



山东大学计算机科学与技术学院

School of Computer Science and Technology, Shandong University

大数据分析与实践课程 经验分享

分享人：赵艺淇
2024年9月12日



C 目录 Content

学无止境
气有浩然



山东大学
SHANDONG UNIVERSITY

01

项目准备

02

项目实践

03

项目选择





项目准备 — 实用工具

1. **Github** : <https://github.com>
2. **Google Scholar** : <https://scholar.google.com>
3. **Papers with Code** : <https://paperswithcode.com>
4. **Navicat** : <https://navicat.com.cn>
5. **Another Redis Desktop Manager** : <https://github.com/qishibo/AnotherRedisDesktopManager>
6. **Hoopscotch** : <https://hoppscotch.io>
7. **Poe** : <https://poe.com>
8. **Ant Design** : <https://ant-design.antgroup.com/index-cn?locale=zh-CN>
9. **Element Plus** : <https://element-plus.org/zh-CN>
10. **阿里巴巴矢量图标库** : <https://www.iconfont.cn>



项目实践 — 项目初期

团队分工，各展所长

作为组长，需要参考各组员水平，擅长内容，意愿等，尽量做到**工作量均匀，分工明确**；作为组员，需要明确自己的工作内容，**不消极不懈怠，按时按量**完成内容

阅读论文，复现代码

阅读项目**相关论文**，看看别人怎么做的，然后可以在**github**上或者**paperwithcode**上找找项目相关代码，复现一下，体会一遍项目流程

制定目标，稳步推进

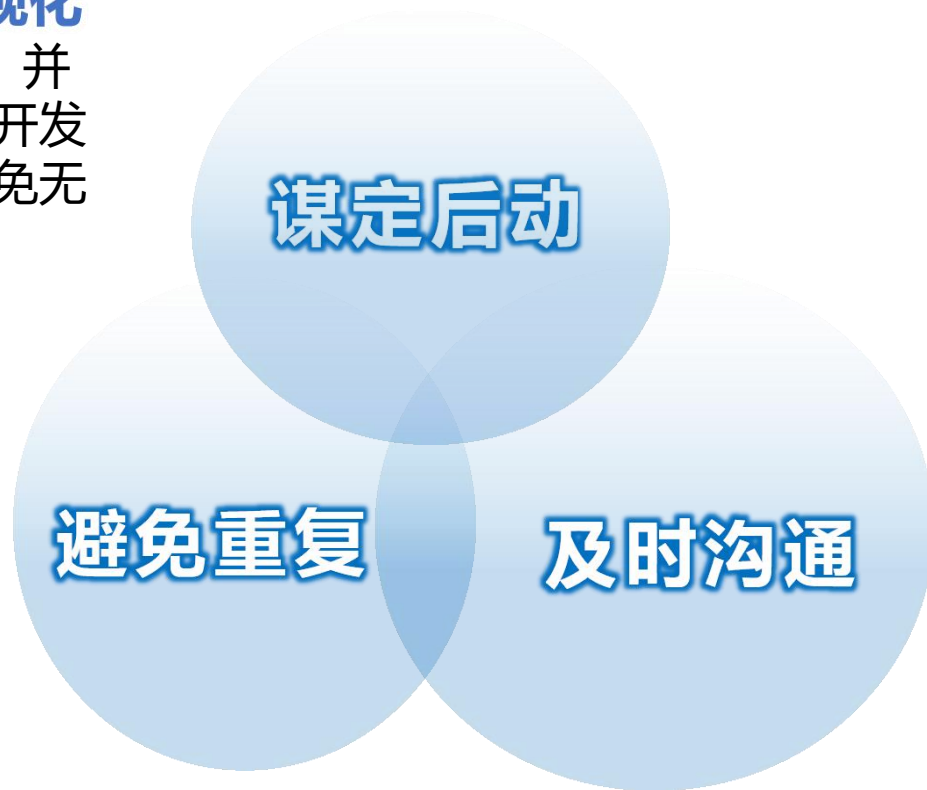
小组可以每周制定一个**里程碑**，每周开**1~2次**会进行同步，组员们按照分工初步了解自己负责模块所需技术，尝试**写写demo**加深体会



