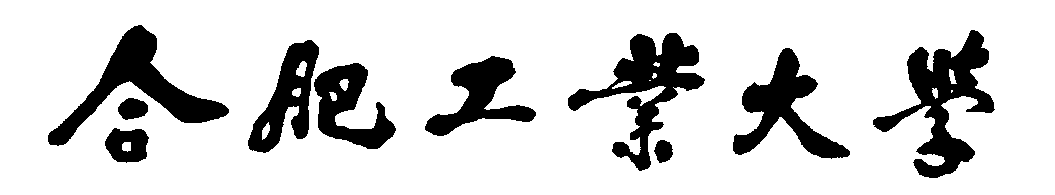
****

分布式程序设计

综合设计报告

|  |  |
| --- | --- |
| 专 业 班 级 | 智能科学与技术 |
| 学生姓名及学号 | 2022214191 曾振杰 |
| 课程教学班号 | 0500020X--001 |
| 任 课 教 师 | 张赞 |
| 实验指导教师 | 张赞 |
| 实验地点 | 翡翠湖校区科教楼C座机房 |
| 2024～2025学年第 一 学期 | |

**综合设计名称：**  基于 Spark 和大语言模型的台风数据分析

时间∶ 2024 年 12 月 24 日

|  |
| --- |
| **预习内容** |
| **一、目的和要求∶**  1.获取台风气象数据  2.使用spark进行数据处理与分析  3.使用大模型进行报告生成  4.可视化化展示 |
| **二、任务∶**  1. 台风模式分析：基于 Spark 的分布式计算能力，分析台风的历史路径模式。计算关键统计信息（如台风强度变化规律、台风路径平均偏移距离等）。对不同年份或区域的台风活动进行趋势分析。  2. 台风路径聚类：使用 Spark MLlib 实现对台风路径的空间聚类，识别常见路径模式。  3. 强度变化预测：训练简单的机器学习模型（如线性回归、时间序列模型），预测台风强度的未来变化。  4. 区域风险评估：基于历史数据计算某一地区的台风风险概率，并结合大语言模型生成区域风险报告。  5. 结果可视化展示。 |
| **三、准备方案，包括以下内容：**  （所采用的系统、组件、工具、核心方法、框架或流程图、程序清单等）  1.采用数据集Asian Typhoon Dataset, 1951-2022 (RSMC Best Track)  2.在服务器上搭建简易环境并完成数据运算  3.完成四个任务：模式分析、路径聚类、强度预测、风险评估  4.使用质谱AI平台调用大语言模式api  5.使用streamlit和github完成线上部署和可视化展示  使用环境：  本地虚拟机（hadoop+spark）用于测试；  本地主机用于项目管理和非spark部分测试；  服务器（使用预编译的spark+hadoop的local模式）；、  开发环境：  vscode+conda+jupyter+streamlit+git  python3.10.6  spark3.5.4    展示：  请访问：  [Streamlit](https://sparktest-v8fad5jyo4bsp3at8mputd.streamlit.app/): （请先确保可以访问streamlit平台#）  [https://sparktest-v8fad5jyo4bsp3at8mputd.streamlit.app/] |

