# “智能问诊”项目说明

## 16211006 徐家兴

## 15241100 张政勋

## 16211020 许志达

## 16211076 李想

### 一．题目描述

#### 1.1选题背景

当前社会经济的发展以及医疗科技的进步促进了现代人寿命的增长和健康状况的改善，同时随着生活水平和民众健康意识的不断提高，人均期望寿命与健康寿命趋异，使得民众对健康服务消费的需求不断增长，并呈现多层次、多样化的结构特点，这就促使医疗机构由“以疾病治疗为中心”向“以健康促进为中心”的医疗服务模式转变中，大力促进健康管理服务发展。Rock Health《2016年数字医疗消费者取向报告》显示，消费者的对于数字医疗的支付意愿相比2015年增长了三倍，其中39％的消费者强烈同意他们愿意支付医疗费用。这与消费者购买高自付额的医疗保险计划（high-deductible health insurance plans）的增长相似，从2014年的20％上升到2016年的29％（Kaiser家庭基金会数据）。目前国际上关于管理健康需求的主要实践包括：需方管理策略、完善服务供给、医保政策导向等，然而目前我国医疗健康管理流程欠规范、缺乏系统的、动态的健康管理服务，在健康管理服务中尚未发挥应有的作用。

而在技术领域，互联网发展如火如荼，据统计，2014年258家互联网企业获得超过200万美元的投资，单季投资额超过100亿美元。随着移动互联网条件的不断成熟，特别是4G时代的来临，未来智能医疗的应用将更加广阔，更好地应对紧急突发状况。从2014年互联网医疗投资获得最多的六大类来看，我们提出的创意分别与消费者参与、数字化医疗设备、远程医疗、个性化医疗以及健康管理类别相关，是今后重点投资的主方向。

#### 1.2选题内容

“智能问诊”项目旨在以人工智能取代用户在医院问诊的这一过程，通过机器学习，对用户提供的症状进行分析，最后给出相应的诊断乃至用药信息。“智能问诊”不但能帮助用户在就医前对自己的病情进行初步的了解，还能辅助医生对病人进行诊断。

