作业6 空间数据应用

请从LBS空间数据库设计、空间索引实现、空间数据分析三个题目中选择一个，可以组队完成，每组最多2人，作业总分50分。

**Option A：LBS空间数据库设计**

**作业目的：**了解基于位置的服务(Location-based Services)和空间数据在日常生活中的应用，熟悉空间数据库设计流程，即需求分析、概念设计、逻辑设计、物理设计、数据库实施等阶段。

**作业内容：**

选择一个基于位置的服务应用，即基于移动终端位置的服务，移动终端例如智能手环、手表、手机和车辆等。基于位置的服务应用可以是已有的应用，例如高德地图、大众点评、滴滴打车、摩拜单车、健康与旅游类APP等，通过使用该应用APP，调研其用户需求和服务器端的数据库设计；也可以是自己设计的应用，分析用户需求和设计相应的数据库。针对选择的应用，完成需求分析、概念设计、逻辑设计、物理设计、数据库实施等。

**作业步骤：**

1. 需求分析。调研该应用的用户需求，分析未来的扩展用户需求。具体描述每一类用户需求，存储什么数据，怎么存储，数据与需求的关系，谁可以访问这些数据。此部分可参考软件工程的需求分析。

2. 概念设计。基于用户需求，分析该应用的实体、实体属性和实体之间的联系，绘制空间扩展E/R图，描述应用场景的约束条件。

3. 逻辑设计。基于空间扩展的E/R图，生成关系和函数依赖，进行相应的关系分解，降低数据冗余和异常，但不能影响数据插入和查询效率，影响用户体验。

4. 物理设计。分析性能要求，描述高峰期用户数据库访问量，选择哪些属性创建索引，针对哪些需求创建视图，针对哪些约束创建触发器。分析每类用户的权限，包括选择、插入、更新和删除等操作，保障数据的安全性。

5. 数据库实施。选择与空间数据相关的应用需求，至少构造1个关系创建语句、1个视图创建语句、1个索引创建语句、1个用户授权语句、1个数据插入语句、3个应用需求相关查询语句等。

**时间安排：**

1. 2022.11.22前完成组队和应用选择，钉钉群登记组队信息。

2. 2022.12.20前上交应用设计文档，包括需求分析、概念设计、逻辑设计、物理设计、数据库实施等内容，提交设计文档到学在浙大(第一次提交，word或pdf，无需压缩，支持系统直接查看)，每位同学都需要提交。

文档内容包括：

1. 需求分析描述
2. 空间扩展E/R图和约束条件
3. 关系及其主键和外键，完全非平凡函数依赖，设计理由
4. 性能描述，索引、视图和触发器，应对高访问量的策略，用户权限描述
5. 应用需求相关的SQL语句
6. 成员分工情况

3. 2022.12.20和2022.12.27课堂报告，每组10分钟报告+3分钟提问，提问的同学有加分，每个问题1分。报告内容与设计文档基本一致，可增加设计文档未要求的内容，课堂报告后将报告PPT提交到学在浙大(第二次提交)，截止日期2022.12.27，每位同学都需要提交。

**考核标准：**基本需求分析完整性、数据库关系与性能设计合理性、SQL语句与问题回答正确性、功能创新性。

**Option B：空间索引实现**

**作业目的：**理解空间数据类型层次结构，熟悉包围盒Envelope在空间查询中作用，熟悉常见的空间计算方法，掌握四叉树和R-Tree的创建，掌握区域查询和最邻近查询方法，理解空间数据查询的过滤和精炼步。

**作业内容：**

1. 基于提供的Geometry、Point、LineString (折线)和Polygon空间数据类型层次结构，完成基本要求(必须完成)：

(1) 实现Point到LineString和Polygon之间的欧式距离计算

扩展要求(至少完成一个)：

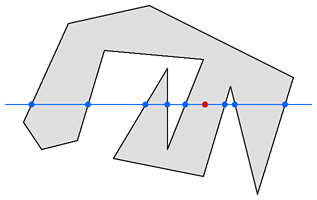
(2) 实现Polygon的内环几何数据存储，并修改Point到Polygon的欧式距离计算

