浙江工业大学

《文本分析与挖掘》

2023/2024(1)

期末综合作业

计算机科学与技术学院

期末综合作业

一、目的

随着智慧医疗的发展，需要对医疗信息数据进行挖掘和分析。医学领域中包含大量非结构化文本，由中文自然语言组成。利用机器读取医学文本，可以显著提高临床科研的效率和质量，并且可服务于下游子任务。要想让机器“读懂”医学数据，核心在于让计算机在大量医学文本中准确提取出关键信息，这就涉及到了命名实体识别、关系抽取等自然语言处理技术。

本实验针对医疗信息数据进行挖掘和分析，将命名实体识别技术与医学专业领域结合。实体抽取是从非结构化医学文本中找出医学实体，如疾病、症状的过程。本实验采用真实数据集，考察学生分析问题，利用文本预处理、文本表示、开源语言模型以及各种算法，解决医疗文本数据中实体识别的综合能力；加深对文本预处理，建模和算法实现整个过程所涉及方法的理解；提高在实际文本分析和处理问题中应用相关技巧的熟练性。

二、提交材料（只需电子版）

- 1 月 12 日 22：00 **前**，提交以下文件到**对应**钉钉群文件夹
- 所有文件均单独命名为：学号+姓名+题目：

  1.实验报告：.pdf；

  2.源代码：.python,命名规则同上。

  3.演示+解说视频（控制时间讲重点）。自评为 A 的同学除了递交视频还要现场演示。

  4.Excel 自评表:自评分数、主要工作。

  注意：**本报告与课内实验报告在格式和内容上都不同**。本报告应该与**毕业设计论文**更为相似，包括：问题的背景描述、对现有相关方法和技术的调研与讨论、明确挖掘目标和任务、提出解决思路和技术方案、模型和系统框架、结果展示和分析、讨论和总结等部分。每部分**内容详实、格式规范**，包含自己的分析、总结和看法。具体可以参考所给出的模板。

三、数据集和具体要求

本数据集 CMeEE-V2 包含训练集（_train 15000 条）、验证集（_dev 5000 条）以及测试集（_test 3000 条）三部分。其中对于实体的类别主要划分为九大类，包括：疾病(dis)，临床表现(sym)，药物(dru)，医疗设备(equ)，医疗程序(pro)，身体(bod)，医学检验项目(ite)，微生物类(mic)，科室(dep)。对数据集的详细描述可以查看数据集压缩包中“标注规范参考.pdf”

评测任务采用严格 Micro-F1 作为主评测指标，即输入语句后要求预测出的**实体的起始、结束下标，实体类型**精准匹配才算预测正确。递交结果示例可参考压缩包中“example_pred.txt”文件。结果提交到阿里云天池平台（https://tianchi.aliyun.com/dataset/95414/submission），将得到的分数截图包含在报告里面。

任务进阶要求：在基础之上实现系统完整化，例如搭建前端页面使系统有交互功能；实现创新性功能等。

基于以上给定数据，编程实现算法和系统，实验工具和平台不限。

四、评分标准

此次作业总成绩：实验报告+系统演示。

成绩评判综合考虑学生的实验设计思路、实现方法的新颖性、编程能力、独立思考能力、实验结果（天池平台给出的评估分数）和实验报告的撰写情况等多种因素。其中，**系统功能完整性、创新性，实验报告的内容详实性、写作规范性**为主要考核因素。

五、进度

自任务发布起：**独自**或**以小组为单位**对数据集进行初步探索、探讨需求、确定选题和具体分析任务，并进行相关方法、工具的调研，自行报名第一阶段汇报。

12 月 29 日：课堂汇报以上第一阶段进展，讨论并汇总小组调研结果。以小组为单位的自行指定汇报人（汇报的同学有加分）。

除了调研阶段可以多人讨论汇总，之后的**系统实现**和**报告撰写**要求每人独立完成（即报告的具体内容不能和其他同学共享）。

1 月 12 日：现场系统演示+验收

自评 A 的同学现场进行系统演示，现场演示不可补。