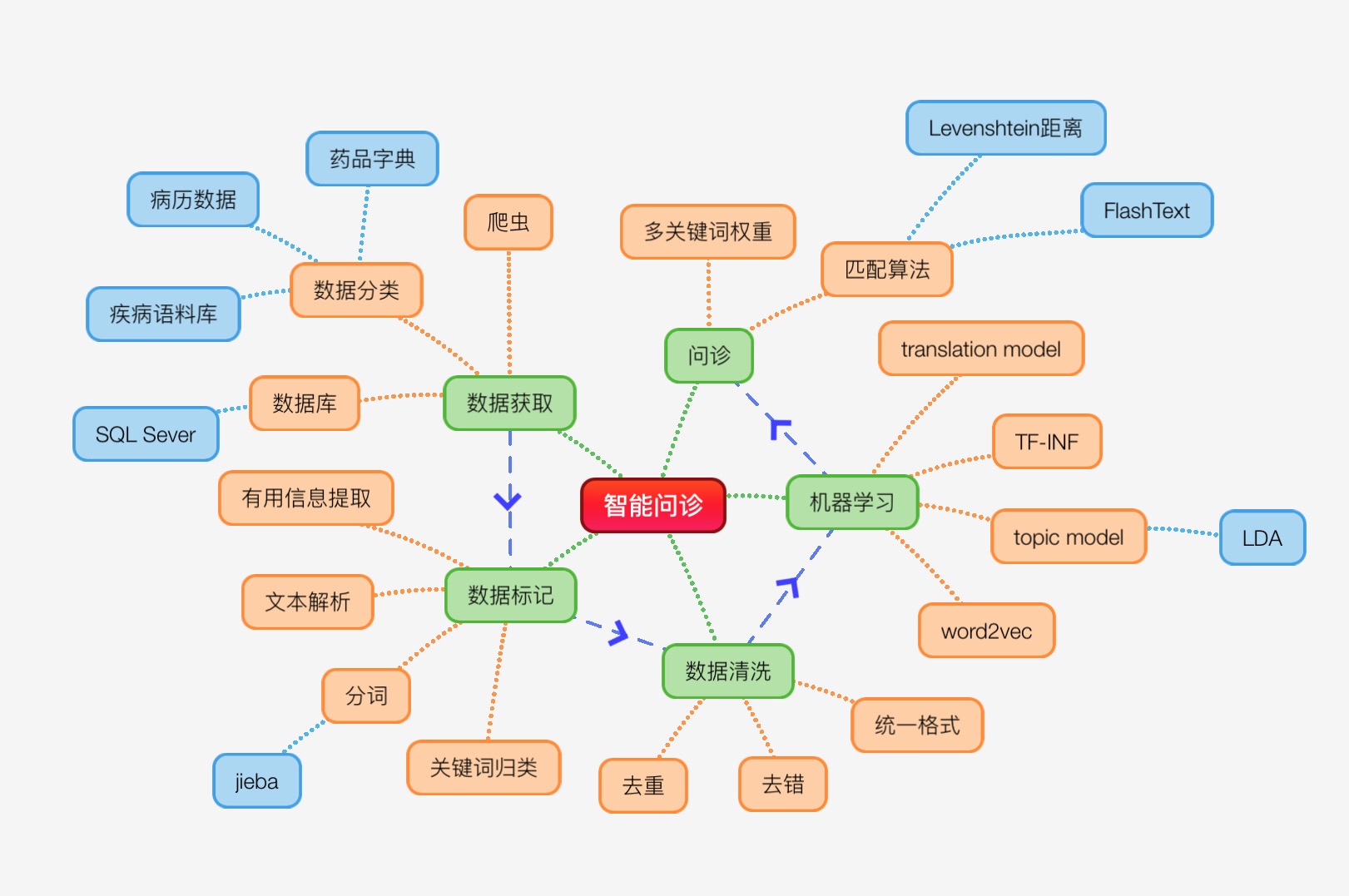
#### 1.3选题价值

早在2016年，习近平总书记出席全国卫生与健康大会时就曾指出，“要把人民健康放在优先发展的战略地位，以普及健康生活、优化健康服务、完善健康保障、建设健康环境、发展健康产业为重点，加快推进健康中国建设，努力全方位、全周期保障人民健康，为实现‘两个一百年’奋斗目标、实现中华民族伟大复兴的中国梦打下坚实健康基础。”

随着“互联网医疗”概念的兴起，如百度“拇指医生”一类的在线问诊网站纷纷出现，但由于我国医疗行业尚无统一的病例、诊断标准，导致各种问诊网站鱼龙混杂，医生资质也参差不齐，而“智能问诊”以机器学习为基础，数据更为可靠，流程更加便捷，与各类在线问诊网站相比有着很大优势。

以上国人身体及医疗管理现状、民众健康意识的提升、科学技术的支持以及国家政策的带动等因素共同促使了“智能问诊”的诞生。

### 二．技术选型



主要涉及爬虫、自然语言处理、数据库、word2vec等技术。

Word2vec

Word2Vec 的基本思想是把自然语言中的每一个词，表示成一个统一意义统一维度的短向量。任何一门语言，都是由一堆的词组成，所有的词，构成了一个词汇表。词汇表，可以用一个长长的向量来表示。词的个数，就是词汇表向量的维度。那么，任何一个词，都可以表示成一个向量，词在词汇表中出现的位置设为1，其它的位置设为0。但是这种词向量的表示，词和词之间没有交集，用处不大。

Word2Vec 的训练模型是具有一个隐含层的神经元网络。它的输入是词汇表向量，当看到一个训练样本时，对于样本中的每一个词，就把相应的在词汇表中出现的位置的值置为1，否则置为0。它的输出也是词汇表向量，对于训练样本的标签中的每一个词，就把相应的在词汇表中出现的位置的值置为1，否则置为0。那么，对所有的样本，训练这个神经元网络。收敛之后，将从输入层到隐含层的那些权重，作为每一个词汇表中的词的向量。比如，第一个词的向量是（w1,1 w1,2 w1,3 ... w1,m），m是表示向量的维度。所有虚框中的权重就是所有词的向量的值。有了每个词的有限维度的向量，就可以用到其它的应用中，因为它们就像图像，有了有限维度的统一意义的输入。

训练 Word2Vec 的思想，是利用一个词和它在文本中的上下文的词，这样就省去了人工去标注。论文中给出了 Word2Vec 的两种训练模型，CBOW (Continuous Bag-of-Words Model) 和 Skip-gram (Continuous Skip-gram Model)。

LDA模型

**1.LDA模型应用：**

LDA可以帮助聚类从而划分一个文档集，主题是聚类中心，文章和多个类簇（主题）关联。聚类对整理和总结文章集合很有帮助。

使用LDA可以帮助判断两种物品的相似性。计算物品的内容相似度时，我们可以先计算出物品在已经构建好的话题上的分布，然后利用两个物品的话题分布计算物品的相似度。如果两个物品的话题分布相似，则认为两个物品相似，反之则认为两个物品不相似。

