# “基于自然语言处理的智能合约自动生成”项目

# 任务书

## 2017.12.3

## 组内成员：

## 徐家兴、林未、黄霁昀、肖军、张凯宁

[一、题目描述 3](#_Toc500098728)

[1.1选题背景 3](#_Toc500098729)

[1.2选题内容 4](#_Toc500098730)

[1.3选题价值 4](#_Toc500098731)

[二、技术选型 5](#_Toc500098732)

[2.1 NLP涉及的部分内容 5](#_Toc500098733)

[2.2 智能合约可能涉及到的人工神经网络 7](#_Toc500098734)

[2.3 区块链技术的部分内容 9](#_Toc500098735)

[三．技术难点 9](#_Toc500098736)

[四． 开发计划 10](#_Toc500098737)

### 一、题目描述

#### 1.1选题背景

有过就医经验的人肯定有过这样的经历：当我们在某家医院接受治疗，出于一些原因有转到另外一家医院时，经常会被要求重新检查、拍片等，这主要是因为之前一家医院的检查结果不能够转到新的医院。导致医疗信息交换屏障主要是因为目前医疗信息是以医疗机构为导向进行管理的，除非患者本人同意，在医院外部进行的一切有关患者信息的交换都会被禁止。这种以个体中心化医疗机构为导向进行的医疗信息管理，不仅导致医疗数据的分散，还严重阻碍了医疗数据的有效利用。

对患者来说，医护人员由于缺乏过往医疗信息，难以为患者提供最佳的医疗服务，还要进行不必要的重复检查和医学成像，增加医疗费用负担。一项调查结果显示，如果医疗信息可以被有效地交换，那么在急诊室进行的临床检查和放射学检查将会减少50%以上。除了增加费用，信息安全本身也受到威胁。目前，每年医院医疗数据遭到黑客入侵的案件都在迅速增加。仅2015一年，在美国遭受到黑客入侵的医疗数量就高达1.12亿条。据报告，因此造成的经济损失超过62亿美金。与医疗相关的福利，如保险，也会造成巨额的损失。不真实可靠的医疗数据，让保险无法准确的衡量，据报告称，全球范围内每年因为虚假索赔造成的保险损失高达4870亿美院，占美国年度总医疗费用的五分之一。

医学研究也同样需要这些信息，现在，医疗数据越来越多的被用于医学研究和护理项目开发。是否能将医疗数据隐去患者个人信息后用于机构和企业的学术研究？在美国，这是允许的，然而问题在于很多信息不难通过社交网络等进行识别。因此，对数据所有权的争议仍在继续，意味着这种数据分享方式有待改善。

医疗数据通常包含了患者的身份信息、治疗方案、治疗费等敏感信息，然后医药行业又是个门槛很高的行业，一旦数据泄露，不法人员利用信息很容易对患者造成健康财产上的伤害。

如果将这些敏感的医疗数据加密后写入区块链，不仅可以通过时间戳来确保数据精度，而且能确保数据安全，不会被没有权限的人看到，并且不会轻易被黑客攻击篡改。16年爱沙尼亚就启动了基于区块链的医疗健康档案安全项目，就是用“独立的法医品质的审计线索”保护这些数据的安全。

#### 1.2选题内容

本题以现代医院问诊为背景，通过NLP分析处理病人就诊过程中的一系列需求，基于智能合约自动生成、触发对区块链数据（病历信息）的全套操作（基于对记录的增、删、改、查）并记录操作轨迹，利用区块链对用户问诊细节和医院接诊情况进行分布式安全有效的保护性管理，信息经病人授权后可用于病历追踪，在必要时用于信息汇总、转发、统计分析，在行业标准下实现一定程度的智能化数据共享，达到节省大量人力物力，提供全面、准确信息的目的。

