2021 年第八届中国可视化与可视分析大会

数据可视分析挑战赛

（ChinaVis Data Challenge 2021）

作品说明文档

参赛队名称：永不加班队

作品名称：AirHub: 全国大气污染分析预测系统

作品主题关键词：大气污染源分析，污染时空态势分析与预测，气象特征分析

团队成员：薛春宇，上海交通大学，Dicardo@sjtu.edu.cn

苏勇文，上海交通大学，yongwensu@sjtu.edu.cn，队长

刁义嘉，上海交通大学，diao_yijia@sjtu.edu.cn，

周轻，上海交通大学，Luff_zz@sjtu.edu.cn

董笑菊，上海交通大学，指导老师

团队成员是否与报名表一致（是或否）：是

是否学生队（是或否）：是

使用的分析工具或开发工具（如果使用了自己研发的软件或工具请具体说明）：D3，EChart，Python (Flask)，Tensorflow (LSTM)，MySQL，Vue，百度地图, Jieba, paddlepaddle

共计耗费时间（人天）：100 人天

本次比赛结束后，我们是否可以在网络上公布该文档与相关视频（是或否）：是

一、作品简介

空气质量，是衡量空气污染程度的重要指标，可以根据空气中污染物的浓度高低进行判断，清洁的空气对于改善我国人民的身体健康，创造宜居的环境至关重要。得益于我国日益完善的空气质量检测网络，空气质量的分析人员可以收集到大量准确的空气质量数据。而如何有效利用这些数据，理解大气污染传播方式，分析大气污染物来源与扩散方式，对于防治大气污染具有重要意义。

此次，我们团队借助 2013-2018 年中国高分辨率大气污染再分析开放数据集，构建了一个全国大气污染物分析与预测系统。该系统通过提取污染物的全国分布与局部分布，结合局部气象指标，对污染物的时空分布趋势进行了可视化与一定时间范围内的预测。该系统的目标用户为空气质量的分析监控人员与污染防控机构，可以用于指挥大厅或宣传科普大屏幕进行展示。

系统在宏观上展现了全国各地主要城市的空气污染情况，并在各个子系统中以可交互的方式展示了地方空气质量相关新闻，污染程度排名，污染物时空变化趋势以及局部气象指标。为用户了解全国与各地空气质量提供了多维度的分析渠道，可以帮助当地政府深入了解污染物来源与历史动态，从而制定出更加合理的空气质量管理政策。

二、数据介绍

作品中所使用的大气污染及相关气象数据来源包括 ChinaVis 2021 挑战赛提供的 2013-2018 年中国高分辨率大气污染再分析开放数据集，阿里云 DataV 提供的地图轮廓以及该时间段内各城市气象局发布于新闻网站的相关舆情数据。原始大气污染数据为 csv 格式，包含经纬度、各污染物含量（PM2.5、PM10、SO2、NO2、CO、O3）和气象指标等多个字段。

本系统利用百度地图 API 提供的数据转化功能，将原始数据中的经纬度坐标转化为城市名。这一功能基于百度地图 API 实现。在原始数据集和预先设定的被观察城市集合的基础上，进行了原始数据的清洗和筛选。最后的 json 文件包含每个城市每日的污染物浓度，以及气象指标。

本系统采用阿里云 DataV 所提供的地图轮廓 GeoJson 数据来绘制各省份中主要城市的环境空气质量指数（AQI）3D 柱状图，并且利用城市名与其地理坐标、污染物指数的映射关系完成了 3D 图表的数值绘制。

本系统收集 2013-2018 年各城市气象局发布于官方网站的环保相关的新闻舆情数据，在进行基于关键词匹配的数据清洗之后，利用 NLP 模型进行分类及词云的生成，如图 2.2 所示。