**2.LDA模型背景**

简称LDA(Latent Dirichlet allocation)中文名为隐含狄利克雷分布，首先于2003年提出，目前在文本挖掘领域包括文本主题识别、文本分类以及文本相似度计算方面都有应用。

LDA是一种典型的词袋模型，即它认为一篇文档是由一组词构成的一个集合，词与词之间没有顺序以及先后的关系。一篇文档可以包含多个主题，文档中每一个词都由其中的一个主题生成。

当我们看到一篇文章后，往往喜欢推测这篇文章是如何生成的，我们可能会认为作者先确定这篇文章的几个主题，然后围绕这几个主题遣词造句，表达成文。LDA就是要根据给定的一篇文档，推测其主题分布。

LDA是一种典型的主题模型，它可以将文档集中每篇文档的主题按照概率分布的形式给出；

同时是一种无监督学习算法，在训练时不需要手工标注的训练集，需要的仅仅是文档集提取出来的关键词以及指定主题的数量k即可；

LDA可以被认为是一种聚类算法：

1.主题对应聚类中心，文档对应数据集中的例子。

2.主题和文档在特征空间中都存在，且特征向量是词频向量。

3.LDA不是用传统的距离来衡量一个类簇，它使用的是基于文本文档生成的统计模型的函数。

**统计式片语翻译模型**

**统计式片语模型应用：**

主要应用在文本翻译中，从一种语言翻译到另一种语言。在跨语言检索的问题上，通常是对查询字词或一段话进行翻译，才能达到翻译的效果。而翻译的结果必须与欲搜寻的文件库有高度相关性，才能达到检索效果。而目前的翻译模型都是一句话到一句话，一个词到多个意思相近的词，如果应用在本实例中，可能不太适合，所以需要稍加改变，只需一句话到多个词即可。

**统计式片语模型背景：**

机器翻译是自然语言处理研究上最重要的课题之一，有助于跨越语言和文化障碍，以前主要用于特定领域文件的翻译，而近年来用于机器辅助翻译和跨语言检索中。在平行语料库为本的统计式做法中，原文S和译文T的翻译概率Pr(T|S),可以分解为下列三个概率函数：

词语翻译概率：

Pr(Tj|Si)

孪生概率：

Pr(a|Si)

位置扭曲曲线：

Pr(j|I,k,m)

其中Si为S的第i个字，Tj为T的第j个字，a为Tj的长度，k为S的长度，m为T的长度。通过反复交替的期望值估计和最佳化演算法，可以得到这三个简单概率函数的统计估计值。期望值估计即在目前的概率函数估计值下，求取所有翻译对应的概率值，而最佳化则是以语料库的对应翻译为根据，估计三个概率函数的最大概似估计值。可以通过贪婪算法，尽可能找到期望值最大的翻译方式。

**levenshtein算法**。

该算法可简化成如下步骤。

|  |  |
| --- | --- |
| 1 | 设置n为字符串s的长度。(“GUMBO”) 　　设置m为字符串t的长度。(“GAMBOL”) 　　如果n等于0，返回m并退出。 　　如果m等于0，返回n并退出。 　　构造两个向量v0[m+1] 和v1[m+1]，串联0..m之间所有的元素。 |
| 2 | 初始化 v0 to 0..m。 |
| 3 | 检查 s (i from 1 to n) 中的每个字符。 |
| 4 | 检查 t (j from 1 to m) 中的每个字符 |
| 5 | 如果 s[i] 等于 t[j]，则编辑代价为 0； 　　如果 s[i] 不等于 t[j]，则编辑代价为1。 |
| 6 | 设置单元v1[j]为下面的最小值之一： 　　a、紧邻该单元上方+1：v1[j-1] + 1 　　b、紧邻该单元左侧+1：v0[j] + 1 　　c、该单元对角线上方和左侧+cost：v0[j-1] + cost |
| 7 | 在完成迭代 (3, 4, 5, 6) 之后，v1[m]便是编辑距离的值。 |