(3) 实现LineString到LineString和Polygon的欧式距离计算

(4) 实现MultiPoint、MultiLineString和MultiPolygon类

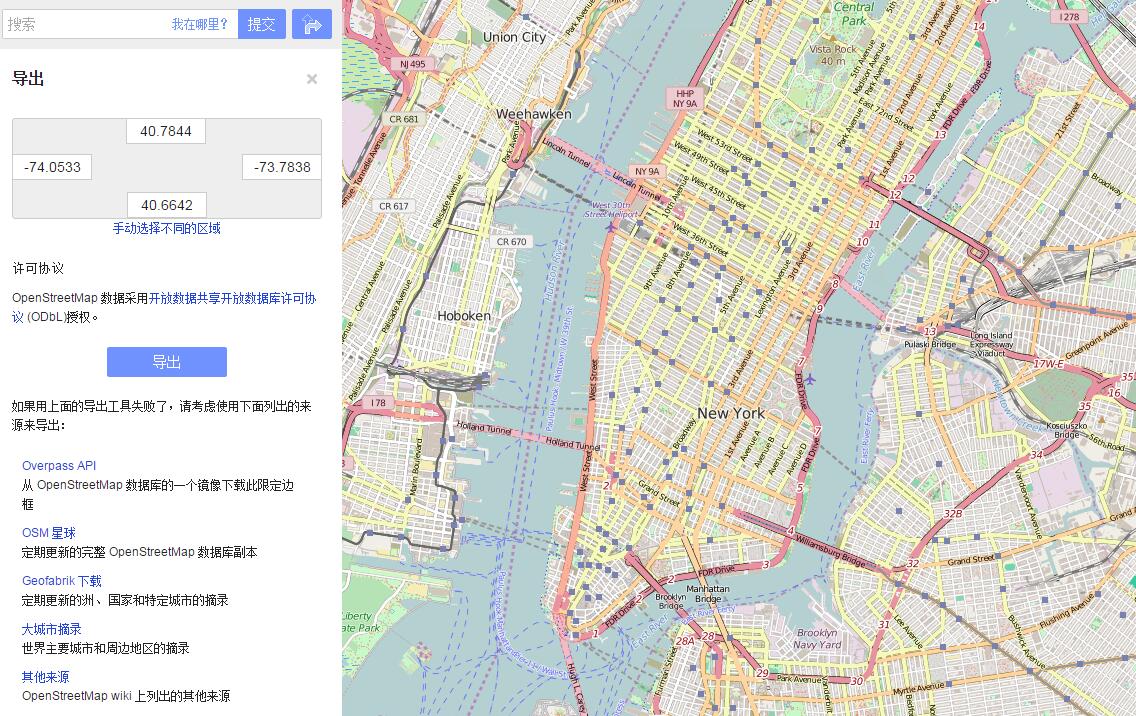
Point到LineString的距离计算分解为Point到每个线段的距离计算，即点P(x, y)到线段[P1(x1, y1), P2(y2, y2)]在二维笛卡尔空间的最短距离。基本思路是通过(P2 - P1)归一化向量和(P - P1)向量的内积，计算P在直线上的投影点，判断该投影点是否在线段上，如果在线段上，计算投影点到P的距离，不在线段上，计算P到线段端点距离的最小值。

Point到Polygon的距离计算关键是判断Point是否在Polygon内部，可以通过射线法判断，参考<https://www.cnblogs.com/luxiaoxun/p/3722358.html>，如果Point在Polygon内部，距离为0，否则计算Point与Polygon边界距离。Polygon定义与PostGIS相同，外环第一个点和最后一个点为同一个点。



2. 基于提供的Envelope类，完成基本要求(必须完成)：(1) contain、intersect和unionEnvelope函数，扩展要求(可选)：(2) 实现Polygon与Envelope相交判断intersects函数。

