**Task: 问题命名实体识别和链接**

1. **任务描述**

本评测任务为问题的实体发现与实体链接，简称QEDL （Question Entity Discovery and Linking）。即对于给定的一组自然语言问题，识别并抽取出与领域相关的实体名称（mention），并将它们链接到给定知识库对应的实体（entity）。实体名称具有歧义性和变异性，也就是同一个实体名称，有可能指代多个实体，需要根据上下文消歧；此外，同一个实体可能有多个实体名称与之对应，比如别名，绰号，昵称等等，需要将这些所有的名称变种都识别出来。

**输入**

输入一段文本。

**输出**

输出文本中的实体名称mention的列表，以及映射到知识图谱的实体entity的列表，列表内元素以|||分隔，两个列表间以\t分隔。若知识库不包含实体名称的对应实体，则该对应实体输出为nil。两个列表的元素必须保持对齐。

**示例：**

输入：

李娜是在哪一年拿的澳网冠军？

cnc手柄不亮了怎么办

输出：

李娜|||澳网\t李娜（中国女子网球名将）|||澳大利亚网球公开赛

cnc手柄\tnil

说明：“李娜”是一个歧义性名词，通过知识工场mention2entity接口，它对应如下若干实体。而在本问句中，显然对应的是“中国女子网球名将”的实体。另外通过知识工场mention2entity接口，可知“澳网”是实体“澳大利亚网球公开赛”的缩写。



对于实体名称“cnc手柄“，mention2entity中并没有相关实体。所以链接结果输出nil。

1. **评价指标：**

本任务的评价指标包括实体名称识别层面的F1，以及实体链接层面的F1。

实体名称识别准确率=标注正确实体名称个数/标注出的实体总数

实体名称识别召回率=标注正确实体名称个数/总的实体名称个数

$$实体名称识别F1=\frac{实体名称识别准确率\*实体名称识别召回率}{实体名称识别准确率+实体名称识别召回率}$$

实体链接准确率=标注正确实体链接个数/标注出的实体总数

实体链接召回率=标注正确实体链接个数/总的实体个数

$$实体链接F1=\frac{实体链接准确率\*实体链接召回率}{实体链接准确率+实体链接召回率}$$

1. **数据集：**

在训练数据发布阶段，我们会发布1400条左右的标注好的问题数据集。在测试数据发布阶段，我们会再发布1400条作用的问题数据集，不含标注结果，作为测试。

本任务使用知识图谱来自知识工场（kw.fudan.edu.cn）的CN-DBpedia。选手需要通过API调用数据库。API相关描述在<http://kw.fudan.edu.cn/apis/cndbpedia/>中。有一个API需要特别注意：

api/mention2entity，输入实体名称(mention)，返回CN-DBpedia的对应实体(entity)的列表,json格式。**选手使用此API获取实体-名称的映射表。**

此外，CN-DBpedia的API描述页面还提供其他API，用以获得实体相关属性信息。包括api/entityAVP、api/entityType、api/entityTag等。**选手使用这些API获取实体的相关属性，用以辅助实体链接。**

**api/mention2entity**

输入实体名称(mention)，返回CN-DBpedia的对应实体(entity)的列表,json格式。

URL

http://knowledgeworks.cn:30001/?p=\*\*

Demo

<http://knowledgeworks.cn:30001/>

Example

<http://knowledgeworks.cn:30001/?p=南京>

如果一个名称对应多个实体，则在Demo中标记“\*primary”的实体为默认最佳匹配。

例：“南京”对应至少三个实体：江苏省省会、同名电影、香烟名称。其中江苏省省会会在Demo被标记为“\*primary”。

1. **任务提交**

每一个参赛队需提交的材料如下：

1. QEDL运行结果文件
2. 相关代码及说明

以上两个文件需在任务提交截止日期前发送至邮箱info.knowledgeworks@gmail.com。邮件的标题为：“CCKS-QEDL-参赛队名称”，例如“CCKS-QEDL-菜刀队”。

结果文件用result.txt命名。

代码及其文档需打包成一个文件（tar，zip，gzip，rar等均可），用code.xxx命名，要求提交所有的程序代码及相关的配置说明，且所得结果与result.txt相符。