# 

# “基于自然语言处理的智能合约自动生成”项目

# 任务书

## 2018.2.1

## **组内成员：**

## 黄霁昀、肖军、徐家兴、林未、张凯宁

（排名不分先后）

### 选题背景及价值

使用过租赁网站的人肯定有这样的经历，作为租出方，除了人为运营的网站数据库外没有可信、无法篡改的证据证明自己发起了交易，作为租入方，也缺乏可靠的平台记录签订的租入合同内容，金钱交易、合法权益难以得到安全保障。而在现实生活中，面对面签订的租赁合约也很可能被损毁、恶意修改，记录合约内容和交易流程的证据随时被单方面控制甚至无处可查，这样的交易也就意味着层出不穷的隐患。

考虑到合约是要有完备的法律效益和执行力的，需要非常完善却没有统一标准、撰写复杂、时常不规范，合约的执行、结束、惩罚都需要人为判别，这中间就会有很多漏洞，没有完整的可信监督机制造成举证难，以及合约一直停留在纸质，没有顺应数字化的潮流等现实背景，推出智能合约势在必行。

任何可以被单方面控制的所谓证据都不能成为异常交易过程有效、可信的举证，这一点电子商务和现实租赁领域都做的不够完善，而引入区块链这一分布式账本后，在交易双方都确认的情况下生成不可删改的节点，使证据的控制权不在任何一方的手中，无疑是推动交易安全性的成功之选。

***总而言之，创新点：***

1、利用自然语言处理技术处理合约文本

2、自动生成java合约代码

3、智能合约与区块链结合推动交易安全性

### 选题内容

本题围绕将自然语言处理后生成的的物品租赁合同代码部署到区块链并监督交易流程、自动对金钱转移作出判断、实时记录行为轨迹展开，安全、可信、全面。

使用自然语言处理技术自动生成智能合约代码，满足用户个性化订制租赁合同需求的同时保证了合同的规范性，准确提取合同关键词、按照合同逻辑生成JAVA代码，部署到区块链上，在双方确认的前提下启动交易监督，一步步触发合同条款并记录操作轨迹，避免第三方恶意干扰或人为判定疏漏，严格按照合同逻辑推动租赁交易，由于交易信息实时记录到区块链上，进行分布式安全有效的保护性管理并设计查询功能，为后期异常交易举证追踪提供方便。

### 开发工具功能简介

1. ***合同订制部分***

录入一份物品租赁合同内容的渠道有三：

①用户根据需要填写已有合同模板（如不填写的内容也有缺省值保证合同逻辑正常运行）

②UI界面设计拖拽、点选等功能录入关键词

③用户个性化提交纸质合同，自然语言处理分析合同语言逻辑，捕捉关键信息待用户确认

以上如有异常用户可以提出异议并微调。

1. ***代码自动生成部分***

根据合同关键信息生成代码部署到区块链

1. ***交易流程详解***

用户注册->确认下单->双方签订合同并确认->提取信息生成代码及合同部署->接收外部信息实时自动判定合同触发条件是否满足->执行相关交易操作（如自动转账）、返回相关信息并在区块链记录轨迹->交易结束用户可查

以上交易过程如报异常随时停止。

1. ***预计触发条件及功能***

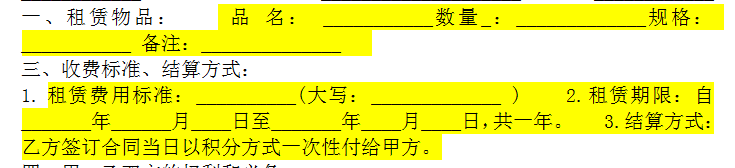
注册、登录、存款、上传合同、合同确认、余额是否充足、确认按时寄出、确认收货、开始租用计时、是否续租（续租申请、续租同意）、是否按时归还、确认完好、申请异常、异常确认、异常赔偿、租用时长查询、是否支付、是否滞后支付、押金是否退还、结束交易、交易历史下载。