该算法可以实现对关键词的相关度匹配。

其次，我们将该算法应用到工程中，首先，对分出的关键词依次匹配，利用其之前得到的相关度作为其权重因子，将该因子乘以其搜索的权重并相加得到某搜索的最终结果。

之后对每个关键词，分别在药品，症状，疾病名称中进行匹配，并对每一列进行加权综合求权值，通过排序选权值最高的几行作为匹配合适的结果。

### 三．技术难点

### 1．关键词提取不精确

### 虽然jieba分词能很好的从一句话里面提取出关键词，但是并不能保证非常好的提取效果，需要再校检。

### 2．药品与疾病匹配的库不健全

### 目前各种疾病用药信息繁杂，没有一个系统的用药与疾病的库，很难找到一个全面可靠的库。

### 3．准确度问题

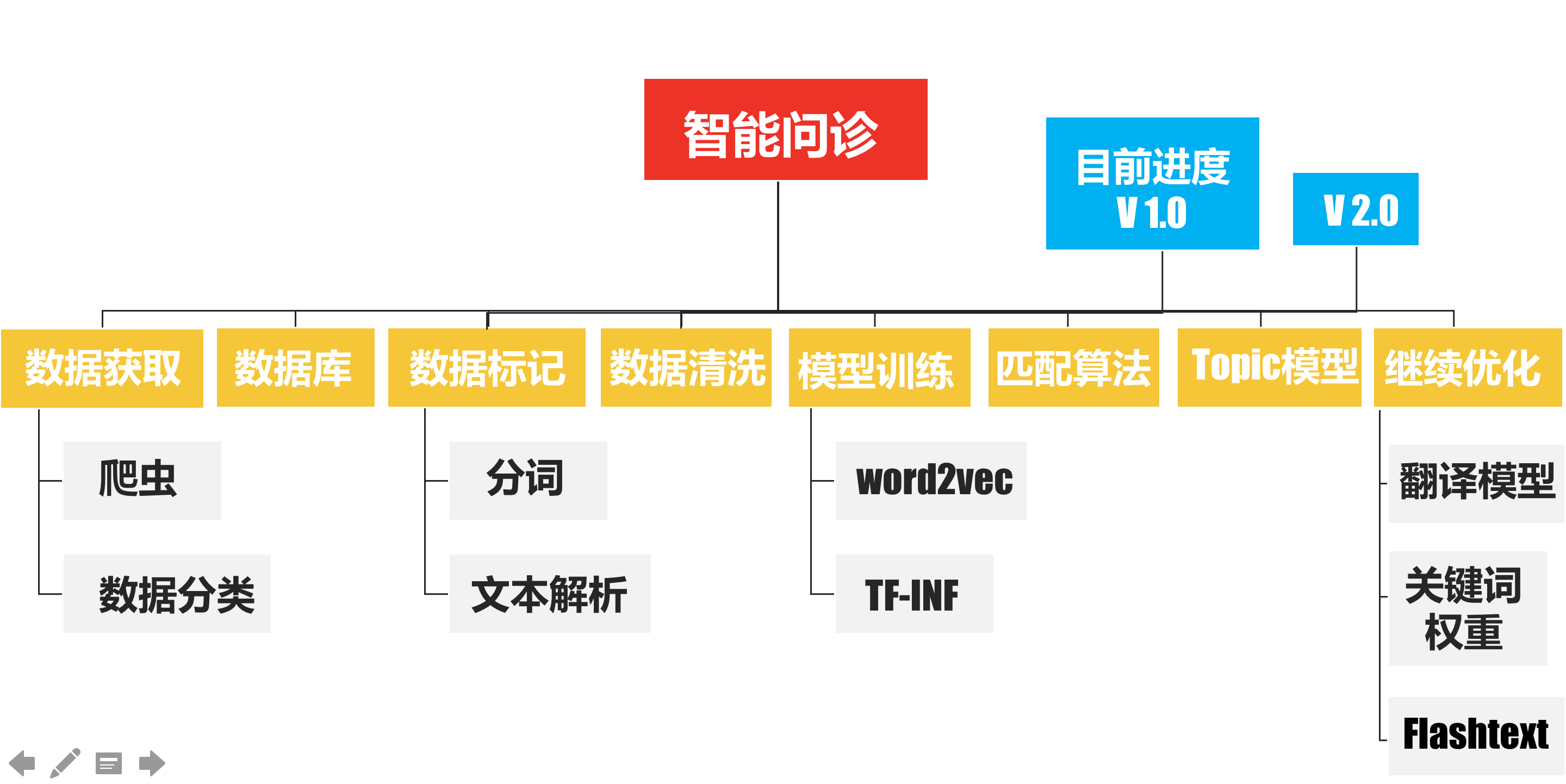
### 词对齐与结构对齐错误累积，可能导致精度不高。一种疾病可能翻译出来的多种药品的作用不一，很难推荐出最适合的。