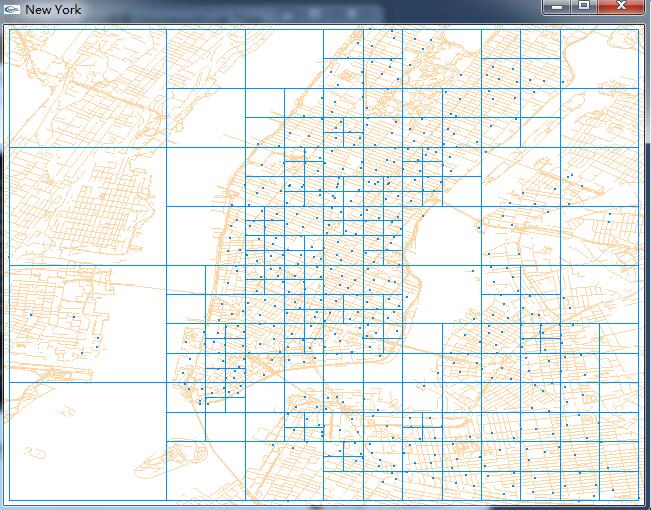
空间索引是基于几何特征的包围盒创建，首先通过判断几何特征的包围盒是否和查询区域相交，降低复杂的几何特征空间关系计算的次数。contain含义和PostGIS中的contain不同，可以是相同的包围盒。



3. 基于提供的QuadNode和QuadTree类实现四叉树创建与查询，完成基本要求(必须完成)：

(1) 实现constructQuadTree和split函数，创建四叉树

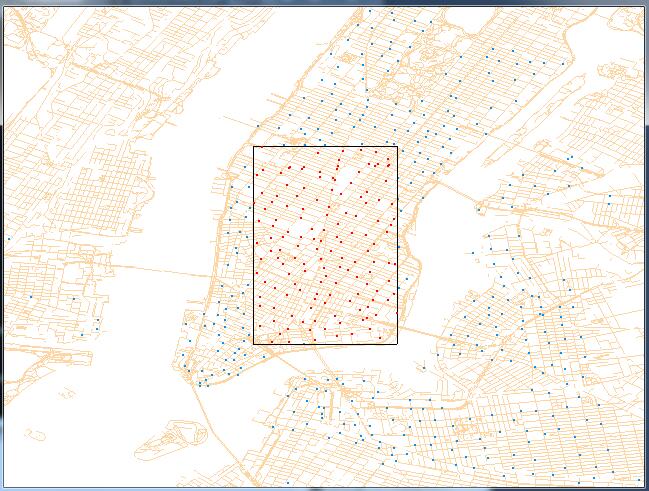
四叉树创建输入一组几何特征，将节点分裂为四个子节点，每个特征加到包围盒重叠的子节点中(即一个特征可能在多个节点中)，删除当前节点的几何特征记录(即所有特征只存储在叶节点中)，如果子节点的几何特征个数大于capacity，递归生成子节点。



(2) 实现rangeQuery函数，完成区域查询

区域查询输入区域rect，查询与区域rect相交的几何特征，存储在features。区域rect如果与当前节点的包围盒bbox相交，递归遍历四叉树，查询哪些几何特征的包围盒和查询区域相交(filter)；再获得可能和查询区域相交的候选几何特征后，精确判断几何特征是否与查询区域相交(refinement)。

通过鼠标选择查询区域，在站点和道路数据上，验证区域查询。



(3) 实现NNQuery和pointInLeafNode函数，完成最邻近几何特征查询

最邻近几何特征查询(NN)输入查询点(x, y)，返回与该点最邻近的几何特征，存储在feature。首先，通过pointInLeafNode查询点(x, y)所在的叶节点，计算查询点(x, y)与该叶节点内的几何特征包围盒的最大距离的最小值minDist，即通过包围盒而非原始几何加速最小距离计算；然后，构造查询区域 (x – minDist, x + minDist, y – minDist, y + minDist)，查询几何特征的包围盒与该区域相交的几何特征(filter)，再查询与查询点(x, y)距离最近的几何特征(refinement)。

通过鼠标移动选择离鼠标最近的几何特征(站点和道路)，验证最邻近几何特征查询。



扩展要求(至少完成一个)：

(4) 基于距离的空间关联 (Spatial Join)