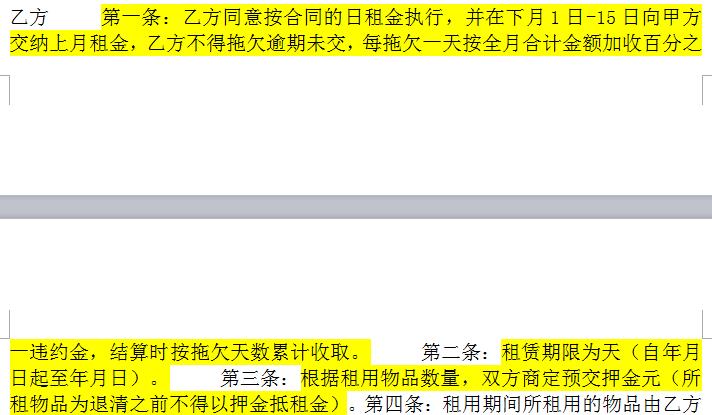
1. ***初始合同模板举例及关键字段提取***

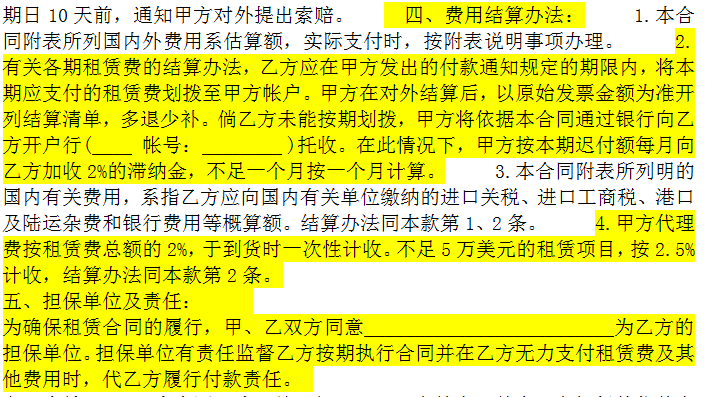
**包括：**

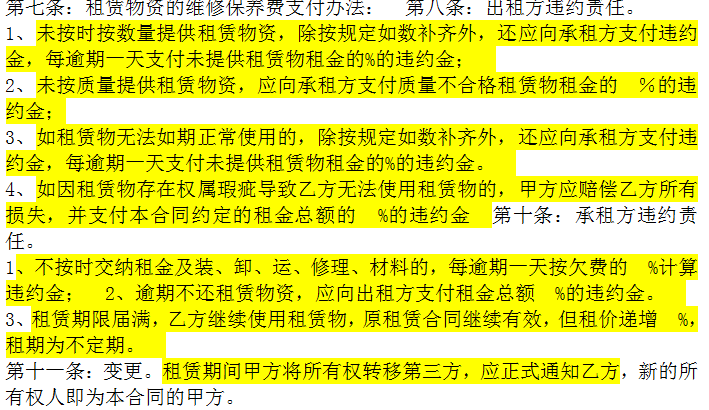
合同双方姓名、电话、住址、身份证号、银行卡号等个人身份信息，租赁物特征及数量、备注、合同签订日期、租赁物交付时间、是否承担快递费用（及费用多少）、使用哪家快递公司、是否开具发票、租赁时段、押金比例、计费方式、滞后发货惩罚、租赁物实物与合同不符赔偿机制、是否允许续租、续租最长期限、续租租金政策、滞后支付惩罚、租物损坏赔偿、租用费用计算政策、最迟归还时间、滞后归还惩罚、归还地址或方式、是否为每日结算、物品是否可以转借他人、是否有担保人、担保人信息等信息块的提取。