项目实践 — 项目中期

一定要先尽量**设计完善**，包括**数据库设计**，**可视化展示**，**框架选择**等等，并且将自己的业务逻辑，开发思路都想好后再写，避免无效开发

要在每个开发阶段**明确自己的工作内容**，不要和队友进行相同的工作，避免**重复开发**；同样的，如果自己完成的快，也可以酌情帮助同组其他成员进行开发



团队内对于业务产生疑惑要**及时沟通，互相商讨**；团队间也可以互相交流，面对某些难题或者共同任务时可以**一起研究**。同学们遇到不会的问题可以先自己尝试解决，积累经验，**不要害怕提问**，有问题解决不了及时求助**老师与助教**



项目实践 — 项目后期

列出目标

列一个**Todo-list**，把所有**待做的，未完成的工作**写到里面，然后是谁来完成就在后面加上谁的**名字**，完成一项就打一个勾，同样的，每新发现一项工作就写到文件里，这样便于**总结**和**统计**

预留时间

要留出各模块的**对接时间**和**整体测试时间**，一般来说，对接可能会耗费很多的时间，并且找到更多的问题；同样，整体测试也是必要的，可以让当前模块的开发人员不被自己的**固有思维局限**住



项目选择



Project Topics

NEW DESIGN

	Topics
1	时空气象数据驱动的臭氧浓度可视分析系统（合作单位：青岛市环境监测中心）
2	学生就业决策系统（合作单位：计算机学院）