中国高分辨率大气污染再分析开放数据集
http://naq.cicidata.top:10443/chinavis/opendata

中华人民共和国生态环境部 http://www.mee.gov.cn/

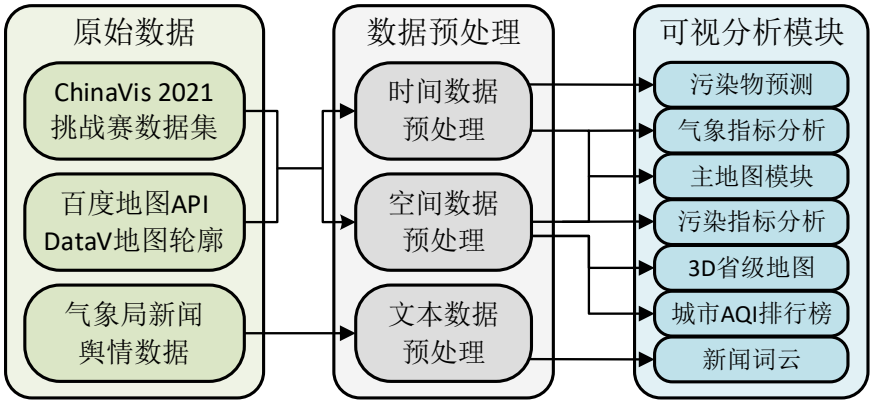
阿里云 DataV 地图轮廓数据链接 https://geo.datav.aliyun.com/areas/bound/id_full.json

三、分析任务与可视分析总体流程

基于本系统已有的数据来源与制作本系统的可视化与可视分析预期效果，

本系统的可视分析任务可以概括为：综合污染物各个指标数据、地理位置信息与文本舆情数据，综合进行时间数据可视化、空间数据可视化与文本数据可视化，在此基础上进行对污染物的时空分布、变化情况、预计变化与舆情关系进行综合分析。

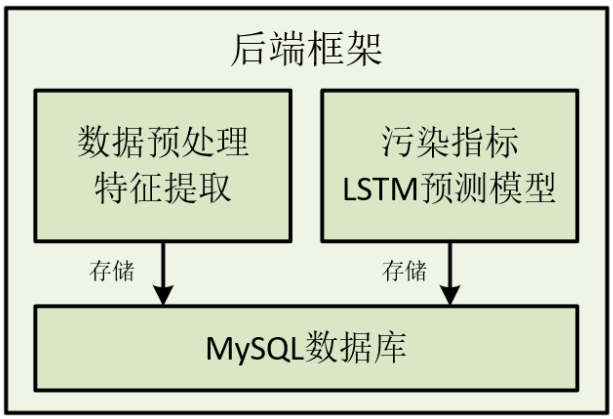
基于该可视分析任务，本系统的可视分析总体流程如下图所示。



首先，获取原始数据，并进行归类。其次，从时间、空间和文本三个角度，分别对相应的原始数据进行预处理。最后，将经过预处理的数据进行相应模块的可视化工作，在此基础上完成可视分析的流程。

需要特别说明的是，本系统采用的原始数据基本都包含时间、空间信息；本系统的可视化效果也基本都包含时间、空间信息。这不与可视分析的总体流程矛盾。对于不同数据和处理方式，可视分析需要突出的重点不同；本总体流程突出每个数据和模块的侧重点。

在可视分析总体流程的基础上，本系统的实现主要分为前端和后端两个部分。后端部分主要执行原始数据获取后的预处理部分，如下图所示；前端部分主要绘制可视化效果。



四、数据处理与算法模型

各城市污染物分布数据提取：

数据集所提供的数据为全国范围内各经纬度地区的每日污染物与气象指标，通过百度 API 提供接口，我们将相关特征与城市名建立映射关系，通过统计各城市的中心位置、城市污染物及气象指标特征向量以及污染程度指数 AQI，构建基于城市的数据集。自建的城市数据集格式主要为 csv 格式，且同时建立 MySQL 数据库来进行存储，以便前后端交互时信息的快速筛选及查找，如图 2.1 所示。

各城市风向频率数据提取：

除去污染物分布数据以外，另一个需要特殊处理的数据类型为风向频率，我们需要统计各城市的全部经向风速与纬向风速，在此基础上，统计八个方向来风频率，并按照风速对其进行分级，最终可以得到各城市全年的风向频率统计图，每个方向的来风按照其风力等级统计其频率。

各城市污染物 AQI 数据提取：

根据环保局所发布的国家环境空气质量标准，每种污染物的浓度与具体 IAQI 数值一一对应，可以提取每个城市某一具体日期时各污染物所属空气质量指数即 IAQI，并取其中最高 AQI 指数为该城市当日 AQI 指数，具有最高 AQI 的污染物即为该城市当日主要污染物。

