数据集介绍

数据集的原始语料位于https://www.klcl.pku.edu.cn/gxzy/231686.htm，需要同学自行下载。语料的格式为每行一个句子，每行的开头为句子的编号。

数据集为原始语料的标注信息，分为三个文件夹，分别是train，validation，test。每个文件夹中都有若干json文件，编码方式为GBK 编码。每个文件记录了一个指代关系。

每个json文件的格式定义如下：

- "taskID"：任务ID
- "pronoun"：代词
  - "id"：原始语料中的句子编号，表示此代词位于这个句子里。
  - "indexFront"：表示实体中的第一个词是句子中第几个词，**请注意是词而非字符**，从0开始计数
  - "indexBehind"：表示实体中的最后一个词是句子中第几个词，**请注意是词而非字符**，从0开始计数
- "antecedentNum"：先行词个数
- "{number}"：这个代词指代的第{number}个先行词
  - "id"：语料中的句子编号，表示此先行词位于这个句子里。
  - "indexFront"：表示先行词中的第一个词是句子中第几个词，**请注意是词而非字符**，从0开始计数
  - "indexBehind"：表示先行词中的最后一个词是句子中第几个词，**请注意是词而非字符**，从0开始计数
- "taskID"：任务号，与文件名中的任务号相同
- "contributor"：标注此数据的同学

以下是一个实例：

```
{
    "taskID": "19980106-11-008 403",
    "0": {
        "id": "19980106-11-008-008",
        "indexFront": 4,
        "indexBehind": 5
    },
    "pronoun": {
        "id": "19980106-11-008-008",
        "indexFront": 7,
        "indexBehind": 7
    },
    "antecedentNum": 1
}
```

- "antecedentNum": 1 表示代词指代了一个实体。
- "0" 和 "pronoun" 的 id 都为 "19980106-11-008-008"。 在原始语料中，对应句子为：

> 19980106-11-008-008/m 两/m 年/qt 来/f ，/wd 杨/nrf 光/nrg 和/c 他/rr 的/ud 战友/n 们/k 用/p 一/m 片/qc 赤诚/an ，/wd 把/p 党/n 的/ud 阳光/n 洒/vt 满/a 人民/n 群众/n 的/ud 心田/n 。/wj

- "pronoun"的起始位置是7-7，对应着句子中的“他/rr” ，"0" 的起始位置是4-5，对应着原本句子中的 "杨/nrf" "光/nrg"。

额外说明

- 如果有级联的指代关系，只需要判断前一个指代关系即可。例如

> 19980118-03-006-004/m 塞纳河/ns 是/vt 一/m 条/qe 母亲河/n 。/wj 人们/n 说/vt ，/wd **巴黎/ns** 是/vt 从/p 塞纳河/ns 的/ud 浪花/n 上{shang5}/f 浮现/vt 的/ud 。/wj 两千/m 年/qt 前/f ，/wd **她/rr** 只/d 是/vl 河/n 中/f 那/rz 不足/vt 半/m 平方公里/qd 的/ud 小/a 岛/n ，/wd 形/Ng 同/vt 小舟/n ，/wd 又/d 似/Vg 摇篮/n 。/wj “/wyz **她/rr** 飘浮/vt 着/uz ，/wd 永不/d 沉没/vi ”/wyy ，/wd 巴黎/ns 的/ud 城徽/n 上{shang5}/f 嵌刻/v 着/uz 这/rz 句/qe 箴言/n 。

句子中 “两千/m 年/qt 前/f ，/wd **她/rr** ” 中的"她"与 " **她/rr** 飘浮/vt 着/uz" 中的“她”都指代**巴黎**。但模型只需要判断第一个“她”指代巴黎，第二个“她”指代第一个”她“即可。不需要考虑祖先的关系。在标注时也没有标注级联的指代关系。

志愿者：梁慧智、邹宇

指导教师：辛欣