浙江工业大学

《文本分析与挖掘》

2021/2022(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

随着移动互联网的发展，在各个领域出现了以海量短文本为挖掘对象的应用。本实验以实际需求以及真实数据集出发，考核学生针对短文本数据挖掘与分析的挑战，综合利用文本预处理、文本表示、预训练词嵌入与模型，以及挖掘算法，包括传统算法和深度神经网络模型解决短文本数据分析与挖掘问题的综合能力。加深其对文本预处理，分析模型和算法及整个过程所涉及技术和技巧的熟悉掌握。

二、提交材料

最终提交以下材料到相应钉钉群文件夹：

- **压缩包**（命名为：学号+姓名+题目）：1.实验报告+.python 源代码。
- 系统演示/解说视频（控制时间讲重点，视频太大不方便存储）。自评为 A 的除了视频还要现场演示。
- **Excel 自评表**：主要工作、自评分数等级（填入给定模板）。
- **实验报告**（参照模板）：包括对问题的分析、明确挖掘和分析目标、解决思路和主要技术、模型和系统框架、结果展示和分析、讨论和总结等部分。实验设置清晰，包含必要的过程性讨论和分析。

三、数据集和具体要求

本实验针对警情数据进行挖掘和分析，实现警情类别预测，类别关键字提取、

事件/特征关联性分析等相关任务。由于类别太多，以及辅警业务能力有限导致人工标注不准确。类别预测的目的是对发生警情进行自动归类，以提高警务人员工作效率和准确率。类别关键字提取以及相关性分析则帮助对不同类别警情进行画像，以及建立知识图谱，从而有助于宣传以及防范同类别事件的发生。

本数据集包含两列非文本信息：发生地域编号和报警人性别，以及两列文本信息：报警内容和出警情况，一共分成 6 个警情类别。数据集已经按 5：1 分成训练集和测试集。

基于以上给定数据，自己选取和设定分析与挖掘任务，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、实验结果情况和实验报告的撰写情况等多种因素。其中，创新性、工作量和系统完整性、实验报告的详实和规范性为主要考核因素。

五、进度

自任务发布起：各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。

12 月 27 日：以小组为单位课堂汇报以上第一阶段进展情况

按照自主学习小组，讨论并汇总各自选题方向和调研结果。每组自行指定汇报人。

12 月 31 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。

所有同学按照前面第二点要求递交相关材料（只需电子）。