# “智能问诊”项目说明

### 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。

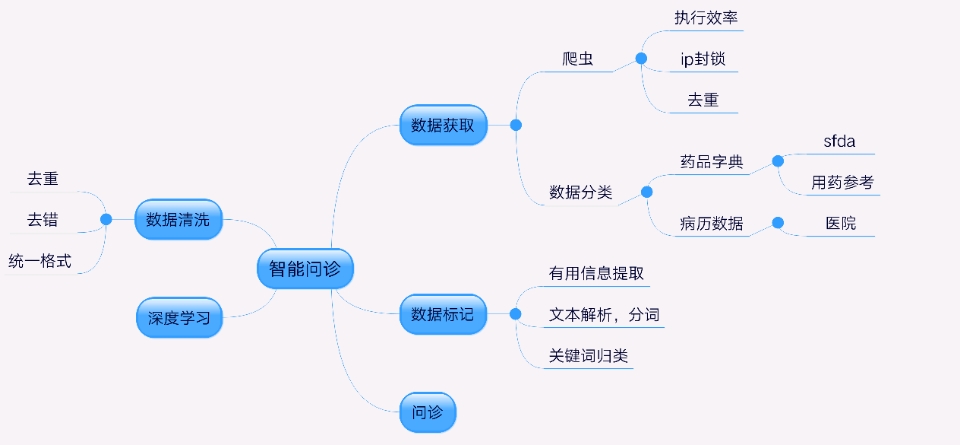
#### 1.3选题价值

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。

具体技术细节及开发日志见“开发技术”部分的博客链接。

### 三．技术难点

1.某些网站可能存在反爬虫机制，需要防止IP封锁。

2.自然语言处理。

3.模型训练，数据专业性。

4.使用合适的字符串复杂度算法并结合项目进行优化

### 四．开发计划

截至目前的进度及成果：

7.20-7.21 利用爬虫从上获取药品的名称、适应症等信息保存到本地，为数据处理做准备。

总结： [“智能问诊”项目——数据获取（1）](http://blog.csdn.net/angus_monroe/article/details/75451886)。

7.24 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。

7.25 将储存在Excel中的数据上传至数据库。

总结：[“智能问诊”项目——数据获取（2）](http://blog.csdn.net/angus_monroe/article/details/75966525)。

7.26 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。

总结： [“智能问诊”项目——数据处理（1）](http://blog.csdn.net/angus_monroe/article/details/76066583)、

[python\_jieba分词的使用。](http://blog.csdn.net/angus_monroe/article/details/76316427)

7.27-7.28 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。

总结： [“智能问诊”项目——数据处理（2）。](http://blog.csdn.net/angus_monroe/article/details/76269587)

7.31 对决策树进行初步的了解。

总结：[机器学习笔记（1）——决策树。](http://blog.csdn.net/angus_monroe/article/details/76397179)

8.1-8.3 了解word2vec算法并进行简单的实用。

8.4-8.7 完成机器学习算法，调试通过。

8.8-8.10 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。

8.11-8.12 决定使用合适的字符串相似度算法，并应用于工程中。首先，我们的思路是寻找和百度搜索引擎类似的算法，但是通过实践发现，类似算法的时间复杂度无法接受，所以我们试图寻找一个足够简化并且时间复杂度可以接受的算法。通过比较我们最后确定了**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。

