城市动态系统多源数据融合分析与应用

本课程作业以“城市动态系统多源数据融合分析与应用”为核心题目，突出**开放性与探索性**特征，允许学生从交通流量预测、空气质量关联、社会事件影响等多个子方向自主选择研究主题。题目设计需紧密结合多源数据特性，鼓励学生通过数据融合方法解决实际城市问题，同时提供灵活的研究路径以满足个性化探索需求。

1. 核心子方向与典型题目案例

1.1 交通动态分析与预测方向

该方向聚焦城市交通系统的多维度数据融合，涵盖流量预测、事故影响分析、通勤行为挖掘等任务。典型题目设计可参考以下案例：

- **大规模交通流量预测**：基于 LargeST 数据集 LargeST – CityMind Lab（包含多城市路网、气象、节假日等多源数据）构建时空预测模型，探索不同时间粒度（5 分钟/1 小时）下的流量变化规律。
- **交通事故影响机制分析**：利用 USAccidents 数据集 US Accidents (2016 - 2023)（含事故位置、天气、道路条件等属性）识别事故高发时段与区域，构建事故严重程度预测模型，并分析气象因素对事故传播范围的影响。
- **多源数据融合的交通管理**：如“首尔市交通流量预测与管理策略效果评估” 서울열린데이터광장，要求学生融合实时交通流数据、历史气象记录及公共交通调度信息，评估限行政策对早晚高峰的缓解效果。

1.2 社会事件与社交媒体影响方向

结合社交媒体数据与城市动态系统，分析社会事件的传播路径及其对城市功能的影响。典型题目包括：

- **社会事件的交通影响传播**：基于 Twitter 事件数据集 Twitter-Dataset（如特定节假日或大型活动期间的推文），构建事件热度与交通流量的关联模型，识别关键传播节点与时间滞后效应。
- **公共情绪与城市服务需求关联**：如“基于酒店预订数据的旅游情绪分析” Hotel booking demand，对比城市酒店与度假酒店的预订需求差异，结合社交媒体情感评分（如 Twitter 用户对目的地的评价），预测季节性入住率波动。
- **谣言检测与城市安全响应**：利用 FakeNewsNet KaiDMML/FakeNewsNet: This is a dataset for fake news detection research 数据集，构建社交媒体谣言检测模型，并分析谣言传播对城市交通、医疗等公共服务的潜在干扰路径。

1.3 多源数据融合的拓展应用方向

结合公开数据集探索城市系统的其他维度，如环境、公共健康等。参考题目包括：

- **气象与城市系统多维度关联**：基于 NOAA 气候数据 Climate Data Online (CDO) - The National Climatic Data Center's (NCDC) Climate Data Online (CDO) provides free access to NCDC's archive of historical weather and climate data in addition to station history information. | National Climatic Data Center (NCDC)与城市交通、空气质量数据集，分析青岛地区 40 年月平均气温变化对居民通勤方式（如共享单车使用量）及呼吸道疾病发病率的影响。

2. 研究问题设计示例与引导方法

为引导学生基于多源数据集定义个性化研究问题，需提供明确的问题构造框架。以下为典型研究问题示例及设计思路：

社交媒体事件的交通影响传播路径分析——以 2023 年上海马拉松赛事为例，利用 Twitter 赛事相关推文（含地理位置、发布时间）与上海地铁客流量数据，通过图神经网络（GNN）识别事件热度向交通节点扩散的关键路径。

个性化问题定义引导步骤：

1. **数据驱动选题**：提供多源数据集清单（如交通类 LargeST、社交媒体类 Twitter 事件数据集、气象类 NOAA 数据等），引导学生通过数据字典分析数据维度，发现潜在关联（如交通流量数据中的“时间戳”与社交媒体数据中的“事件发布时间”可构建时间关联）。
2. **问题结构化训练**：要求学生按“**研究对象（城市子系统）+ 数据类型（多源数据组合）+ 核心方法（融合/预测/关联）+ 应用价值（实践意义）**”四要素定义问题，例如：“基于芝加哥交通碰撞数据集（含天气、时间属性）与空气质量指数（AQI）数据，分析 PM2.5 浓度与事故严重程度的关联性，为恶劣天气下的交通管制提供依据”11。
3. **参考平台案例迁移**：鼓励学生借鉴 Kaggle、阿里天池等平台的成熟项目框架（如“Hotelbookingdemand 酒店预订需求分析”的多维度对比思路），结合课程数据集进行创新性调整，避免简单重复已有研究。

3. 数据集要求

为满足开放式大数据原理与应用课程的实践需求，本课程作业数据集采用**基础数据层-关联数据层-融合应用层**三级架构设计，覆盖结构化时序数据、非结构化文本数据及多源融合数据类型，确保学生可通过标准化路径获取并开展从数据预处理到融合分析的全流程实践。

基础数据层聚焦交通流量与气象两类核心结构化时序数据，提供时空维度完整、指标体系规范的基础数据源，支撑学生掌握时序数据处理与时空分析方法。

关联数据层提供标注完善的社交媒体事件文本数据，支持自然语言处理（NLP）基础任务（如情感分析、事件提取），并可与基础数据层关联分析外部事件对交通系统的影响机制。

融合应用层通过预关联或可关联的多模态数据集，展示基础数据与关联数据的耦合方式，引导学生掌握多源数据融合分析方法，探索跨模态信息对交通系统的影响规律。

典型融合模式与数据集案例

气象-交通时空关联

数据集组合：NOAA 气象数据+PeMS 交通数据

- **来源**：https://www.ncdc.noaa.gov/cdo-web/ + https://pems.dot.ca.gov/
- **数据量**：气象数据 350 万条（400+站点×10 年×hourly）+ 交通数据 2 亿条（5 分钟间隔）
- **耦合方式**：基于经纬度与时间戳时空对齐，例如将加州某高速公路传感器（34.05°N, 118.24°W）的 5 分钟交通流量数据与 10 公里内气象站的小时降水数据关联，分析雨天对车速的影响