3	世界范围大气污染时空经济效益可视分析（合作单位：环境学院）
4	大模型引导的冷冻电镜数据可视分析系统（合作单位：数学交叉中心）
5	社会稳定风险预警系统（合作单位：政治与管理学院）
6	房价预测与决策系统（合作单位：政治与管理学院）
7	智慧政管全景平（合作单位：政治与管理学院）台
8	IEEE VIS or CCF大数据挑战赛赛题
9	自拟题目，要求涵盖数据管理、数据分析、数据呈现、数据交互等流程



项目选择 — 时空气象数据驱动的臭氧浓度可视分析系统

充分的代码基础

01

两位学长**四个月**间断性开发的代码基础，满足各种要求的**输入输出以及可视化呈现**，模型现有效果还算不错

简单的上手难度

03

若不更换模型则在LightGBM上继续**构建特征工程**以期达到更好结果；换模型的话可以尝试深度学习，并且**大部分代码无需改动**

02

丰富的试错经验

学长们曾经尝试过诸多模型 (**LightGBM**, **Istm**, **Informer**, **Arima**等等, 以及**Kaggle**上部分表现优越模型等等)并保留下来大部分源码可供参考。

04

项目的实际价值

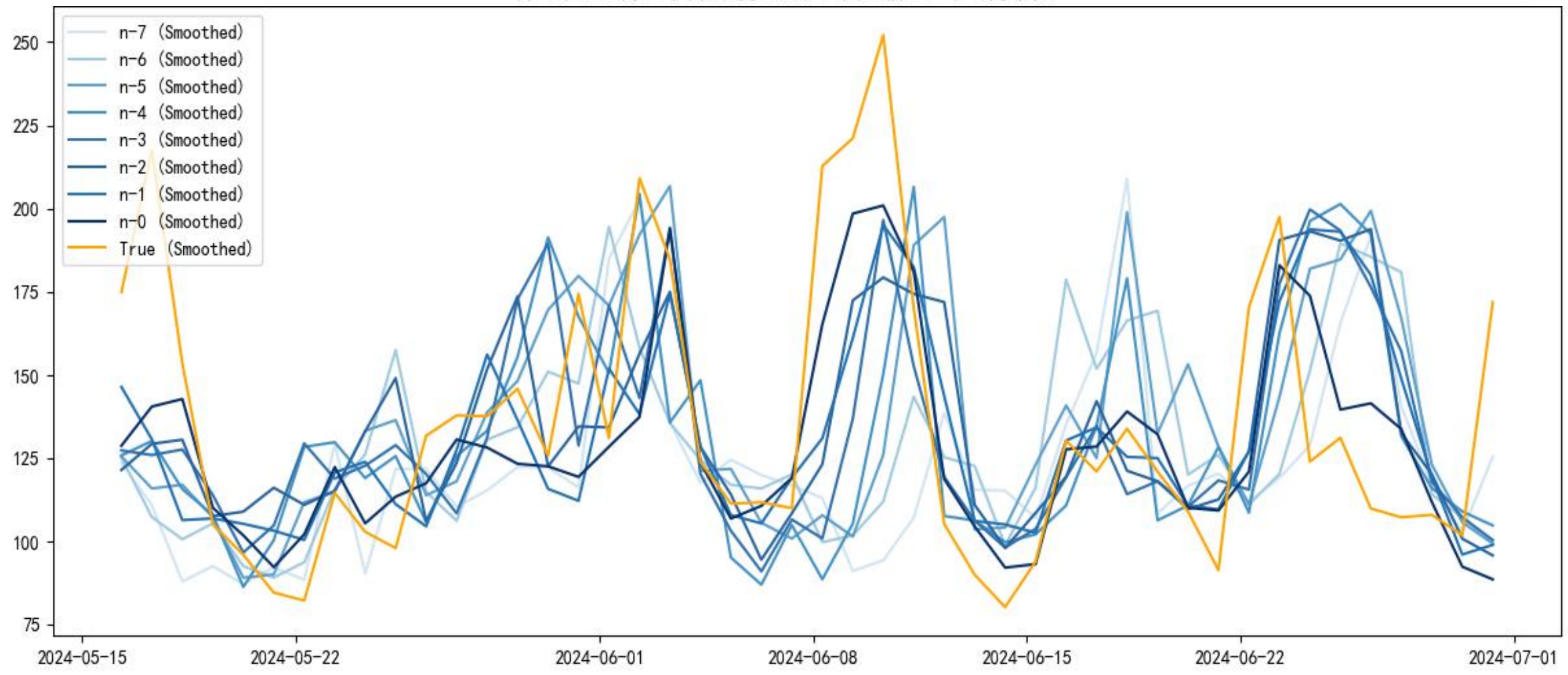
这个项目是与**青岛市环境监测中心**对接的，具有很强的**实践意义**，能提高大家对**机器学习**，**深度学习**的了解和认识，并在简历上留下浓墨重彩的一笔



