2024.08.04报告（07.29-08.04）

本周主要完成以下两个任务.

1. VAST数据集分析
    - 背景介绍：VAST数据集的来源、目的和应用领域
    - 研究目标

      VAST数据集的主要目标是支持零样本（zero-shot）和少量样本（few-shot）立场检测任务.传统的立场检测方法依赖于大量数据，然而现实中有成千上万的主题，每个主题都收集足够的标注数据既耗时又昂贵.因此，VAST数据集及其提出的模型旨在克服这一困难，通过广义主题表示（Generalized Topic Representations）来隐式地捕捉主题之间的关系，从而提升模型在未见主题上的泛化能力.
    - 数据集目的
    -
        1. **支持零样本和少量样本立场检测的研究**：通过提供包含大量不同主题和多种表达方式的数据集，研究者可以评估模型在未见主题上的性能.
        2. **改善立场检测的泛化能力**：通过提出新的模型（TGA Net），该数据集展示了如何利用广义主题表示来提升模型在复杂语言现象（如讽刺）上的表现，并减少对情感线索的依赖.
    - 应用领域
        1. **社交媒体分析**：在社交媒体平台上自动检测用户对特定话题的立场，帮助理解公众情绪和观点动态.
        2. **新闻内容分析**：分析新闻报道或评论中的立场倾向，辅助新闻报道的客观性和公正性评估.
        3. **政策制定和公众意见调查**：了解公众对政策提案、社会问题的立场，为政策制定提供数据支持.
    - 数据集结构
        - 样本数

          VAST数据集包含23,525个样本，每个样本由一个评论（document）、一个主题（topic）和一个立场标签（stance label）组成.
        - 特征数

          数据集主要包含三个特征：

            1. **评论（d_i）**：文本数据，表示评论的具体内容.
            2. **主题（t_i）**：文本数据，表示与评论相关的特定主题短语.
            3. **立场标签（y_i）**：分类标签，表示评论对主题的立场（支持、反对或中立）.
        - 数据类型

          数据集中主要包含文本数据（评论和主题）和分类标签（立场标签）.文本数据用于模型输入，分类标签用于模型训练和评估.
        - 数据字典

特征名称	数据类型	描述
d_i (文档)	文本	评论的具体内容，作为立场检测的主要输入.
t_i (主题)	文本	与评论相关的特定主题短语，是立场检测任务的目标主题.
y_i (立场标签)	分类标签	评论对主题t_i的立场，分为支持（pro）、反对（con）或中立（neutral）.

通过这些特征和标签，VAST数据集为立场检测任务提供了一个全面且具有挑战性的数据集，特别是在零样本和少量样本场景下.

2. 师兄论文项目 SentKB 复现
   - 环境搭建
     1. 使用Anaconda创建一个虚拟环境，名称`SentKB`，`python=3.9.6`：

        ```
        conda create -n SentKB python=3.9.6
        ```

     2. 安装第三方包

        项目包含requirements.txt文件，给出了代码所需第三方包的版本.

        ```
        python==3.9.6
        pytorch==1.12.1
        transformers==4.32.1
        torch-scatter==2.0.9
        pytorch-geometric==2.3.1
        spacy==3.5.3
        requests==2.27.1
        responses==0.13.3
        tqdm==4.65.0
        gensim==4.3.0
        networkx==3.1
        ```

        我并未使用`pip install -r requirements.txt`直接安装所有包，因为我要更换安装源，使安装过程更快速.于是我单独安装每一个包，添加`-i https://pypi.tuna.tsinghua.edu.cn/simple`参数来更换安装源.

     3. 调整代码

        根据我实际硬件情况，调整师兄的代码以正常运行.

   - 运行代码

     根据`README.md`，直接运行`run_vast.py`.

     ```
     python -u "E:\YCJH\Project\2_2024.07.29-08.04\SentKB-main\run_vast.py"
     ```

     `-u`参数用于强制Python的标准输出和标准错误流（stdout和stderr）为无缓冲模式，可以使实时显示进度条或日志信息等直接显示在终端上，而非储存在缓存中.

   - 问题解决

     在运行代码的过程中，我遇到了一个大问题：我无法访问 huggingface.co . 这是一个储存模型文件的网站，`transformers`库的`AutoTokenizer`会从这个网站下载模型文件.

为了解决这个问题，我搜索了网络，找到了huggingface.co的一个国内镜像站 hf-mirror.com. 然而，我无法替换transformers库的默认下载源. 这困扰了我很久，我尝试手动下载所有的模型文件，但是工作量太大了，我需要一个一个的下载.

这个问题师兄也无能为力，他也很忙. 只有下周继续研究了.

以上即为本周的学习情况.