***自然语言处理物件租赁合同模板举例：***

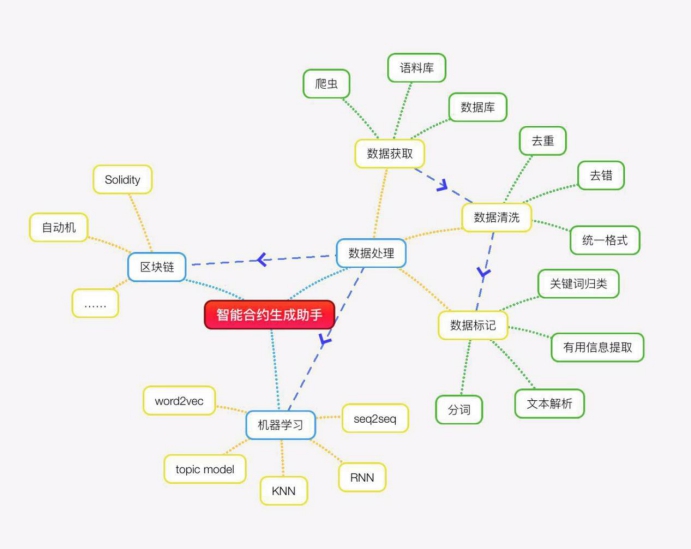








### 技术选型



**主要涉及数据处理、机器学习、函数设计、接入区块链等方面的内容。**

**实验环境：**

**Python 3.6**

**PyQt5**

**urllib,urllib2,json,pymssql,word2vec,Lenvenshtein,jieba,jieba.analyse,xlrd,xlwt**

其中：

#### 4.1 NLP涉及的部分内容

**举例：**

***（1）jieba分词***

jieba分词 是一款基于Trie树结构的中文分词工具。使用jieba分词，将自然语言分解为一个个的词组，然后再经过筛选去除没用的词语，留下有用的作为关键词。用本地Excel进行分词以及数据清洗的操作。从搜狗输入法里下载相关词库，再将词库的scel文件转换为TXT文件，然后用jieba分词根据词库删选爬下来的Excel里面的词语。但是程序运行有部分缺陷，一部分词未能删选出来，需要手动删除，比如的，与，之等词。然后将删选好的表格保存下来，然后依据这个表格来统计词频。

其主要的处理思路如下：

加载词典dict.txt

从内存的词典中构建该句子的DAG（有向无环图）

对于词典中未收录词，使用HMM模型的viterbi算法尝试分词处理

已收录词和未收录词全部分词完毕后，使用dp寻找DAG的最大概率路径

输出分词结果

jieba分词在本项目中主要用于文本分词。

***（2）word2vec***

word2vec 是word embedding（词向量）的一种浅层神经网络训练方法。利用机器学习和现有数据对模型进行训练，得到满足要求的反馈，初步实现所需功能。

word2vec本质上来说就是一个矩阵分解的模型，简单地说，矩阵刻画了每个词和其上下文的词的集合的相关情况。对这个矩阵进行分解，只取每个词对应在隐含空间的向量。

所以word2vec适合的情况就是对于一个序列的数据，在序列局部数据间存在着很强的关联。典型的就是文本的序列了，邻近的词之间关联很强，甚至可以通过一个词的上下文大概预测出中间那个词是什么。学习到的词向量代表了词的语义，可以用来做分类、聚类、也可以做词的相似度计算。此外，Word2vec本身的层次分类器或者采样方式实际上对热门item做了很大的惩罚，所以不会像一般的矩阵分解一样，最后算出来语义接近的都是热门词，这也是word2vec很好的一个特性。

word2vec在本项目中主要用于计算词语相似度，对同义词、相关词进行处理。

***（3）topic模型***

topic model 是一种针对文本隐含主题的建模方法，主要用于：

计算文本的相似性，考虑到文本语义，更好的刻画文本相似性，避免多义词，同义词的影响

文本聚类，用户聚类(RS)

去除噪音，只保留最重要的主题，更好的刻画文档

topic模型在本项目中主要用于进行文本聚类，对相同话题进行归纳。

***（4）KNN算法***

K近邻算法（k Nearest Neighbors, kNN）是一种基于实例的学习，通过计算新数据与训练数据特征值之间的距离，然后选取K个距离最近的邻居进行分类判断。

kNN由3个要素决定：

距离度量方法

k值

分类决定规则

kNN算法的过程:

选择一种距离度量方式, 通过所有的训练实例与输入实例的距离；

将距离按递增次序进行排序，选取与当前距离最小的k个点；

确定前k个点所在类别的出现频率；

将出现频率最高的类别作为输入实例的预测分类。

KNN算法在本项目中主要用于计算新数据与训练数据特征值之间的距离，并进行分类判断。

***（5）seq2seq(RNN)***

RNN（此时被称为 encoder）将输入句子表示为一个向量， 再使用另一个 RNN（此时被称为 decoder）解码这个向量获取输出。如在英汉机器翻译任务中，先使用 encoder RNN 处理英文句子获取语意向量，将该向量作为 decoder RNN 的初始输入，按顺序解码每个英文单词获取中文。

同样的，如果将自然语言作为encoder，代码作为decoder进行训练，即可实现代码自动生成。

#### 4.2 智能合约代码自动生成的函数设计

#### 4.3 区块链技术的部分内容

基于区块链技术的智能合约不仅可以发挥智能合约在成本效率方面的优势,而且可以避免恶意行为对合约正常执行的干扰。此过程的主要运作原理是将智能合约以数字化的形式写入区块链中,接着由区块链技术的特性来保障存储、读取、执行,整个过程透明可跟踪、不可攥改。同时,由区块链自带的共识算法构建出一套状态机系统,使得智能合约能够高效地运行。