实时新闻文本提取：

获取到的新闻为非结构化信息，我们需要对其进行数据清洗和处理。首先观察新闻数据结构找到有价值的信息标签，随后使用正则表达式匹配提取有价值的新闻内容，放到给定的文件目录下供新闻展示模块使用。

NLP 模型分词：

将获取到的新闻文本放入指定 txt 文件，然后调用开源 python 库 Jieba 对新闻文本进行分词

Jieba 分词算法：

- 基于前缀词典实现高效的词图扫描，生成句子中汉字所有可能成词情况所构成的有向无环图 (DAG)
- 采用了动态规划查找最大概率路径，找出基于词频的最大切分组合
- 对于未登录词，采用了基于汉字成词能力的 HMM 模型，使用了 Viterbi 算法

为了提高分词精度，我们启动了 paddle 模式，利用 PaddlePaddle 深度学习框架，训练序列标注（双向 GRU）网络模型实现分词。

对分词处理后的数据进行观察发现部分高频词汇是无意义的虚词和代词，因此我们加入词汇类型过滤逻辑，过滤出高频的动词、名词、形容词等具有特征意

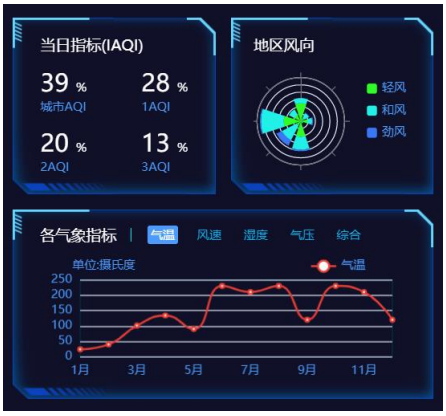
义的词汇，从而使可视化的词云更加具有意义。

五、可视化与交互设计

总体大图：

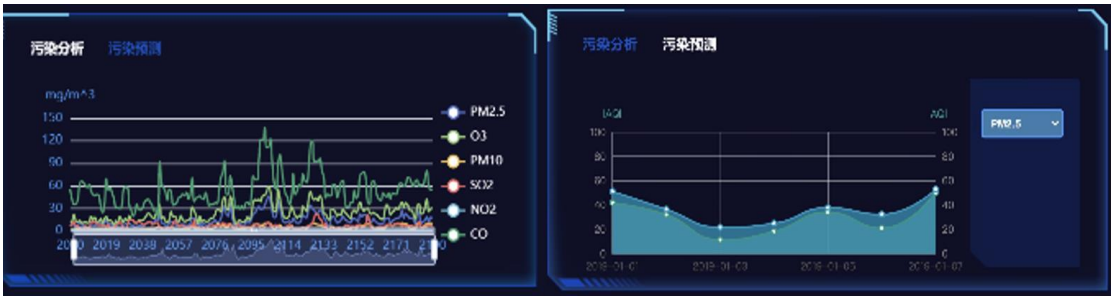
总体大图通过日期选择框和污染物选择框与后端进行交互获取到对应的污染物数据，并调用 echarts 的 map 组件展示。污染物大图设计为用三种颜色表示污染的严重程度，同时用点的直径表示污染物数值大小，因为需要防治出现过大的点造成遮蔽，因此需要对数据进行归一化预处理，将原数据映射到一定范围之内。最终较好的展示了中国地图上大气污染物的分布，并且随着缩放等级的变化可以改变数据点的疏密程度，实现更灵活的控制和舒适的观看体验。

气象指标：



如上图所示，气象指标的可视化由三个子组件构成，分别描述了当前城市 AQI 指数，全年风向频率以及温度，气压，湿度等气象指数随时间的变化情况。其中，当日指标指示了当前城市 AQI，主要污染物与次要污染物和其 IAQI，地区风向图不再赘述，而气象指标提供气温，风速，湿度，气压数值随时间变化趋势的可视化功能。在此基础上，我们还提供了将各气象指标集中在一张图中的功能，便于用户分析单个指标的变化规律与各指标之间的潜在联系。

