## 1. 采用的方法是基于模板+字典匹配的方法

如果不需要字典匹配可以找到NER.java中相关部分注释掉就可以），因此属于基于规则的电子病历命名实体识别方法。

## 2. 命名实体抽取的粒度不是很细

根据网上相关的这方面的论文（在线医疗文本中的实体识别研究.pdf），也基本是基于这四个命名实体做的，所以没有增加其他的命名实体。

## 3. 文件说明

**3.1. 源数据**

医疗专业相关实体的字典是网上找的，分别是category，relation，reserved三个文件中的数据，做了一定的数据清洗工作，生成了0NER（症状)，1NER（药品），2NER（疾病），3NER（检查）和4NER（比较杂的一些医学实体词汇），这部分工作在KeyWord.java中实现。

一种投机的办法就是，遇到识别错的，即未识别成为正确的对应实体，那么把这个词加入到字典文件中，就可以正确识别了。

**3.2 traindata**

traindata文件是提供的电子病历文本，通过调用第三方的分词工具，Bosonnlp在线分词，感觉这个分词精确一些（seg.py）,注意这是python程序，如果你需要用需要去申请自己的API token密令，免费的。上bosonnlp.com申请，把你自己的token填入seg.py的33行your token处。

有一点点麻烦，先要装python环境，python3最好，然后在命令行输入pip install requests，安装好requests组件，再命令行运行python seg.py即可

为什么选用bosonnlp分词：在线工具，背后有大数据支撑，相比固定的开源分词工具包具有更好的扩展性，更专业的中文分词工具，号称是综合考虑了多种因素得到的分词结果。

**3.3 trainseg.txt**

为得到的分词结果文件，命名实体提取部分都是基于这个文件的数据的。

**3.4 test.txt文件**

测试用的单个病历文件，在main函数处修改输入测试文件即可，在用户界面中，必须选择txt文件，否则无法处理！

**3.5 GUI程序**

使用Java AWT + Swing做了简单的用户界面，双击jar文件直接运行，但是要注意，把字典文件（0NER~4NER）和一些图片文件（主要是用户界面上的图标）需要在**同一目录**下方可正确运行。中文如果出现乱码，需要正确设置测试文件text.txt的编码为utf-8格式。（使用notepad++文本编辑器，格式菜单，转化为UTF-8无BOM编码格式）

可执行文件在DianZiBingLiNER\out\artifacts\DianZiBingLiNER\_jar目录下！

## 4. 项目说明

Java项目，Main.java是主程序的入口。

可以用Intellij IDEA直接导入这个工程，如果用Eclipse的话，把src下的代码拷贝进去，其他相关的数据文件也放在项目目录中（如果只是用命名实体识别部分，一个trainseg.txt加上字典文件0NER,1NER,2NER,3NER,4NER就行了）。

## 5. 命名实体识别测试

5.1

在程序中提供了几个小片段，结果直接输出在命令行了，效果还可以。药品方面的识别效果比较好。命名实体识别的效果评价，主要是和方法有关，这里主要是和医学名词字典和我们制定的规则有关的，毕竟不是医学专家，不能写出完美的规则来匹配所有的命名实体。

5.2

测试方法：需要调用bosonnlp的接口，或者直接在他们的网站上在线分词，原始文本需要用bosonnlp的在线工具分词，或者用python程序直接得到结果。得到分词结果作为文本输入再进行命名实体识别的流程

5.3

当然，为了展示你也可以自己写一些电子病历文本，自己标注一下，然后进行命名实体识别。提供的病历有些杂乱，所以推荐自己稍微修改一下，然后去bosonnlp的网站上分词，存一下再去做测试

5.4

字典是否要采用可以自己调整，虽然会引入一些不是我们需要的实体，但是个人认为全面些好。如果发现这些词典中有些不合适的，直接删掉就可以了。目前注重提高recall，宁可错杀，不可漏杀的思想。

## 6. 为什么采用这种方法

首先，我们没有标注好的数据，开源工具（Opennlp，HanNLP)和命名实体识别常用的CRF工具，都要求有标注好的训练数据，以及未标记的测试数据，但是时间仓促，没来得及标注这么多的数据，而且提供的数据质量也不是很好，都是儿科的病历，有许多重复的内容和无关内容，范围覆盖不够。

这样做的**优点**：不需要标注数据，比较节省人力物力财力，而且命名实体识别的速度很快。如果有新的规则，也可以立即添加进来，从而进一步提高准确率。

**缺点**：准确率一般，而且与编写的规则有很大关系，与人为设计的规则质量好坏直接相关，人工的影响比较大，如果有医学专家参与应该能取得更好的效果。

## 7. 结果分析

综合来说，效果一般，对于药品类的实体识别效果较理想，症状、疾病这两种实体之间有一些交叉，所以召回率还可以，准确率（precision）较低，这也是采用基于词性的规则匹配带来的许多杂质。检查这一实体算是一般吧，主要是基于关键字的规则匹配，问题在于命名实体的边界不好确定十分准确的规则。

具体的原因已经在qq上说过了。其实最关键的三点：分词本身不会完全正确，基于规则的方法需要较好的规则来区分实体的边界，数据本身的问题（无关内容，自然语言的歧义，医学专业性较强，医疗电子病历本身的语言组织多以短句、缩句为主，风格各异，不算是十分正式、规范的文本）

## 8. 代码部分都写了一些注释和说明，有什么不理解的可以再问，也可以提供远程帮助。