|  |
| --- |
| **内容** |
| **一、仪器、设备：**  宿主机：计算机（AMD Ryzen 7 6800H+RTX3050），操作系统： windows11  虚拟机：vmware，操作系统：Ubuntu 22.04.4 LTS  服务器：学校算力平台服务器（感谢室友提供的账号） |
| **二、内容与步骤（过程及数据记录）：**  1.环境设置  需要在服务器上搭建临时的spark环境，大致方法同实验一，这里我直接使用的with-hadoop的预编译版本，通过wget下载到服务器中。  环境变量的设置直接创建一个临时的.bashrc，每次手动执行。  项目结构如下图：    spark设置为:    其中 spark.rapids.sql是允许sql操作替换为使用GPU加速的方法，当然还有其他GPU加速设置，但由于在服务器上不方便安装RAPIDS Spark因此这里并不涉及。      2.数据选择与整理  （1）数据选择  此次设计我选择了RSMC 最佳轨迹数据集，该数据集包含与台风相关的天气信息，原始数据是由日本气象局（JMA）以txt格式提供的。  数据集来源于[Kaggle](https://www.kaggle.com/datasets/skyil7/asian-typhoon-dataset-1951-2022-rsmc-best-track/data):    包含了1951 – 2022年的68624条台风数据。    （2）数据项  每条数据都有17列数据，说明如下：   |  | | --- | | 1. International number ID  类型：整数  描述：台风的国际编号。前两位数字表示台风发生的年份，后两位是从 1 开始的递增整数索引。如果前导数字为零，则省略。例如，2004 年的第二个台风编号为 402；1960 年的第十个台风编号为 6010。  2. year  类型：整数  范围：1951 到 2022  描述：台风发生的年份。  3. month  类型：整数  范围：1 到 12  描述：台风发生的月份。  4. day  类型：整数  描述：台风发生的日期。  5. hour  类型：整数  范围：0 到 23  描述：台风发生的小时。  6. grade  类型：分类数据  描述：台风的等级。可能的选项包括：  'Tropical Depression'  'Tropical Cyclone of TS intensity or higher'  'Extra-tropical Cyclone'  'Just entering into the responsible area of RSMC Tokyo-Typhoon Center'  'Severe Tropical Storm'  'Tropical Storm'  'Typhoon'  7. Latitude of the center  类型：整数  描述：台风中心的纬度，缩放因子为 10。例如，如果实际纬度为 25.3，则保存为 253。  8. Longitude of the center  类型：整数  描述：台风中心的经度，缩放因子为 10。例如，如果实际经度为 135.7，则保存为 1357。  9. Central pressure  类型：整数  描述：台风中心的气压，单位为百帕（hPa）。中央气压是衡量台风强度的重要指标之一，通常气压越低，台风强度越强。  10. Maximum sustained wind speed  类型：整数  描述：台风中心附近的最大持续风速，单位为节（kt）。最大持续风速是衡量台风强度的另一个重要指标。  11. Direction of the longest radius of 50kt winds or greater  类型：字符串  描述：50节（kt）或更大风速的最长半径的方向。可能的方向包括北（N）、东北（NE）、东（E）、东南（SE）、南（S）、西南（SW）、西（W）和西北（NW）。  12. The longest radius of 50kt winds or greater  类型：整数  描述：50节（kt）或更大风速的最长半径，单位为海里（nm）。这是台风风场的一个重要指标，表示台风影响范围的大小。  13. The shortest radius of 50kt winds or greater  类型：整数  描述：50节（kt）或更大风速的最短半径，单位为海里（nm）。这是台风风场的另一个重要指标。  14. Direction of the longest radius of 30kt winds or greater  类型：字符串  描述：30节（kt）或更大风速的最长半径的方向。可能的方向包括北（N）、东北（NE）、东（E）、东南（SE）、南（S）、西南（SW）、西（W）和西北（NW）。  15. The longest radius of 30kt winds or greater  类型：整数  描述：30节（kt）或更大风速的最长半径，单位为海里（nm）。这是台风风场的一个重要指标，表示台风影响范围的大小。  