污染物历史变化与预测数据可视化：

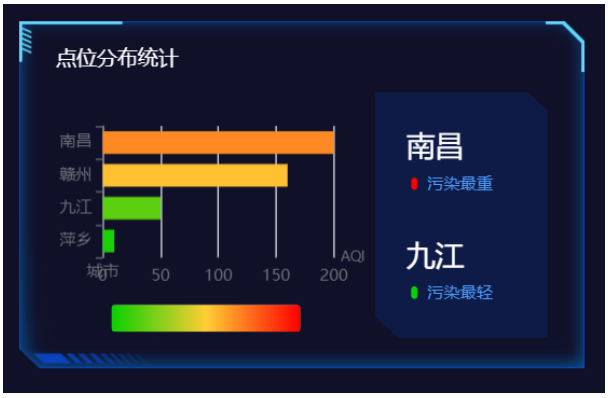


在这一模块，我们提供两个按钮为用户在两个可视化模块之间提供切换功能。其中，污染分析模块展示了各污染物浓度随时间变化趋势，用户可以通过勾选图例选择想要了解的污染物进行显示并分析，而非将 6 种污染物全部显示。

污染预测模块则提供了在当前城市，通过 LSTM 模型所预测的未来 7 天各种污染

物的 IAQI 指数以及该地区对应的每日 AQI 指数的可视化效果。用户可以通过右边的下拉勾选框选择某一污染物的预测结果进行显示，当 IAQI 曲线与 AQI 曲线有公共点时，说明此时当前城市的主要污染物即为选中的污染物种类。

省级行政区内各城市空气质量排名可视化：



如上图所示，为了展示省市内各城市相对空气质量，我们绘制了按各城市 AQI 指数进行排序的纵向柱状图，用户可以在右侧的面板看到一省内空气质量最差与最好的城市，同时利用 echarts 对各城市 AQI 进行视觉映射，形象地展示了 AQI 所代表的空气质量。

省份 3D 地图与主要城市环境空气质量指数可视化：

如上图所示，本部分基于省份的 3D 地图，在主要城市的中心点，对该城市当天的 AQI 进行柱状图可视化。3D 地图具有较强的交互性，通过旋转、缩放以及视角拖拽，可以清晰地反映出省份内城市 AQI 与地域的关系。当鼠标在特定城市附近悬浮，会显示出该城市的具体污染物指标，使交互信息更加丰富。

污染物大图与新闻模块交互：

通过点击污染物大图上的城市或者省份，设置点击触发函数并且传输点击到的地点名称到后端，从而获取当前城市的新闻，通过更改 html 的方式显示到前端，从而实现实时播报特定城市新闻。

污染物大图与词云模块交互：

与新闻模块交互类似，不过不同之处在于需要服务器对新闻文本进行清洗以及分词和统计词频，打包成特定的格式后发送给前端，前端调用 echarts-wordcloud 进行词云的绘制。用户可以通过词云发现当前新闻舆情的关注点与特色内容。

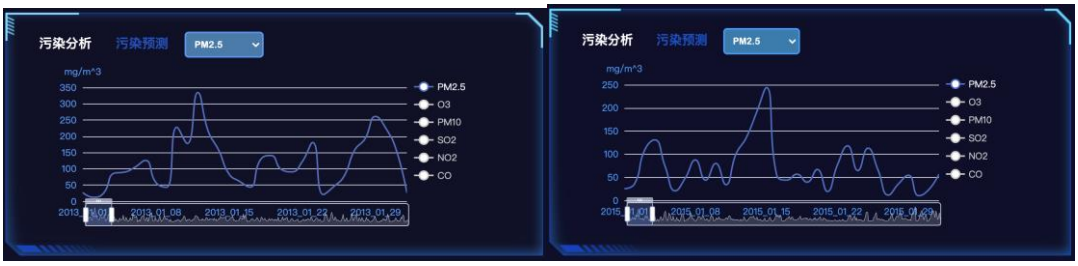
六、案例分析

案例 1：北京市 2013 年以来雾霾(PM2.5)治理情况

雾霾常见于城市，是特定气候条件与人类活动相互作用的结果。在 2013 年，

“雾霾”成为年度关键词。这一年，在北京，仅有 5 天不是雾霾天，通过本小组的污染物分布图可以看到，在 2013 年，北京市的主要污染物为 PM2.5，在全年的大部分时间，北京市的空气质量均较差。严重的雾霾天气引起了政府的重视，在 2014 年 1 月 4 日，国家减灾办，民政部首次将危害健康的雾霾天气纳入 2013 年自然灾情进行通报。