#### 1.3选题价值

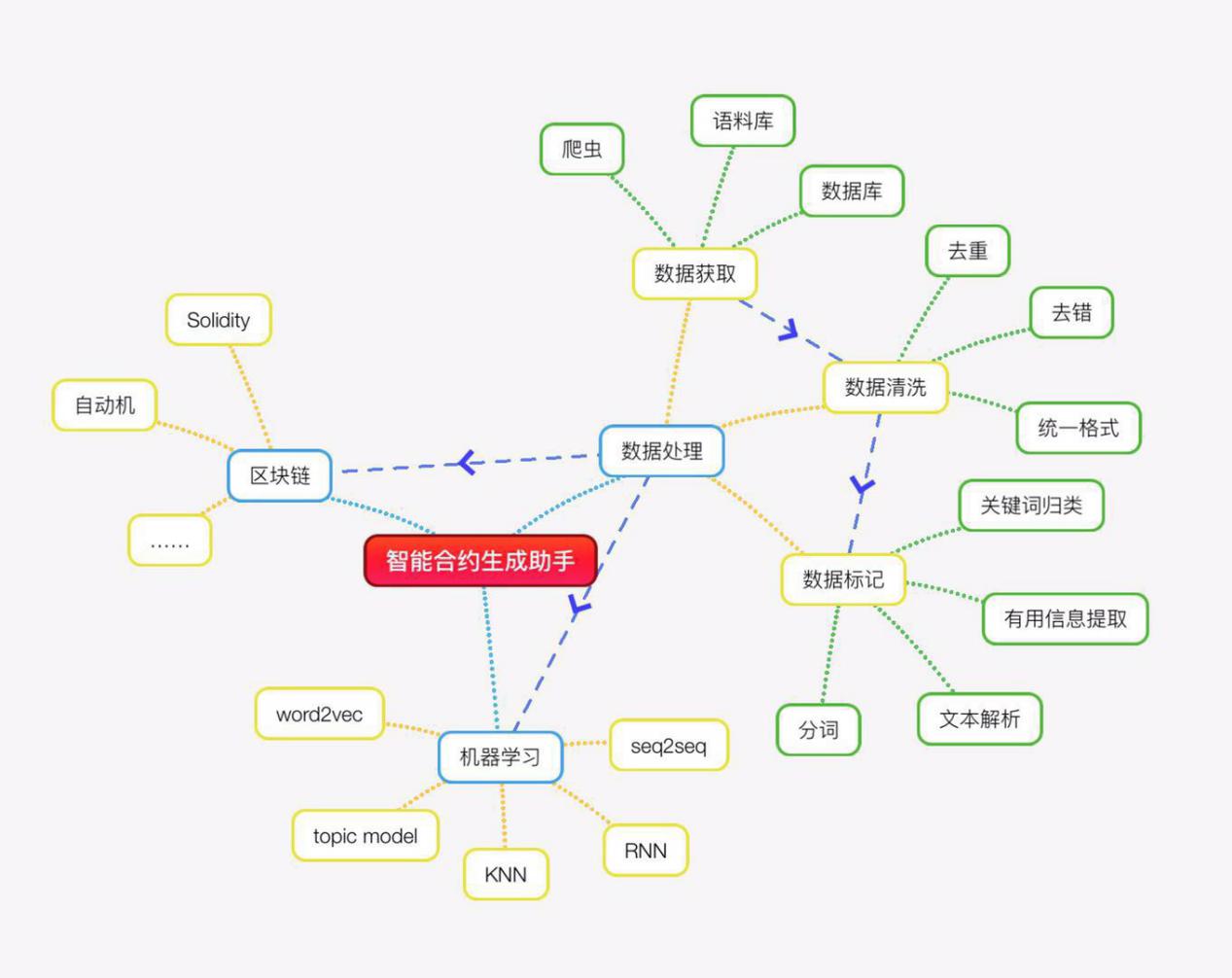
**基于自然语言处理的智能合约自动生成主要有以下几个优点：**

1. **安全性**。普通数据库操作使数据可以被单一数据库所有方保存、修改甚至恶意篡改，不能保障安全性，而引入区块链后，数据分块存储的特性使任何一方都不能成功篡改历史记录，区块链的信息是可信赖的、可靠的。
2. **自动化**。传统就诊流程可能涉及到多方参与协商调度，耗费人力物力，信息收集也难免存在遗漏，引入智能合约后，在医患双方达成协定、行业标准加以规范的情况下，病历信息可以实现部分智能共享，并且就诊过程涉及的信息操作自动化，提高了效率，也大大降低了人为操作的失误。
3. **可视性**。简单的医院数据库不够宏观，病人的个人就诊记录不够具体、准确，而区块链内的信息可以以一种统计的思想从医院视角和患者的多维视角详细列出，信息非常全面、直观可靠，有条理，便于医患双方总结、分析、比较，其支持的数据统计功能、趋势分析功能有利于下一步的决策。

而反观目前的医患现状，开发此题的意义重大：区块链技术支持的智能合约为问责机制提供了可信的保障，一方面对逃避责任的院方有了惩戒依据，一方面也防止了医闹的发生，可以说是双赢的选择。自动化设计避免了不必要的就诊细节、降低了失误率，为更有效、更便捷的就医提供了可能。可视性的信息统计反馈和病例信息共享可以为医患双方的改进和决策保驾护航。

区块链技术将会创造一个连接医药健康产业的新框架，将所有医疗平台的重要数据连接到一起。区块链技术保证了数据的有效性和安全性。对于医疗行业而言，这使得医院、保险公司和实验室能够实时连接并且即时无缝分享信息，而无需担心信息被泄露或者被篡改。区块链能够建立一个数据记录和身份管理的公开标准。一个全球化的医疗健康区块链能够将每个病人包含本地医院和医生的记录信息关联匹配一个ID。基于区块链技术的通用医疗健康ID能够减少患者诊疗过程中的医疗错误并保护病人隐私。