12.10 添加topic模型的LDA算法实现对长文章的主题分类，从而达到根据已经训练好的主题类，对新添加的文章划分归类的效果。

下面简要介绍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。

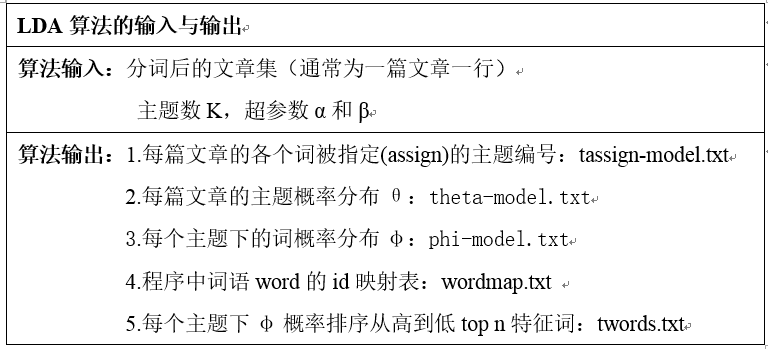
3.LDA不是用传统的距离来衡量一个类簇，它使用的是基于文本文档生成的统计模型的函数。

**3.LDA模型优势**

LDA在处理长文本信息时准确性较好。

LDA的另一个优点则是，对于每一个主题均可找出一些词语来描述它。

**4.LDA生成模型使用方法**

****

**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。