基于距离的空间关联输入两类几何特征的四叉树和空间距离约束条件，输出满足空间距离条件的所有几何特征对，通过公共自行车站点和道路进行验证。

(5) 四叉树性能分析

使用纽约taxi上车点数据分析四叉树性能。四叉树的性能决定于参数capacity，即每个叶节点最多的几何特征数目，数值太多，每个叶节点判断次数太大，数值太小，树的层次太高。当capacity在[70, 200]时，计算四叉树的高度和叶节点数目，评估100000次随机最邻近几何特征查询的时间。分析不同参数下的性能，可以得出什么结论？

4. R-Tree的创建与查询，完成基本要求(必须完成)：

(1) R-Tree创建

R-Tree构建可以采用几何特征逐个插入，基于节点新增面积越小越好的原则选择节点，当超过节点所能存储的最大几何特征时，基于二次分裂(quadratic split)算法，选择最左和最右两个几何特征作为种子点，对几何特征进行分组。R-Tree也可以采用类似四叉树的递归构造方法。

(2) 区域查询 (Range Query)

(3) 最邻近几何特征查询 (Nearest Neighbor)

扩展要求(至少完成一个)：

(4) 基于距离的空间关联 (Spatial Join)

(5) data目录提供了纽约城市道路(highway)、自行车站点(station)和出租车上下车位置(taxi)数据，分析四叉树与R-Tree在不同数据下的性能差异。

5. 扩展要求(可选)：如已实现多边形的相关函数，特别是包围盒与Polygon相交函数，测试多边形数据集的区域查询和最邻近查询。

**作业步骤：**

1. 根据README文档环境配置，完成各任务的代码，并进行测试验证。

2. 完成作业报告，包括原理、实现、测试、分析等，并提供成员分工情况。

**时间安排：**

1. 2022.11.22前完成组队，钉钉群登记组队信息。

2. 2022.12.20前上交代码和作业报告，在学在浙大系统提交文档，每位同学都需要提交。

3. 2022.12.20和2022.12.27课堂报告，每组5分钟报告+3分钟提问，提问的同学有加分，每个问题1分。报告内容包括实现功能列表、算法改进或新功能(如有)、查询性能分析、程序演示等，课堂报告后将报告PPT提交到学在浙大，截止日期2021.12.27，每位同学都需要提交。

**考核标准：**基本要求和扩展要求功能实现正确性、程序测试完整性、空间查询高效性、课堂报告问题回答正确性等。

**Option C：空间数据分析**

该选题需老师同意，具体内容讨论决定，时间安排与前两个选题相同。

1. 地铁数据分析

地铁是目前城市交通出行的主要工具之一，地铁站突发客流量激增或意外情况时，容易造成安全隐患，并对市民出行造成极大影响。因此，以“地铁系统韧性评估”为课题，基于地铁站的历史刷卡数据，分析和可视化站点的重要性以及站点之间的相互影响关系，帮助地铁运营部门和公安机关实时监测和评估地铁系统的运营状态，提前部署站点安保措施，并不断优化地铁线路，最终实现智慧城市公共交通建设。实现的系统类似<https://rdcu.be/b5Kxk>，具体分析内容包括（不限于）：

* 不同站点随时间演化的重要性及异常分析
* 站点之间随时间演化的关联性分析
* 某个站点的异常对于其他站点随时间演化的影响分析

2. GeoQueryVis

参考QueryVis (<https://queryvis.com/)>，实现包含空间函数的查询语句可视化，具体可视化内容包括（不限于）：

* 查询过程可视化（理解查询流程，性能分析，索引建议）
* 拓扑关系可视化（空间数据可视化，拓扑关系示意图，修正错误的拓扑关系）
* 查询结果可视化

3. 其他选题

查看VLDB、SIGMOD等数据库会议近五年的论文，选择一篇空间数据相关的论文，实现或改进论文方法，或设计新方法解决同类问题，例如基于神经网络的空间索引构建<https://vbn.aau.dk/en/publications/effectively-learning-spatial-indices>