4. 技术实现要求

技术实现的工具链以 Python 生态为核心，覆盖数据获取、预处理、建模、分布式计算全流程，具体包括以下模块：

4.1 开发环境与核心库

开发环境采用 Python 3.9 及 Jupyter Notebook，支持交互式代码编写与可视化分析。核心库包括：

- **数据处理**：Pandas（数据清洗、类型转换如 pd.to_datetime 处理日期数据）、NumPy（数值计算）、NLTK（文本预处理，如停用词移除、词干提取）；
- **建模工具**：Scikit-learn（传统机器学习，如 Logistic 回归、SVM）、PyTorch（深度学习模型构建）、Hugging Face datasets（高效数据管理）；
- **分布式计算**：PySpark（亿级数据分区处理、多源数据 Join 操作）；
- **数据获取**：Requests（HTTP 下载，如交通数据集批量获取）、Tweepy（Twitter API 访问，如抓取特定用户推文）、Wget（中等规模文件下载）。

领域专用工具

- **文本处理**：TextBlob（情感极性计算）、WordCloud（词云生成）；
- **时空数据**：SQL Server（PeMS 交通数据集存储）、Neo4j（图数据管理，支持 Cypher 查询）；
- **深度学习**：Sentence-transformers（预训练文本嵌入模型，如 paraphrase-multilingual-minilm-l12-v2）。

4.2 算法设计

算法设计需基于数据类型匹配原则，结合任务目标选择组合方案，具体包括以下两组典型示例：

组合示例一：时序预测+图计算

- **适用场景**：交通流量预测、舆情传播分析等时空关联数据。
- **算法选择**：
  - **时序预测**：采用 LSTM 或时空融合模型（如 DCRNN、ASTGCN）处理时间序列特征，例如基于 PeMS04/08 数据集预测路段车流量；

    - **图计算**：结合图神经网络（GNN）分析路网拓扑结构或社交网络关系，如通过 Twitter 关注者网络提取信息传播子图。
- **选择依据**：交通数据具有时间连续性（需 LSTM 捕捉时序依赖）和空间关联性（需 GNN 建模路网连接），二者结合提升预测精度。

组合示例二：分类算法+聚类分析

- **适用场景**：文本情感分类、用户群体划分等非结构化数据任务。
- **算法选择**：
    - **分类算法**：SVM 或朴素贝叶斯用于情感极性判断（如基于 TextBlob 的 polarity 得分将推文分为正面/负面/中性）；
    - **聚类分析**：结合 K-means 对分类结果进一步分组，如将负面情感推文按主题聚类（如“交通拥堵”“事故通报”）。
- **选择依据**：文本数据需先通过分类算法标注情感标签，再通过聚类挖掘潜在主题，二者形成“分类-聚类”递进分析链

4.3 可视化方案

可视化方案区分静态与交互式两类，结合 Python 库实现多维度数据呈现：

静态可视化

通过 Matplotlib、Seaborn 等库生成固定图表，适用于报告输出或静态展示：

- **趋势图**：使用 matplotlib.pyplot 绘制交通流量日变化曲线（x 轴为时间，y 轴为车流量），直观展示早高峰（7:00-9:00）与晚高峰（17:00-19:00）特征；
- **热力图**：通过 Seaborn 的 heatmap 展示区域事故密度，以颜色深浅映射事故频次，辅助识别高危路段 2。

交互式可视化

基于 Plotly、Folium 构建动态仪表盘，支持用户实时探索数据：

- **地理信息仪表盘**：使用 Folium 在地图上标记交通事件位置，点击弹窗显示详细信息（如事故时间、伤亡人数）；
- **时间滑块动画**：通过 Plotly 的 animation_frame 参数制作年度流量变化动画，滑块控制时间维度，直观展示流量随季节/政策的变化趋势。

5. 提交内容

报告采用 IEEE 期刊格式撰写，包含摘要、引言、方法、实验、讨论及结论等核心章节。讨论部分需客观分析研究局限，例如“数据集未覆盖特殊天气事件（如暴雨、暴雪），可能影响极端条件下模型泛化能力”；未来改进方向可包括引入多模态数据融合（如实时气象数据）及注意力机制优化等。

可视化成果区分交付形式：静态图表（如模型性能对比柱状图、交通流量时空分布图）提交高清 PNG 格式（分辨率 300dpi）；交互式仪表盘采用 Dash 框架开发，支持数据筛选（如时间范围、事故类型）与动态预测结果展示，交付物为独立 HTML 文件；同时需附 3 分钟内操作说明视频，演示仪表盘核心功能（如数据查询、模型预测结果导出）。

交付物清单

1. **代码工程**：含模块化源代码及单元测试脚本
2. **数据集**：原始数据及预处理后的数据文件（CSV/Parquet 格式），附数据集描述文档

3. **实验报告：**IEEE 格式 PDF 文档，含研究局限与未来改进方向(代码全部上传至 github，附在报告中)
4. **可视化成果：**静态 PNG 图表、Dash 交互式仪表盘（HTML）、操作说明视频（MP4 格式）
5. **环境配置文件：**requirements.txt 及环境搭建说明文档

所有交付物需打包为压缩文件，命名格式为“课程名称-学号-项目名称-v1.0.zip”，确保文件结构清晰且无冗余内容。所有交付物发至助教邮箱 lzx77883322@gmail.com。