### 二、技术选型



**主要涉及区块链、机器学习、数据处理三方面的技术。**

**实验环境：**

**Python 3.6**

**PyQt5**

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

其中：

#### 2.1 NLP涉及的部分内容

***（1）jieba分词***

jieba分词 是一款基于Trie树结构的中文分词工具。

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。

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；

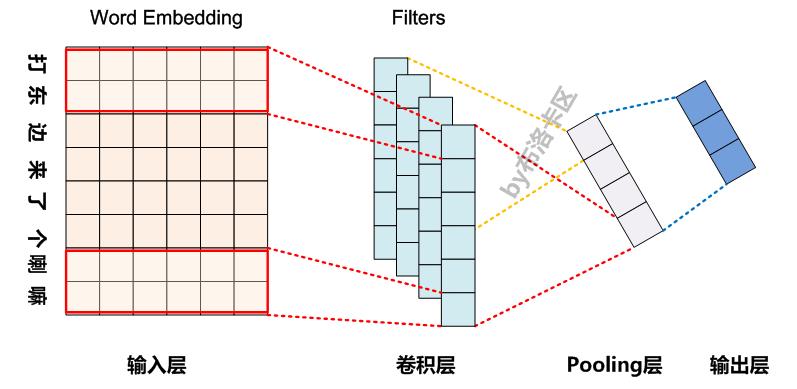
确定前k个点所在类别的出现频率；

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。

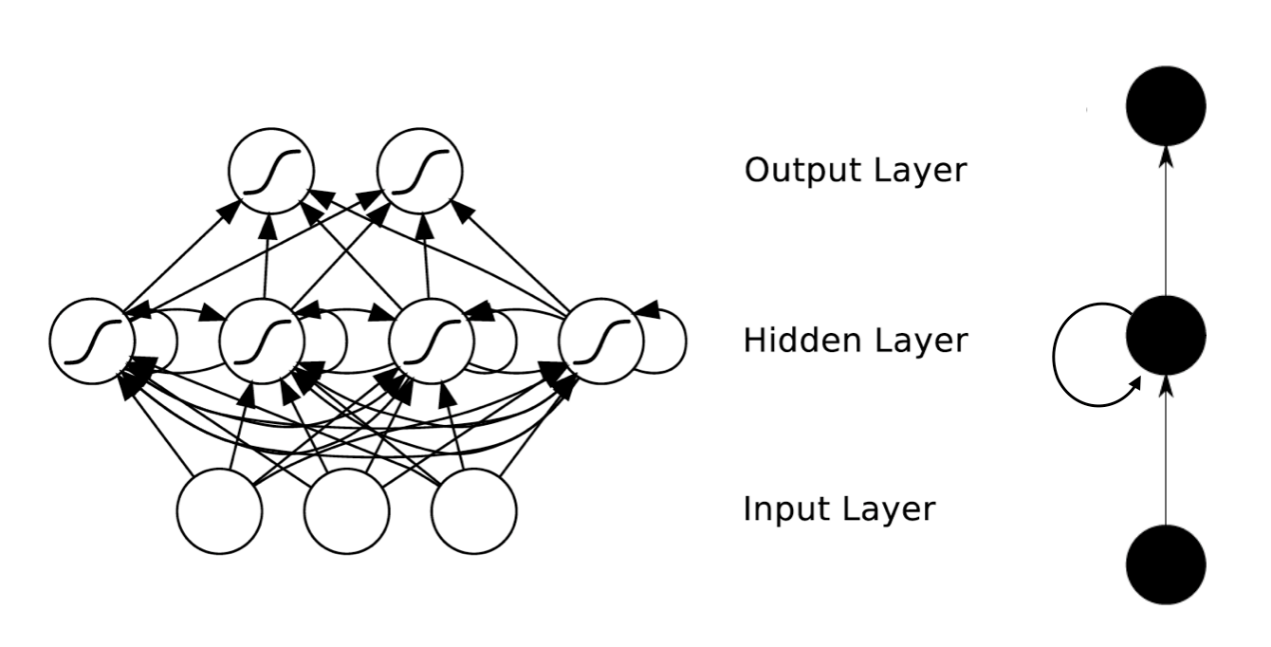
#### 2.2 智能合约可能涉及到的人工神经网络

***（1）卷积神经网络（CNN）***

下图展示了在NLP任务中使用CNN模型的典型网络结构。一般而言，输入的字或者词用Word Embedding的方式表达，这样本来一维的文本信息输入就转换成了二维的输入结构，假设输入X包含m个字符，而每个字符的Word Embedding的长度为d，那么输入就是m\*d的二维向量。