区块链技术的特点：

（1）分布式多镜像储存，任一节点（镜像）保有全部记录

（2）记录权去中心化，通过某种措施竞争记录权

（3）对已被认可的记录，所有网络参与者无条件的共识认同

（4）公开、可回溯、防伪（有条件的）

**主要学习区块链存储和查询操作方法的调用。**

### 技术难点

### **数据获取方面：**有些网站可能存在反爬虫机制，需要防止IP封锁。

### **自然语言处理方面：**

### 模型训练。

比较不同模型的准确度并进行优化。

词语实体边界界定。在自然语言中词与词之间通常是连贯的，而正确划分、界定不同的词语实体是正确理解语言的基础 。这个问题对于汉语尤其突出。界定字词边界通常使用的办法是取用能让给定的上下文最为通顺且在方法上无误的一种最佳组合。

词义消歧。词义消歧包括多义词消歧和指代消歧。多义词是自然语言中非常普遍的现象。指代消歧是指正确理解代词所代表的人或事物。例如，在复杂交谈环境中，“他”、“it"到底指代谁。词义消歧需要对文本上下文、交谈环境和背景信息等有正确的理解。

方法的模糊性。自然语言方法常常会出现模棱两可的句子，即一个句子可能会解析出多棵语法树。

语言行为与计划。一个句子常常不只是字面上的意思而人类往往更注意其潜在的含义。

**3、代码函数设计方面：**需要注意合同的逻辑关系并在代码中对应、函数缺省值的配置、虚拟货币的正确转移等。

**4、区块链使用方面：**首次接入区块链的代码适应修改工作

### 六、开发计划

**6.1 分工情况和时间轴**

|  |  |  |
| --- | --- | --- |
| **阶段** | **主要负责人** | **预计时间段** |
| **数据获取和关键字段确认** | **黄霁昀、肖军、张凯宁** | **最好估计：2.2**  **最差估计：2.2** |
| **围绕提取合同关键字段功能的自然语言处理模型选取与训练** | **林未、徐家兴** | **最好估计：2.2-2.12**  **最差估计：2.2-2.14** |
| **选择并学习适合处理合同信息生成智能合约代码的语言** | **黄霁昀、肖军、张凯宁** | **最好估计：2.2-2.6**  **最差估计：2.2-2.7** |
| **后台搭建函数设计和基于Excel和数据库的初级测试** | **黄霁昀、肖军、张凯宁** | **最好估计：2.7-2.13**  **最差估计：2.7-2.14** |
| **自然语言处理与功能性代码自动生成对接测试** | **黄霁昀、肖军、张凯宁、徐家兴、林未** | **最好估计：2.20-2.21**  **最差估计：2.20-2.22** |
| **前端设计** | **黄霁昀、肖军、张凯宁** | **最好估计：2.23**  **最差估计：2.23-2.24** |
| **区块链的接口设计与接入** | **徐家兴、林未、黄霁昀、肖军、张凯宁** | **最好估计：2.24-2.25**  **最差估计：2.25-2.26** |
| **整个工具细节的测试与改进** | **黄霁昀、肖军、张凯宁、林未、徐家兴** | **最好估计：2.26-2.28 最差估计：2.27-2.28** |

**（注：主要负责人对该过程起主要组织作用，全组成员各负其责共同参与该项工作）**

**6.2 项目阶段概述**

1. **数据获取和关键字段确认**

搜集语料库，确认函数所需关键字段，尽量多下载合同模板，填入不同内容、不同表达方式的信息，为之后的自然语言处理设计、模型选取、训练做准备。

1. **围绕提取合同关键字段功能的自然语言处理模型选取与训练**

学习、选用不同的模型，对比自然语言处理效果，达到可以精准筛选出合同关键信息的目的。

1. **选择并学习适合处理合同信息生成智能合约代码的语言**

**（4）后台搭建函数设计和基于Excel和数据库的初级测试**

对关键信息进行进一步的处理，基本功能函数的设计，对自动生成的代码进行基于excel和数据库的信息输出比对、货币转移等功能测试与改进。

**（5）自然语言处理与功能性代码自动生成对接测试**

将两个阶段衔接起来，发现问题并修改代码对接，消除技术断层。

**（6）前端设计**

使用pyqt设计出简洁美观的UI界面，功能齐全，一目了然，用户使用方便。

**（7）区块链的接口设计与接入**

改进代码、调用方法生成独立有效的区块链接口，将之前在数据库测试过的信息成功存入区块链，测试区块链增加记录和查询记录功能的有效性。

**（8）整个工具细节的测试与改进**