Web 信息处理与应用：实验 2
信息抽取部分
实验于 2019 年 11 月 20 日开始，为期四周，两人一组进行分组实验。
请于 2019 年 12 月 18 日前将实验报告发送至课程邮箱：ustcweb2019@163.com

实验总体要求：
给定若干数量的医疗文档，其中，每条文档含有数量不一的医疗相关命名实体。
具体包括六类命名实体：疾病与诊断、检查、检验、药物、手术、解剖部位。
请为给定的每条医疗文档，找出其中所有属于上述六类的医疗命名实体，并准确判断其类别。

必备考核内容：命名实体识别（基于词典/规则或统计方法均可，方法自选）
可选考核内容：中文分词（是否对整个句子进行分词自定）
主要评价实体识别（NER）结果，同时考察两个子任务（即实体边界切分和实体类别判断）。
严禁抄袭代码，一经查实本实验作 0 分处理。

数据文件格式说明：
训练数据包含“subtask1_training_part1.json”一个文件。
其中，每一行为一条医疗文档，具体内容格式如下：
{"originalText":"患者行肝肿瘤切除术后出院。", "entities":[{"end_pos": 9, "label_type": "手术", "overlap": 0, "start_pos": 3}]}

在该条文档中，原始文本为“**患者行肝肿瘤切除术后出院**”部分。
所包含的实体信息形如：{"end_pos": 9, "label_type": "手术", "overlap": 0, "start_pos": 3}
其中包含如下信息：
实体类别：“手术”
起始位置：3
结束位置：9
重叠信息：0（在本实验中，该部分信息无意义，可忽略）

- 注意：其中条目的顺序可能会有变化，但任何一个实体均包含这四项元素。
- 位置计算从正文开始，第一个字为 0，以此类推，所有标点符号和数字均视作一个字。
- 结束位置代表实体结束后第一个字的位置。例如在本文中，起始位置为 3，结束位置为 9，意味着“肝肿瘤切除术”这一实体。

训练数据获取方式；
百度网盘：https://pan.baidu.com/s/1H7F0yTZ8_pNHTjuo9R8zGQ 提取码: pf2c

测试数据包含“实验二测试数据集.json”一个文件
其中，每一行为一条医疗文档，且仅包含训练数据中的 originalText 部分。
同时，在每行的末尾，包含“textID”信息作为该文档的编号。

测试数据获取方式；
百度网盘：https://pan.baidu.com/s/1-lf1MWfRM2r3yBssqC8hYA 提取码: 49aj

评价标准说明：

命名实体识别（NER）包含两个子任务，即实体边界切分和实体类别判断。

本实验采用严格匹配的计算方式，即两个任务全部完成才判定为正确。

换言之，只有在某个命名实体的**开始位置、结束位置与实体类别**均与标准答案完全一致，才认为该命名实体被准确识别。

基于上述判定，计算结果的准确率（Precision）与召回率（Recall）。最终分数以 F1 值指标为准，结合实验报告进行评判。

提交文件的数据格式如下：

提交文件为一 csv 格式文件，其中每一行代表一个命名实体， 共有 4 列，分别是 'textId', 'label_type', 'start_pos', 'end_pos'。

提交文件格式如下图所示：

textId	label_type	start_pos	end_pos
0	解剖部位	137	139
0	药物	291	300
0	手术	237	239
0	解剖部位	422	425
0	疾病和诊断	382	385
0	疾病和诊断	134	136
0	疾病和诊断	10	12
0	解剖部位	96	108
0	疾病和诊断	117	120
0	疾病和诊断	181	195
1	解剖部位	135	136
1	解剖部位	93	95
1	疾病和诊断	151	152
599	疾病和诊断	311	312
599	影像检查	318	321
599	解剖部位	247	248
599	解剖部位	114	115
599	解剖部位	137	141
599	解剖部位	186	224
599	解剖部位	290	295

textId 对应于测试数据集中每个文本的 textId，范围为 0~599。label_type 表示实体类型，start_pos,end_pos 分别是所识别出的实体，在原始文本的起始,终止索引。其中 textId，start_pos，end_pos 为 int64 类型，label_type 为一字符串。提交文件的行数即是你所识别出实体数量。

和实验 1 一样，提交时请务必确保文件格式正确，列名正确，每列的数据类型正确，不要忘记加上列名。

请将文件命名为 姓名_学号_lab2_submission_N.csv，通过 http://118.25.90.40:8081 进行评测。

在最终结果提交前，将安排若干次测试提交并反馈结果（只反馈指标，不反馈正确答案），第一次提交时间请等候课程群内通知。

本说明文档将根据实验进行不断更新。更新时将通过课程主页、课程 QQ 群及课上等渠道进行通知，敬请关注。