这里可以看出，因为NLP中的句子长度是不同的，所以CNN的输入矩阵大小是不确定的，这取决于m的大小是多少。卷积层本质上是个特征抽取层，可以设定超参数F来指定设立多少个特征抽取器（Filter），对于某个Filter来说，可以想象有一个k\*d大小的移动窗口从输入矩阵的第一个字开始不断往后移动，其中k是Filter指定的窗口大小，d是Word Embedding长度。对于某个时刻的窗口，通过神经网络的非线性变换，将这个窗口内的输入值转换为某个特征值，随着窗口不断往后移动，这个Filter对应的特征值不断产生，形成这个Filter的特征向量。这就是卷积层抽取特征的过程。每个Filter都如此操作，形成了不同的特征抽取器。Pooling 层则对Filter的特征进行降维操作，形成最终的特征。一般在Pooling层之后连接全联接层神经网络，形成最后的分类过程。

***（2）循环神经网络（RNN）***

RNNs的目的使用来处理序列数据。在传统的神经网络模型中，是从输入层到隐含层再到输出层，层与层之间是全连接的，每层之间的节点是无连接的。但是这种普通的神经网络对于很多问题却无能无力。例如，你要预测句子的下一个单词是什么，一般需要用到前面的单词，因为一个句子中前后单词并不是独立的。RNNs之所以称为循环神经网路，即一个序列当前的输出与前面的输出也有关。具体的表现形式为网络会对前面的信息进行记忆并应用于当前输出的计算中，即隐藏层之间的节点不再无连接而是有连接的，并且隐藏层的输入不仅包括输入层的输出还包括上一时刻隐藏层的输出。理论上，RNNs能够对任何长度的序列数据进行处理。但是在实践中，为了降低复杂性往往假设当前的状态只与前几个状态相关，

上图便是一个典型的RNNs

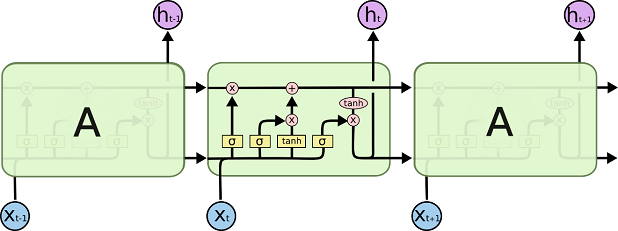
***RNN应用举例：***

***seq2seq(RNN)***

Seq2seq 模型使用 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。

***（3）长短期记忆神经网络（LSTM）***

LSTMs明确设计成能够避免长期依赖关系问题。记住信息很长一段时间几乎是它们固有的行为，而不是努力去学习。所有的递归神经网络都具有一连串重复神经网络模块的形式。在标准的RNNs中，这种重复模块有一种非常简单的结构。LSTMs同样也有这种链状的结构，但是重复模块有着不同的结构。它有四层神经网络层以特殊的方式相互作用，而不是单个神经网络层。

在上面的图中，每条线表示一个完整向量，从一个节点的输出到其他节点的输入。粉红色圆圈代表逐点操作，比如向量加法，而黄色框框表示的是已学习的神经网络层。线条合并表示串联，线条分叉表示内容复制并输入到不同地方。LSTMs的关键点是单元状态，就是穿过图中的水平线。

#### 2.3 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）

### 三．技术难点

1、基于四种基本操作封装函数模块时为保证功能齐全，分类要科学。

2、多维视角中医患用户端权限既有差别又有联系。

3、些网站可能存在反爬虫机制，需要防止IP封锁。

4、自然语言处理。

5、模型训练，数据专业性。

6、比较不同模型的准确度并进行优化。

7、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。

8、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。

9、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。

10、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。

### 开发计划

**4.1 2018年1月29日之前（大约两周时间）**

1. **数据获取**

搜集语料库，利用爬虫获取医疗信息，为之后的病历数据处理做准备。

**（2）后台搭建和基于Excel和数据库的初级测试**

完成对区块链上病历信息的增、删、改、查四种基本函数的设计以及以此为基础的函数模块的划分，之后完成函数类的封装等任务。对数据进行进一步的处理，先将数据调整格式后储存在本地的Excel中，然后再由Excel导入数据库

**（3）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。

**（4）区块链的接口设计**

保留独立有效的区块链接口，便于后期插入。

**4.2 2018年1月29日之后（将近20天的时间）**

**（1）数据处理**

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医学方面的词库，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。

**（2）机器学习NLP相关算法的学习和应用**

深入学习机器学习技术，了解word2vec等算法并进行简单的应用。完成机器学习算法，将NLP和本题病历信息处理结合并调试通过。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。

**（3）接入区块链**

**（4）测试与改进**