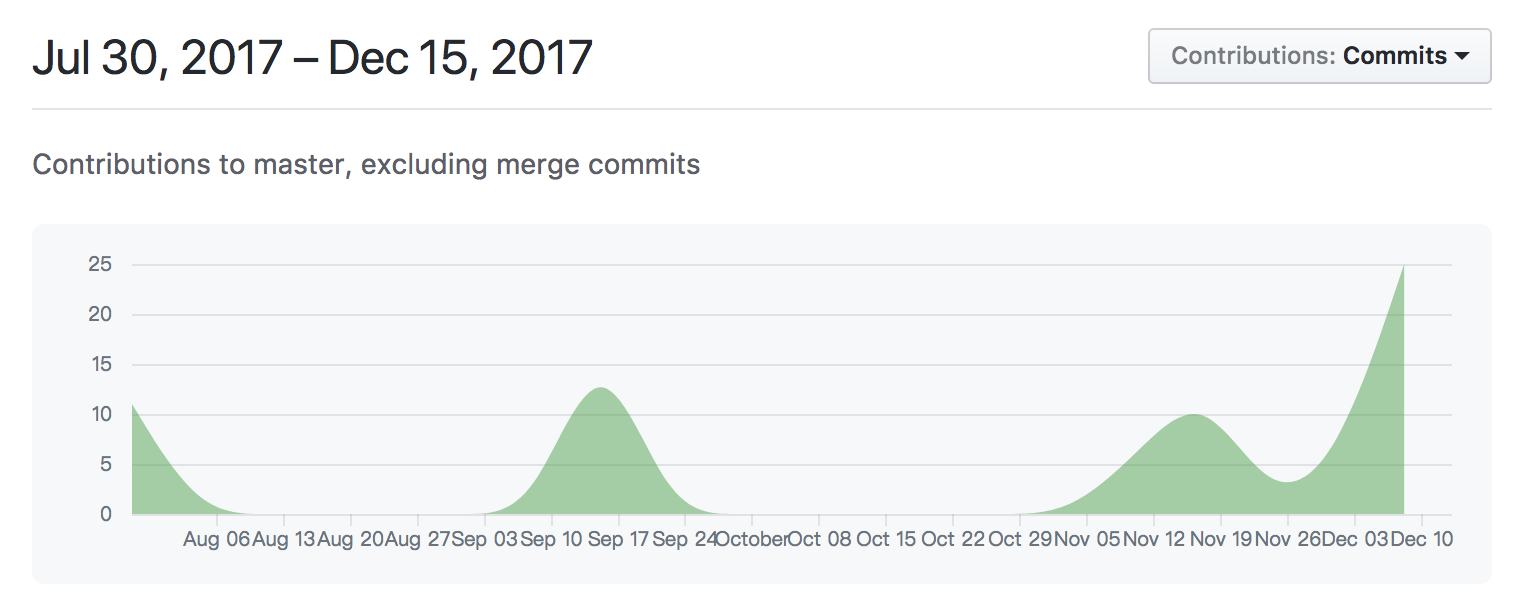
### 四．开发计划

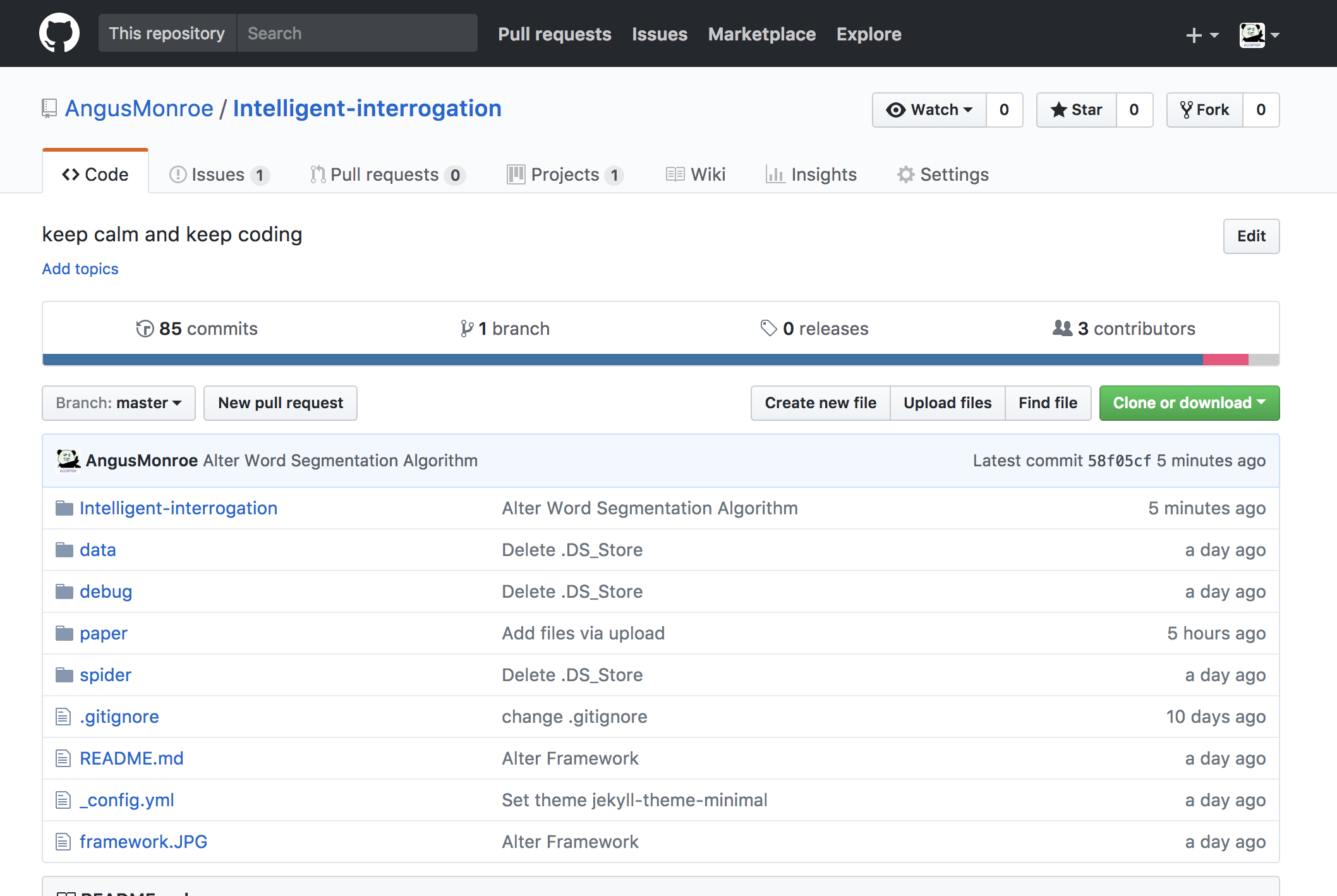
项目分工：



