《文本数据挖掘》课程作业

课程编号：081104M05019H　　**课程属性：**专业普及课　　**学时/学分：**40/2

一、作业目的：

通过本课程作业加深对文本数据挖掘基础理论的认识和了解，锻炼和提高分析问题、解决问题的能力。通过对具体项目的任务分析、技术调研、数据准备、算法设计和编码实现以及系统调试等几个环节的练习，基本掌握实现一个文本数据挖掘系统的基本过程。

二、作业题目：

1. 实现一个文本聚类系统

文本聚类实现文本集合的自动划分。要求在公开的文本分类数据集上删除类别标签，采用不同的聚类算法对文本集合进行聚类，对照有标签的文本集合评价不同聚类算法的优劣。

2. 实现一个情感分类系统

分别利用基于情感词典的规则方法、传统的统计学习方法（例如朴素贝叶斯、支持向量机等）以及神经网络模型（如卷积神经网络、自注意力机制网络和预训练语言模型）分别实现基于文本的情感分类系统，在公共数据集上对三类方法进行对比分析。注：英文情感词典和中文情感词典都可以分别从网上公开获取。

3. 实现句子的向量表示方法

实现句子的语义向量表示方法，至少与三种神经网络模型（例如循环神经网络、卷积神经网络和自注意力机制网络）在公共数据集上进行实验对比，分析优缺点。英文训练集、开发集和测试集可从下面的网站下载：
http://ixa2.si.ehu.eus/stswiki/index.php/STSbenchmark
中文数据集可从下面的网址下载：
http://www.nlpr.ia.ac.cn/cip/dataset.htm

4. 实现一个汉语命名实体自动识别系统

命名实体一般指如下几类专用名词：人名、地名和组织机构名。要求使用不同方法实现三类实体的识别，并在公开数据集上分别对比总体的识别效果以及三类实体各自的识别效果。

5. 实现一个实体关系分类系统

实体关系分类是指预测两个实体之间的关系。要求至少使用两种方法（例如统计学习方法和神经网络方法，或两种神经网络方法）实现实体关系分类，并在公开数据集上分别对比总体的分类效果以及不同关系类型各自的分类效果。

6. 实现一个主题模型

主题模型就是在文档集中学习、识别和提取隐藏在文档背后的主题，从而挖掘文本背后的语义。请实现概率潜在语义分析和潜在狄利克雷分布两类主题模型，利用公开的文本分类训练数据分析并对比两类主题模型。

7. 实现一个话题检测系统

话题检测的目标是从连续的报道数据流中检测出新话题或此前没有定义的话题。要求从社交网络中爬取相关数据或者下载公开数据集，实现话题检测功能，并且分析话题检测效果。

8. 实现一个自动摘要系统

自动摘要实现从海量冗余数据到核心信息的压缩过程。要求使用抽取式方法和生成式方法实现自动摘要功能，并在公开数据集上利用自动评价指标对比不同自动摘要方法的性能。

9. 实现一个基于预训练大模型微调的文本挖掘系统

当前，ChatGPT 等预训练大模型具有很强的零样本（Zero-Shot）学习能力，即可以依据提示完成各种任务。现有的 Colossal-AI 等一些训练框架可以在 RTX 2060 6GB 普通笔记本能训练 15 亿参数模型。要求利用十亿级别的预训练模型在多个文本数据挖掘任务上进行多任务指令微调，对比不同任务在公开数据集上的性能。

三、基本要求：

(1) 每人可以选择其中的一个题目，也可以几个人合作完成一个题目，原则上合作人数不应超过 3 人，彼此之间必须有明确的分工和要求。除了上述题目，也可以自由选择其他与文本数据挖掘有关的题目，但需要事先说明。

(2) 任何一个题目，不限定所采用的方法，鼓励方法创新，但是需要有方法介绍、对比和分析，必须有理论根据或实验数据依据。

(3) 完成一份技术报告，报告内容包括：项目目标、国内外相关工作、自己在本项目中承担工作的不同点、实现系统（或模块）的核心思想和算法描述、系

统主要模块流程、实验结果对比及分析，并列出主要的参考文献以及每位参与者的具体贡献。

(4) 提交系统源代码和可执行程序，以保证实验系统可以正常编译和运行。如果是多人合作完成的，应提交最终集成的系统。

(5) ？月?日（北京时间 24:00）之前提交技术报告、系统代码和可执行程序。请留下作者的姓名、单位、联系电话和邮件地址。可以通过电子邮件提交或直接提交光盘。

四、特别声明：

(1) 鼓励充分使用网络资源和其它一切可以利用的资源（包括数据、语料、软件工具和论文资料等），但严禁侵害他人知识产权，技术报告中必须明确说明所用资源的真实来源。

(2) 鼓励相互交流、相互合作，但严禁抄袭他人工作，严禁伪造结果。