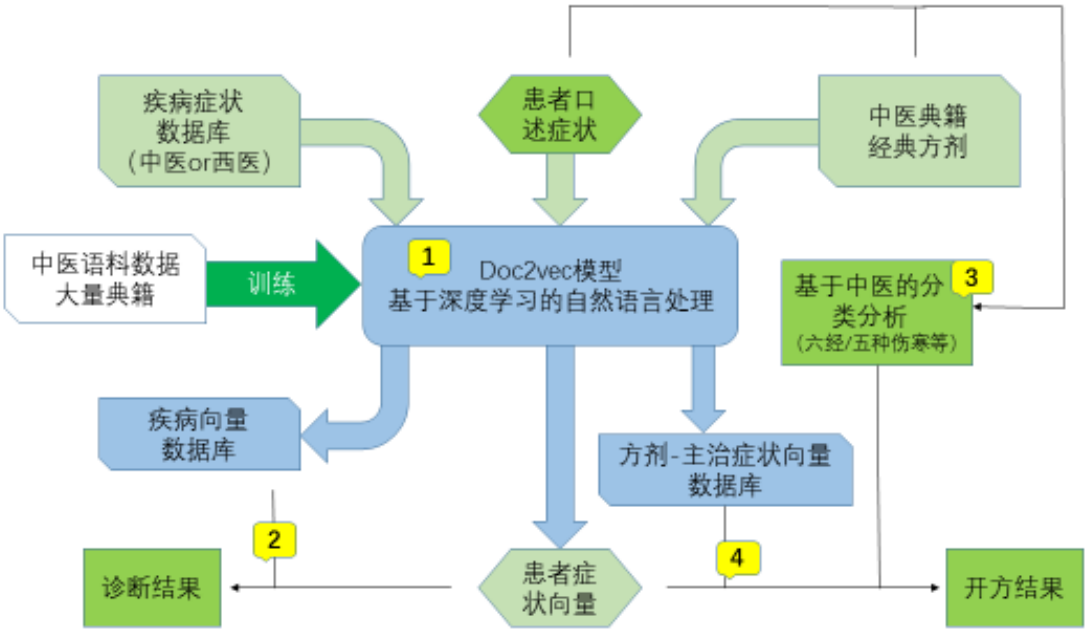
doc2vec for 中医 第一版本

第一版本的基本框架：
（这个图是为了组内交流临时画的草图 暂时没有考虑美观 如果要做报告用的话我重新找模板画一下）



说明：

anagraph.npy 存储方剂主治向量库
diseases.npy 存储疾病症状向量库
model.4.0.1.md 初步训练的 doc2vec 模型
result.model 初步训练的朴素贝叶斯疾病分类模型

diseases_out_description.txt 目前暂时使用的疾病症状数据
yaofang.txt 从《伤寒杂病论》中提取的结构化方剂数据
（这两个 txt 文件不要修改，否则会影响程序运行）

将患者主述症状读取入 input.txt 中，运行程序 zy_5.0.0.py，即可在 output_1.txt 中输出基于疾病症状数据库的诊断结果（默认输出前 5 位），在 output_2.txt 中输出基于典籍的中医疾病分类和开方结果（默认输出前 5 位）。

1 关于 doc2vec 模型训练

目前使用的训练数据是《黄帝内经》、《难经》、《伤寒杂病论》、《神农本草经》、《圣济总录》和疾病数据库中的所有描述，总字数大约 60 万。在目前字数下调整 doc2vec 参数对结果的影响并不显著。
doc2vec 存在一个“天生”缺陷就是对于短句（大概 10 字以下）的理解能力很差，短句主要出现在古书典籍中（因为很多文言句子往往只有寥寥几个字），这部分有可能还需要做关键词扩增。
在这个模块，下一步工作是扩充 doc2vec 训练数据。

2 关于疾病症状数据库以及基于该数据库的诊断

这部分数据库中目前存在很多“非传统中医”的疾病，例如“一氧化碳中毒”“雷诺氏病”等。
我个人认为这部分用西医的疾病症状也并非不可，因为西医对于疾病的系统、细致性其实有着自己独特的优势，而且西医的诊断结果也可以与后面中医的分类、开方结果形成参照。（是这样吗？）
在这个模块，下一步工作是完善、扩充并且系统化疾病症状数据库。

3 关于症状分类

症状分类的初衷其实是在于对《伤寒杂病论》的阅读中，发现该书的思路是“辨”+“治”，而且辨别得到疾病大类之后，还需对于各种因人而异的不同症状针对性开方。例如，“太阴病”可以通过“腹满而吐，食不下，自利益甚，时腹自痛，若下之必胸下结鞕”判断，而其下对于“脉浮而缓，手足自温者”、“大便反鞕，腹中胀满者”等不同具体症状分别进行了药方的说明。因此可以先基于口述症状和典籍中记载的不同类型病症的症状进行初步的分类。
在这个模块中我也存在一些疑惑，主要在于伤寒杂病论对于疾病的分类上。伤寒分为五种，即中风、伤寒、湿温、热病、温病，而《伤寒杂病论》也分别用一个章节讲述了这五种类型疾病对应的方剂。但同时其又基于外感疾病发展、变化过程中产生的各种证候分出六经病，即太阳病、阳明病、少阳病、太阴病、少阴病、厥阴病，而且分别设章节讲述药方。这些分类是否存在交叉呢？如果存在交叉的话《伤寒杂病论》中又是怎样将方剂放入不同章节中的呢？如果有交叉的话那这一版本程序的疾病分类就可能存在不妥之处，需要改进。
分类目前使用的是朴素贝叶斯方法，下一步可以尝试训练神经网络以进行改进。

4 关于开方

在分类确定后，可以在该分类的药方中进行检索。目前最主要的问题是提取药方的方式，对于《伤寒杂病论》中药方数据的结构化工作，很大一部分是我纯手工完成的，比较耗费时间，以后有更多数据时很难推广，所以急需要一些结构化的数据。

此外，之前提到过的文言短句的扩增也是需要解决的问题。下一步需要学习相关的自然语言处理技术进行改进。
第一版本的开方基本上是基于检索实现的，在下一阶段的工作里，将会尝试构建开方的神经网络，实现对于主药、辅药的预测（类似由电脑产生新的方剂）。