开发进度：







目前本项目的主体部分已经完成并可以使用，接下来我们将对现有模型进行优化：

1. 模型优化

尝试translation、Seq2seq等模型，与现有模型效果进行对比。同时通过增加高质量语料优化现有模型

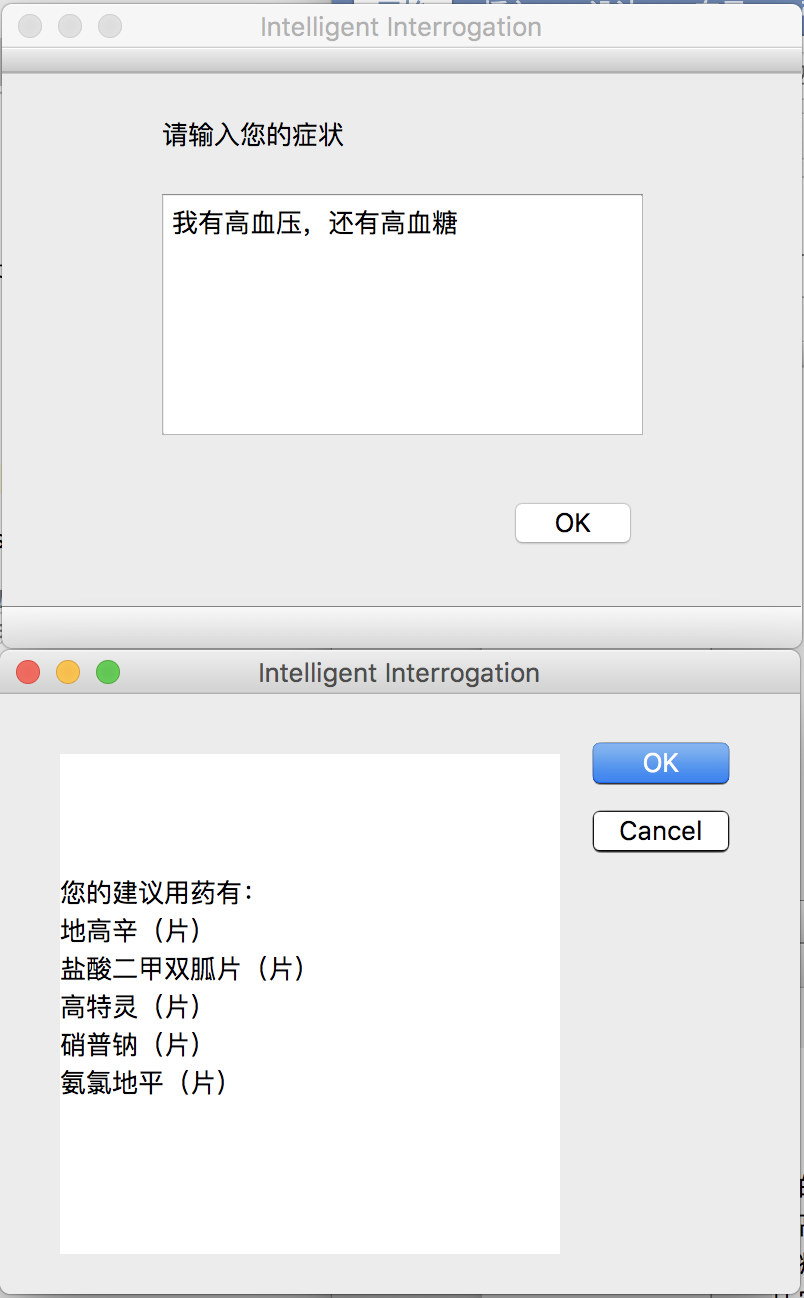
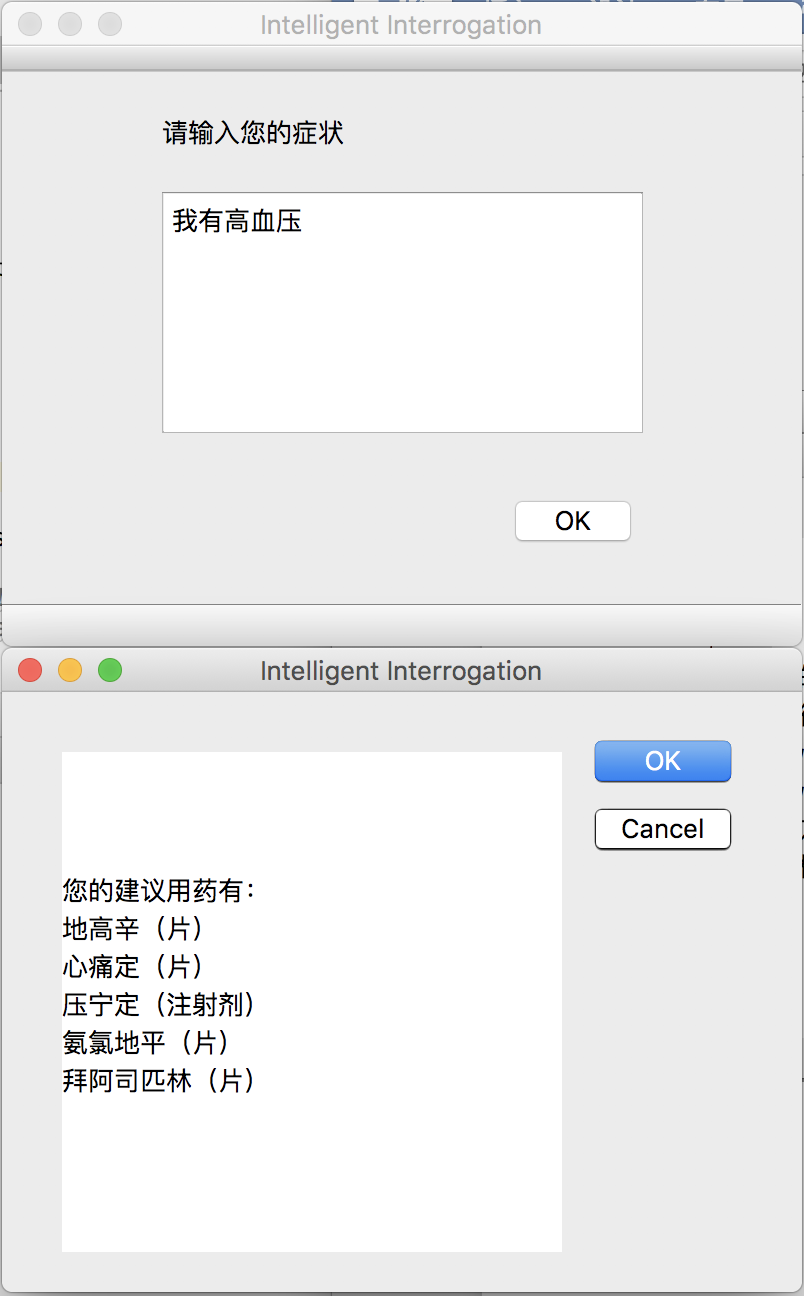
1. 增强对长文本的处理能力