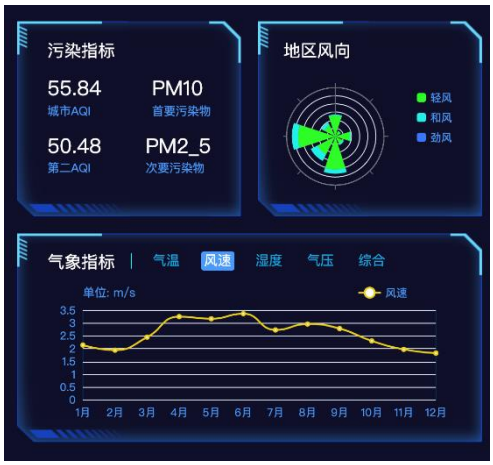
利用本小组的可视化系统，我们可以对雾霾较为严重的北方城市如北京，天津和陕西，山西等省市的空气污染情况进行追踪调查。以北京为例，如下图所示，在 13 年 1 月，北京的空气质量较差，PM2.5 峰值接近 300mg/m^3,而经过政府的有意识管控与相关政策的出台，在 15 年 1 月，PM2.5 浓度有显著减低，而在 2018 年，其浓度大大减少，这与政府推行的政策有很大关联。



案例 2：以甘肃省兰州市为例，分析沙尘暴形成规律(PM10 颗粒物质)及其与风向指标的联系

本案例选取甘肃省兰州市进行分析。兰州市在每年 3 到 5 月多发沙尘天气；可以通过 PM10 污染指标来反映沙尘天气。如下图所示，在 2015 年 3-4 月，兰州市的 PM10 污染指标在一些天中超过 20mg 每立方米，说明在这些天中发生了沙尘天气。

对于这两段时间的污染成因，可以通过气象指标进行辅助分析。由下方图片可以看出，风速与风向是导致沙尘天气多少以及严重程度的重要成因。在 2015 年 4 月，风速相较于 3 月有一定程度上升；该年 4 月沙尘天气天数以及 PM10 污染指标峰值均高于三月。作为对比，2016 年 3、4 月风速基本相近，该年 3、4 月沙尘天气、PM10 污染指标变化比较相似。



通过以上分析，得出结论：兰州市 3-4 月的沙尘天气与风速相关，风速越大，出现沙尘天气的概率越大。

案例 3：主要污染物变化

随着机动车辆迅猛增加，中国部分城市的大气污染特征正在由烟煤型向汽车尾气型转变，NOx、CO 呈加重趋势，从 SO2 为主的污染物转向以 CO，NOx 为主，在本案例中，本小组选取上海进行分析：

如上两图所示，在 2013 年 7 月，上海的主要污染源为 PM2.5 以及 PM10，而在 2018 年同一时期，上海的主要污染源则变为 NO2 以及 PM10，可以看到，首要污染物从 PM2.5 转向 NO2，可以说明，随着机动车车辆数量的迅速增加，上海的主要大气污染源正在从 PM2.5 向汽车尾气所产生的 NO2 气体转变。

七、讨论与总结

伸缩性：

计算的伸缩性：系统中耗时较长的计算，如预测模型训练、时序数据特征提取、污染数据预测、新闻爬取与分析都是线下计算完成保存的。其余前端组件和后端都是实时交互的。

可视化的伸缩性:我们目前采用的是 echarts 的 map 组件，一次绘制全部城市内容过于拥挤且加载速度较慢，因此我们采用根据缩放级别更改显示城市的数量，从而控制屏幕上的点数量大致均衡。

可复用:

我们的系统侧重于对污染数值的分析预测、舆情的可视化分析。因为我们使用的预测模型和可视化框架适配于新数据，因此理论上，新近的大气污染数据可以便捷、准确地融入已有的可视图表中。

1. 新闻文本处理过程中可能因为结构变化导致正则提取式失效，此时需要进行对应的修改方能正确提取出有价值的新闻事件。

2. 新加入的数据可能无法对预测模型起到显著的训练效果，这是由于新加入的数据量可能较少或者模型较大训练较为缓慢。

总结：

我们的系统较好地实现了第一章提出的任务，可以通过实时的可视化互动反映污染态势变化、污染数据预测、新闻舆情分析等重要信息，基本达到了设计目标，并且在美观度上进行提升，最终实现了实用性和美观度的统一。