《自然语言处理》课程作业

**课程编号**： U10M11164

**课程属性**： 专业选修课

**学时/学分**：32/2

**预修课程**：线性代数、微积分、数学优化、概率论、信息论

**一、作业目的：**

通过本课程作业加深对自然语言理解基础理论的认识和了解，锻炼和提高分析问题、解决问题的能力。通过对具体项目的任务分析、技术调研、数据准备、算法设计和编码实现以及系统调试等几个环节的练习，基本掌握实现一个自然语言处理系统的基本过程。

**二、作业题目：**

**1．实现一个面向语义相似度匹配的临床术语标准化方法**

临床术语标准化方法是医学统计中不可或缺的一项方法。临床上，关于同一种诊断、手术、药品、检查、化验、症状等往往会有成百上千种不同的写法。标准化（归一）要解决的问题就是为临床上各种不同说法找到对应的标准说法。有了术语标准化的基础，研究人员才可对电子病历进行后续的统计分析。本质上，临床术语标准化方法也是语义相似度匹配任务的一种。但是由于原词表述方式过于多样，单一的匹配模型很难获得很好的效果。本次评测任务主要目标是针对中文电子病历中挖掘出的真实诊断实体，并以《国际疾病分类 ICD-10 北京临床版v601》词表为标准进行语义标准化。在给定一诊断原词，要求给出其对应的诊断标准词。以(诊断原词，标准词)作为基本单位计算Micro-F1得分。如测试集有m对(诊断原词，标准词)组合，预测了n对(诊断原词，标准词)组合，有k对组合是预测正确的。

P = k / n

R = k / m

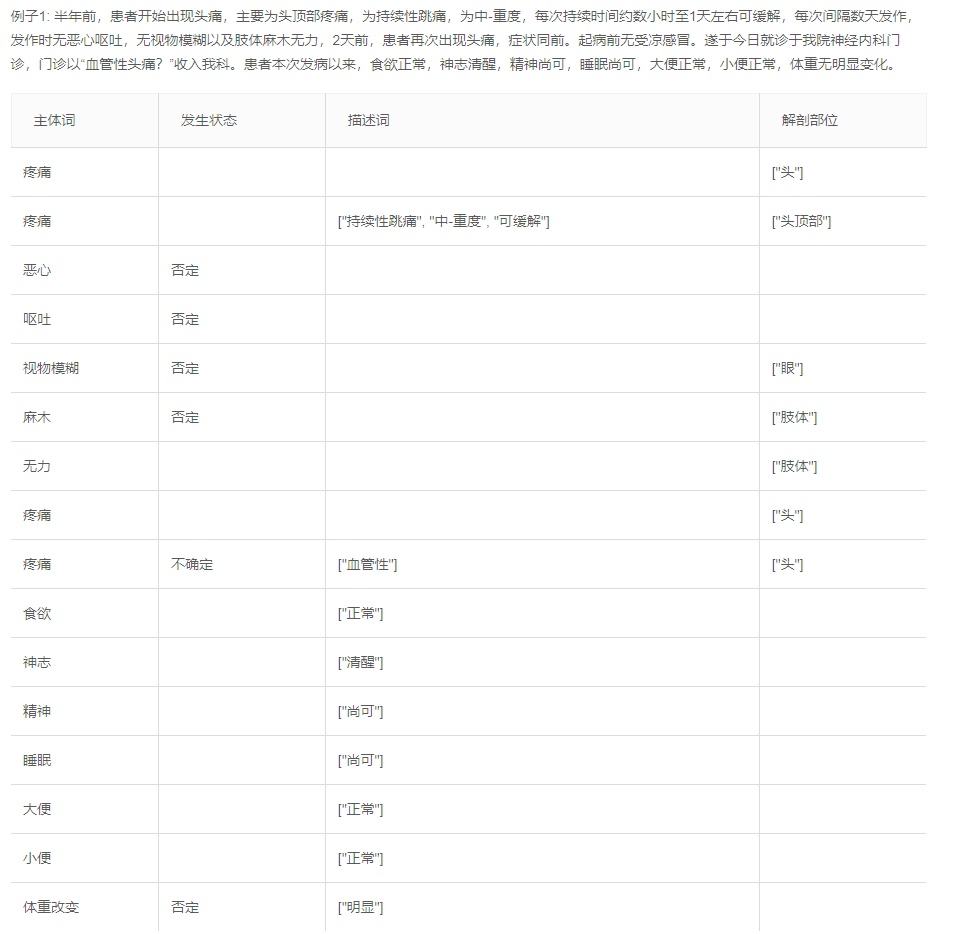
F1 = 2 \* P \* R / (P+R)

样例如下所示：



**2． 实现一个面向关联关系发现的临床发现事件抽取方法**

临床发现指的是疾病的表现，泛指患者不适感觉以及通过检查得知的异常表现，主要包括症状、体征。临床发现事件抽取是医学数据处理中的一项任务，本题目要求从病历中抽取例如临床发现事件的“解剖部位、主体词、描述词，以及发生状态”四个维度的属性，需进行实验分析。以Micro-F1为评测指标，算分逻辑以每个四元组为统计粒度，要求四元组的每个属性值完全匹配才算正确。样例如下所示：



**三、基本要求：**

(1) 每1-2人可以组成一组选择其中的任何一个题目，鼓励创新方法并结合一定的理论根据或实验数据依据。完成一份技术报告，报告内容包括：项目目标、国内外相关工作、自己在本项目中承担工作的不同点、实现系统（或模块）的核心思想和算法描述、系统主要模块流程、实验结果及分析(具体内容参照课程作业模板的word文档)。

(2) 提交系统源代码和可执行程序，以保证实验系统可以正常编译和运行。

(3) 2023年10月25日（北京时间 23:59）之前提交课程大作业报告。

**四、严正声明：**

(1) 鼓励充分使用网络资源和其它一切可以利用的资源（包括数据、语料、软件工具和论文资料等），但严禁侵害他人知识产权，技术报告中必须明确说明所用资源的真实来源，否则本课程作业计零分。

(2) 鼓励相互交流、相互合作，但严禁抄袭他人工作，严禁伪造结果。