项目选择 — 时空气象数据驱动的臭氧浓度可视分析系统

当前结果

青岛市北宅街道 臭氧浓度 预测-真实比较(每1小时为单位)





项目选择 — 时空气象数据驱动的臭氧浓度可视分析系统

注意事项

千万仔细检查代码，不要有任何形式的**数据泄露**，包括但不限于：**滞后特征**中原有的真实数据未被替换；自行构建的**统计特征**中部分真实数据也被统计但未被替换。

解决方法

1. **直接预测法**：构建1-N个模型
2. **递归预测法**：将前一轮的预测值作为当前待预测值的特征
3. **直接-递归混合预测法**：前两种的混合

参考文章：<https://zhuanlan.zhihu.com/p/390093091>



项目选择 — 世界范围大气污染时空经济效益可视分析

学科交叉，拓宽视野

可以让我们了解**更多领域**的知识，丰富专业技能，**拓宽**我们的**视野**，并将课堂所学实际应用到其他领域之中，发挥更大的作用

拓宽技术栈

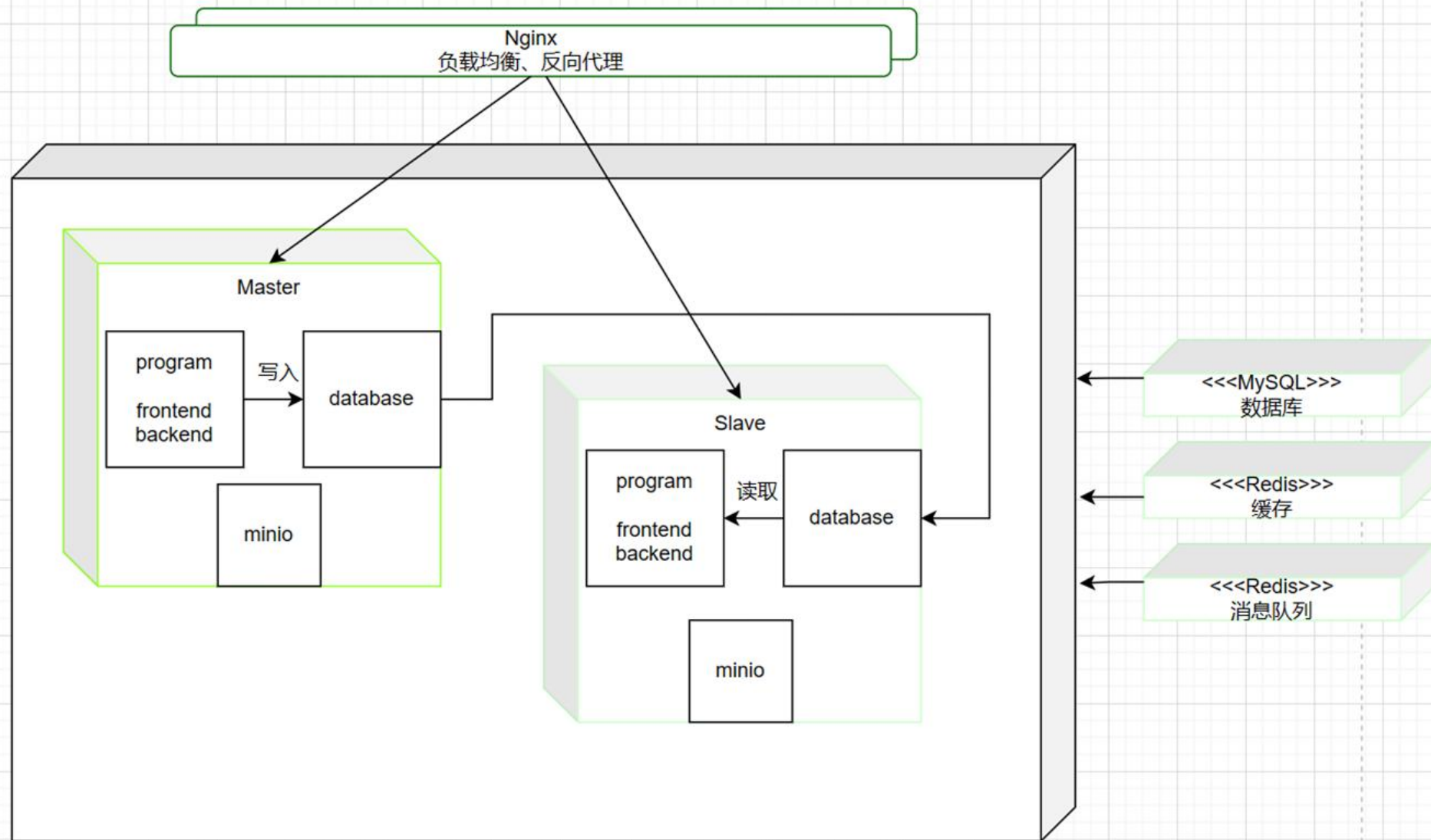
项目涉及**平台搭建**，**前后端开发**，**数据分析**，**可视化展示**，**框架构建**，**机器学习预测**，**负载均衡**，**Docker虚拟化**，**区块链**等等

强拓展性，未来可期

同学们如果做出一份不错的项目，可以拿着投**互联网+**等比赛，后面再包装包装，学科交叉的项目很容易获得省奖；也或者与其他**企业或者组织**进行对接，帮助他们进行某些工作



