## 1. 基于百度知识图谱（全科）的卒中知识补全

### 1.1 功能概述

     在医学知识图谱的构建过程中，自动识别**非结构化数据**（中文电子病历）中的**实体**，实现**实体对齐**，并将相关实体链接或补全到知识图谱的相应节点上；识别病例中的**关系**，并补全到已有知识图谱的关系中；最终**统计**并展示补全效果。

### 1.2 功能要求

1. **医学命名实体识别**

基本要求：从非结构化文本中识别并抽取出脑卒中病例中的相关实体。必需识别的实体包括疾病、症状。

可拓展点：识别药品、检查、手术等（具体参考百度知识图谱的schema）。

举例：

文本：“脑血管病患者3天前无明显诱因出现头晕不适,伴行走不稳”

抽取实体：“脑血管病”（疾病）、“头晕不适”（症状）、“行走不稳”（症状）

1. **实体对齐**

将同一实体的不同描述合并到一起（同义词识别）。

举例：

“脑卒中”《===》“中风”《===》“cerebral stroke”

“人工牛黄甲硝唑胶囊”《===》“牙痛安”

1. **关系抽取**

基本要求：从电子病历文本中识别出“疾病-症状”关系，并计算关系权重（统计共现概率）。

可拓展点：识别“疾病-药物”、“疾病-手术”、“疾病-检查”等关系（具体参考百度知识图谱的schema），并补全到已有知识图谱中。

举例：

文本：“脑血管病患者3天前无明显诱因出现头晕不适,伴行走不稳”

抽取实体：“脑血管病 - 头晕不适”、“脑血管病 - 行走不稳”

### 1.3 验收要求

1、实现非结构化数据中医学实体的识别抽取

2、对齐医学实体

3、实现非结构化数据中医学关系的识别抽取

* 1. 病例中已识别出id、性别、年龄、职业、主诉、现病史、既往史、个人史、家族史、体格检查、辅助检查、入院诊断、出院诊断、入院情况、出院情况、诊疗经过、出院医嘱、病程记录等字段。要求对每一份病例保留一份文件，包括id、性别、年龄、职业，以及识别出的医学实体和关系。
  2. 需要将识别出的实体、关系对齐后补全到百度全科知识图谱中，并统计每种实体、关系的数目。

### 1.4 技术参考

了解知识图谱构建的基础知识，可阅读Nature论文[1]。

1. 实体识别与抽取

**a. 直接与已有的知识库匹配的实体抽取方法**

**b. 基于规则的实体抽取**

**c．基于条件随机场（或其他机器学习方法）的实体抽取**

**d．基于深度学习的实体抽取**

2、 实体对齐

**a. 基于词典或语言学模式的实体对齐**

“复方乙酰水杨酸片”又称“阿司匹林” ——实体

**b. 基于支持向量机（或其他机器学习方法）的实体对齐**

训练语料（百度全科知识图谱） -> 特征选取 -> 模型训练

**c．基于深度学习的实体对齐**

3. 关系抽取

基本要求参考Nature论文[1]即可。

### 1.5 参考资料

所需的百度全科知识图谱、脑卒中病例等资料确认选题后提供。

[1] Rotmensch M, Halpern Y, Tlimat A, et al. Learning a Health Knowledge Graph from Electronic Medical Records.[J]. Scientific Reports, 2017, 7(1).