**3统计片语式模型优势**

可以很好地应用于本例中，而且不用过多考虑表达方式，都是从中文到中文，只需提取关键词，然后搜索对应的目标词汇即可，这些都可以通过加以训练提高精确度。

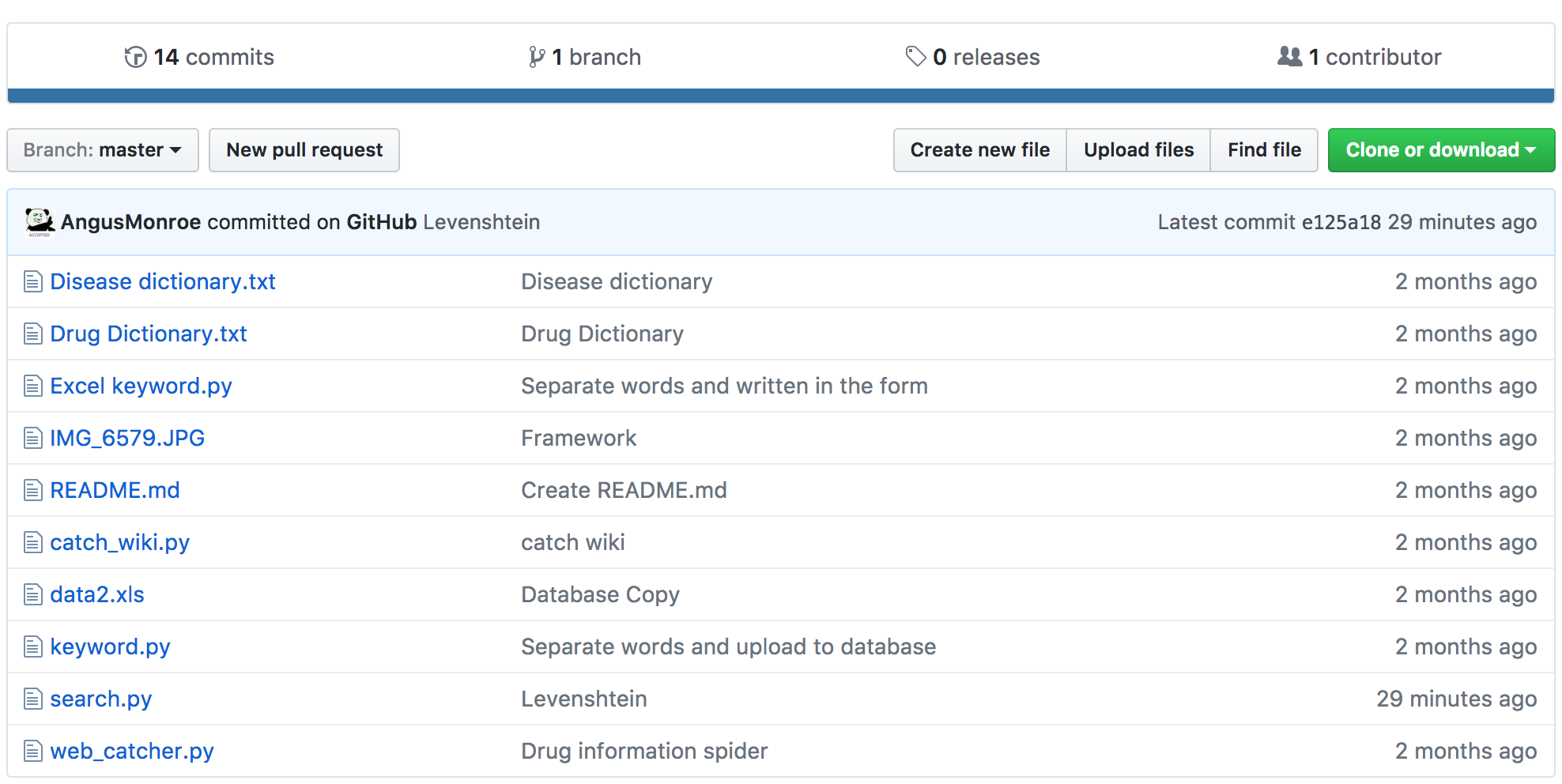
**4统计片语式模型使用方法**

1 提取关键词，获得源文库和对应的译文库

2 通过期望值估计和最佳演化法，尽可能得到最大的期望值。

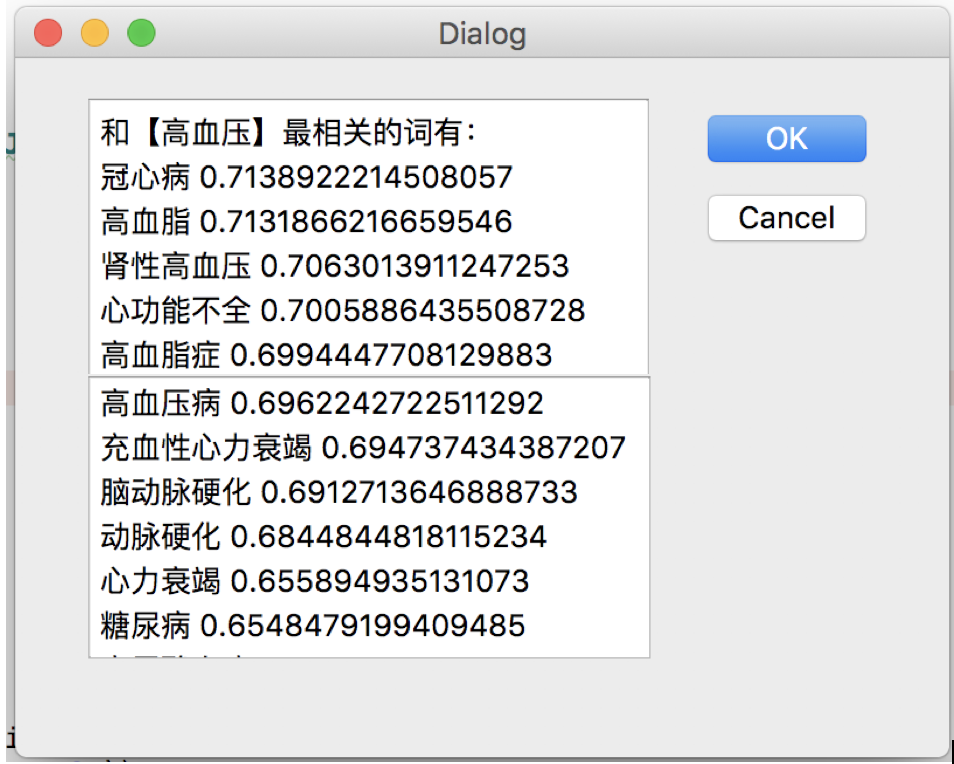
3 反复训练模型，逐步改正并优化算法





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。