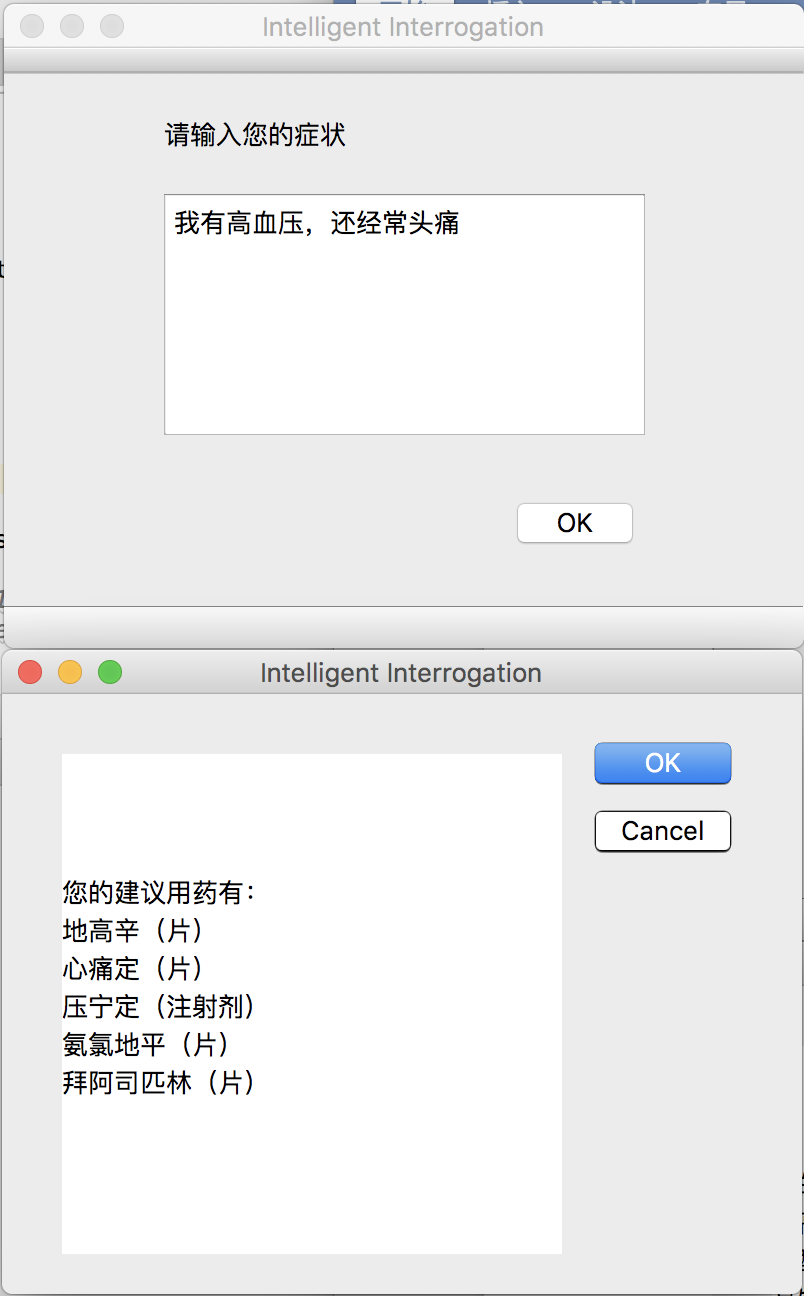
使用topic模型提取出长文本的话题，从而使得本产品可以以大段文本作为输入。

1. 匹配算法优化

目前使用的levenshtein算法 对于搜索来说差强人意，我们希望能在此基础上尝试Flashtext等更加精确高效的搜索算法

截至目前的效果：





可以看出，目前的模型效果仍然不够理想，所以接下来我们要做的主要工作是继续优化模型。

此外，关于数据库中的搜索算法，我们还没有一个明确的想法，在这方面还需要进行一下研究。