项目架构





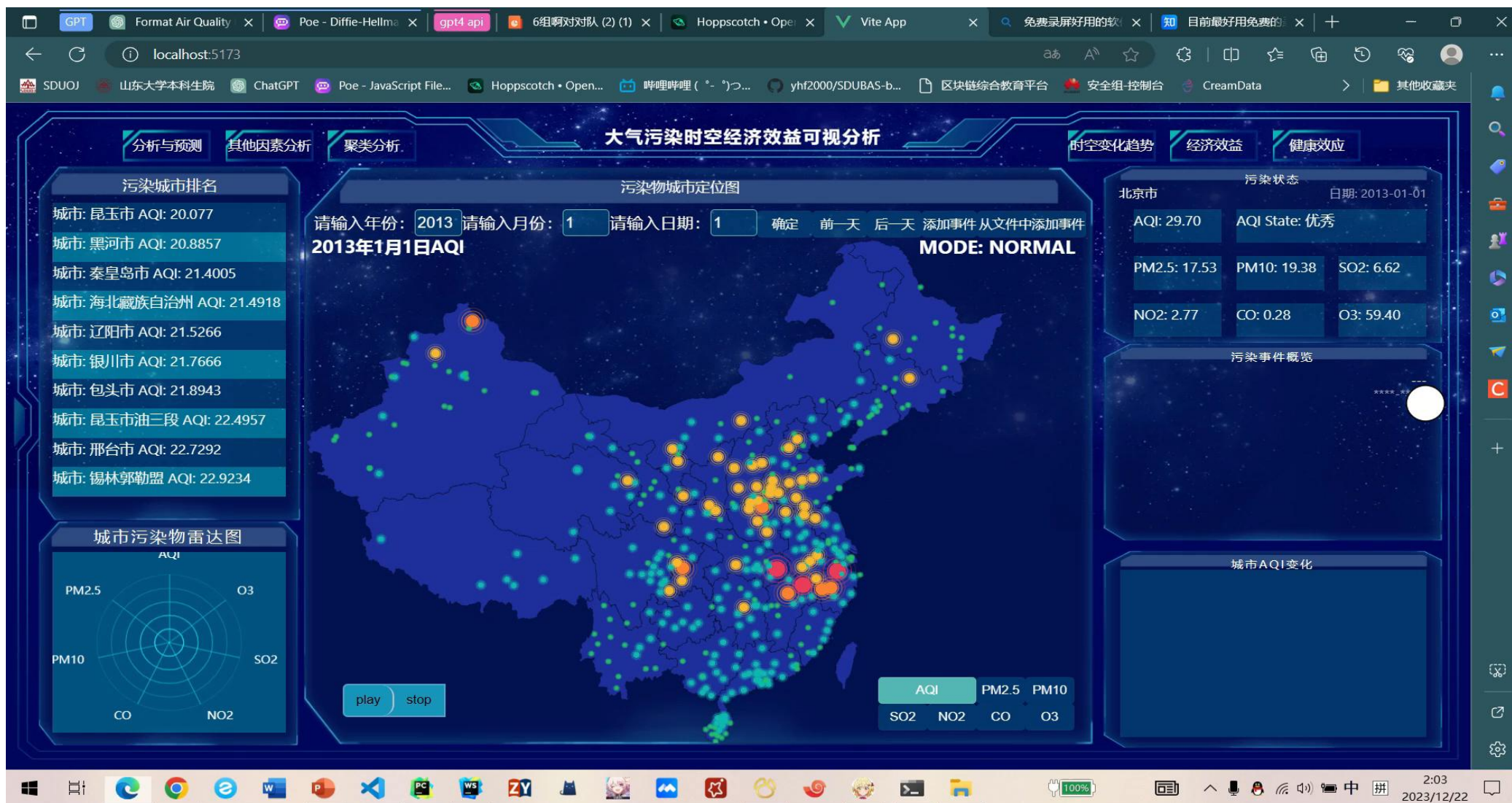
项目选择 — 世界范围大气污染时空经济效益可视分析

平台整体架构





项目选择 — 世界范围大气污染时空经济效益可视分析



前端仓库地址: https://github.com/mozhong123/Visual_analysis_of_air_pollution-frontend

后端仓库地址: https://github.com/mozhong123/Visual_analysis_of_air_pollution-backend



项目选择 — 世界范围大气污染时空经济效益可视分析

1. 可以进一步拓展为一个**智能决策系统**，通过人工智能**辅助分析与决策**的方式，按照用户的要求展示数据分析后与可视化的结果，并给出针对污染治理的**建议与决策**，以供决策者参考。
2. 我们不只局限于一个系统，更多的是希望基于我们的系统形成一个**大气污染可视分析的社区**，聚集更多人的智慧为治理大气污染这一全人类的共同问题献出一份力量。用户可以调用我们的接口**爬取**最新的天气数据贡献到系统中，并在管理员审核通过后上传最近的**污染事件**参与贡献。
3. 实现用户操作的**溯源**，即实现用户可以查看自己的所有贡献，具体一点便是对自己上传过的文件可以进行下载，也可以查看自己爬取的数据与添加的事件等。为了保护用户隐私，我们也可以对用户上传的文件作**混合式加密**，更可以引入**区块链**对用户的操作进行上链，真正实现防篡改。
4. 实时爬虫只能爬取精确到天的污染物数据，没有风向等数据。未来将通过**爬虫获取更多维**的数据，不只局限于大气污染方面，还有联网舆情、医学等大数据，与时俱进地更新与展现我们系统的各个模块。
5. 随着数据量的增大，未来考虑多引入几台从服务器，现在一主一从体现不出**主从框架**与我们设计的负载均衡的优势，或者考虑将架构升级成**微服务**。



山东大学计算机科学与技术学院

School of Computer Science and Technology, Shandong University

气有浩然  学无止境

2024年9月12日

