hidden)

```
from ltp import LTP
ltp = LTP()
sents = ltp.sent_split(["他叫汤姆去拿外衣。", "汤姆生病了。他去了医院。"])

# [
#   "他叫汤姆去拿外衣。",
#   "汤姆生病了。",
#   "他去了医院。"
# ]
```

```
seg, hidden = ltp.seg(["他叫汤姆去拿外衣。"])
pos = ltp.pos(hidden)
# [['他', '叫', '汤姆', '去', '拿', '外衣', '。']]
# [['r', 'v', 'nh', 'v', 'v', 'n', 'wp']]
```

具体内容参考：https://ltp.readthedocs.io/zh_CN/latest/

相关工具——jieba

- 三种分词模式：
  - 全模式、精确模式、搜索引擎模式
- 支持自定义字典
- 支持繁体字分词
- 其他功能：词性标注、关键词提取……

```
# 导入 jieba
import jieba
import jieba.posseg as pseg #词性标注
import jieba.analyse as anls #关键词提取
```

THANKS