16. The shortest radius of 30kt winds or greater  类型：整数  描述：30节（kt）或更大风速的最短半径，单位为海里（nm）。这是台风风场的另一个重要指标。  17. Indicator of landfall or passage  类型：字符串  描述：指示台风是否登陆或经过陆地。可能的值包括空格（' '）和井号（'#'）。井号（'#'）表示台风中心已经到达陆地。首次出现井号是在1991年。 |   我们主要需要用到的信息包括台风编号、日期信息、台风经纬度、台风等级、中心气压、最大风数、是否登陆。    （3）数据处理  数据处理主要包括两方面，一个是选择，一个是转换。  比如对于台风的模式分析，用到的数据为编号、日期、经纬度、中心气压、等级，对于路径聚类，用到的只有编号、日期、经纬度。具体来说各任务用到的部分都不同，因此会在各任务中进行更多的选择和处理。  转换主要是针对要使用的数据，比如日期数据，需要从年、月、日、时当中获取date格式和时间戳格式的信息，因此可以：    针对经纬度，还需要将其缩放为原始尺度,如:    针对是否登陆，替换为0和1表示：    对于轨迹的可视化和聚类，还需要提取更多的数据，比如两条记录之间的距离、时间差和速度、方向等，具体在对应任务中讨论。  3.模式分析  （1）获取台风路径  首先获取需要的信息，可通过lag获取先后的数据信息：    地图的实现是通过folium库完成：      这部分是对于单个台风的位置信息的检查，下面是对于总体的分布的检查.  （2）获取台风分布  同样使用folium，可以绘制台风的热力图。    这是根据年份选择对应范围内的台风，将其轨迹以热力坐标的形式添加到地图上：    可选参数包括年份范围和热力图的点半径和模糊量。  （3）台风趋势  对于台风的总体情况，可以查看各类别台风趋势图和强度图：      4.路径聚类  （1）提取特征  聚类需要使用一定的特征，最直接的是每个台风的轨迹信息，即每个经纬度位置点，但每个台风的观测数很可能不一致，这样不容易作为输入特征给模型训练，因此这里我提取了以下特征：  总移动距离：    平均速度：    经纬度方差：    经纬度协方差    移动方向：    如下：    （2）聚类分析  使用pyspark.ml完成聚类，模型为KMeans。  步骤包括选择特征、合成特征向量、标准化、确定聚类数、聚类。  如下：      这里使用了剪影法（Silhouette Method）或者叫轮廓法，它是一种用于评估聚类结果质量的方法。它通过计算每个数据点的剪影系数来衡量聚类的紧密度和分离度。    a(i) 代表样本点的内聚度，b（i）为分离度，具体公式略。    然后训练、预测：    k=2时效果最好，从可视化图上看，两种类簇的模式差距还是很明显的，k为其他值时的结果也有，但区别不会很明显。        4.强度预测  （1）强度分析  先查看历史上的台风强度，可以通过饼图查看各级别的台风分布：    其中最多的是温带气旋和热带低压。    此外通过每年的平均中心气压也能反映出台风强度的变化趋势：    这里的风速是1977年才有的，那时JMA改变了观测方法，并加入了分速这项数据，因此我们的预测数据选择从1977年开始。  （2）时间上的的强度预测  如果以时间作为输入变量，中心气压作为输出，那么查看的则是年代与台风强度的关系，我尝试直接使用线性回归模型进行预测：    结果如下图：    发现结果并不好，分析是因为输入特征只有年份，这样的信息对于预测强度来说是不够的，正常情况下应该考虑时间序列预测模型，比如ARIMA或者Prophet，考虑到我们是spark的设计，并且鉴于时间较紧，这里就不改了。  对于下面的单台风分析我们还是用线性回归模型，但以当前的日期、经度、维度、中心气压作为输入特征，以下一个时间点的经度、维度、中心气压作为输出。  （3）单台风的路径和强度预测  这是整个项目最耗时的地方，此数据集有1881个台风，每个台风单独预测需要消耗大概4s，因此总体需要接近3h的时间，为了节省计算时间，我只取从1990年后的有登陆过的台风，其数量为123。  当然有另一种策略是在可视化化时现场训练和预测，但考虑到可视化平台难以支持spark，只能改用sklearn，不太符合这次的目的。  对于特征的抽取，需要处理数据获得日期和时间戳，还要对每一条数据获取下一个时间点的经纬度和气压数据，可以通过lead实现：      最后一个时间点的数据没有下一个时间点的信息，我们把它作为预测的输入，然后训练：    预测的过程是通过输入最后一个时间点信息，获取预测值，预测值通过一定的格式修改，比如时间和列名，可以视为下一时间点样本，再次作为输入可继续预测之后的时间点信息。      结果如图：      5.大语言模型风险评估  （1）api调用  （2）数据依据  （3）报告生成 |

|  |
| --- |
|  |
| **三、结果分析、思考题解答∶**  1.可视化展示  2.git和github上传  3. streamlit部署 |
| **四、感想、体会、建议∶** |
| **综合设计成绩∶**  **指导教师签名：**  年 月 日 |