# 实验四：学科论文数据的关系型数据库建模与分析

## 一、实验目的

本实验的目标是将多个学科的科研论文指标数据从CSV文件导入MySQL数据库，并通过规范化设计构建合理的关系型数据库Schema，以支持多维度的科研数据分析。实验最终实现了对学校、学科、国家及区域层面的科研表现进行统计与分析。

## 二、实验环境

1. 操作系统：Windows 10  
2. 开发工具：Python 3.12, MySQL 8.0, MySQL Workbench  
3. 主要库：pandas, pymysql, chardet, python-docx  
4. 数据文件路径：C:\Users\xze97\Desktop\Programming\python\lab\第四次作业\data

## 三、实验步骤

（1）数据导入：

编写 import\_to\_mysql.py 脚本，自动检测CSV编码、清洗列名并导入MySQL表 disciplines 中。

（2）数据库规范化：

设计三张关系表：subjects（学科表）、institutions（机构表）、metrics（指标表），满足第三范式（3NF）要求，消除数据冗余，优化查询性能。

（3）数据迁移：

编写 normalize\_data.py 脚本，从原始 disciplines 表中自动拆分出唯一学科和机构，插入对应表并建立外键关联，将科研指标数据写入 metrics 表。

（4）SQL分析：

通过编写SQL语句，实现以下三个核心查询：  
① 获取华东师范大学在各个学科的排名；  
② 获取中国（大陆地区）大学在各个学科的整体表现；  
③ 分析全球不同区域在各个学科的科研表现。

## 四、数据库Schema设计

1. subjects表

subject\_id INT AUTO\_INCREMENT PRIMARY KEY  
subject\_name VARCHAR(255) UNIQUE

2. institutions表

institution\_id INT AUTO\_INCREMENT PRIMARY KEY  
institution\_name VARCHAR(255)  
country VARCHAR(255)  
region VARCHAR(255)

3. metrics表

metric\_id INT AUTO\_INCREMENT PRIMARY KEY  
subject\_id INT FOREIGN KEY → subjects(subject\_id)  
institution\_id INT FOREIGN KEY → institutions(institution\_id)  
rank INT, wos\_documents INT, cites INT, cites\_per\_paper FLOAT, top\_papers INT

## 五、SQL查询与分析

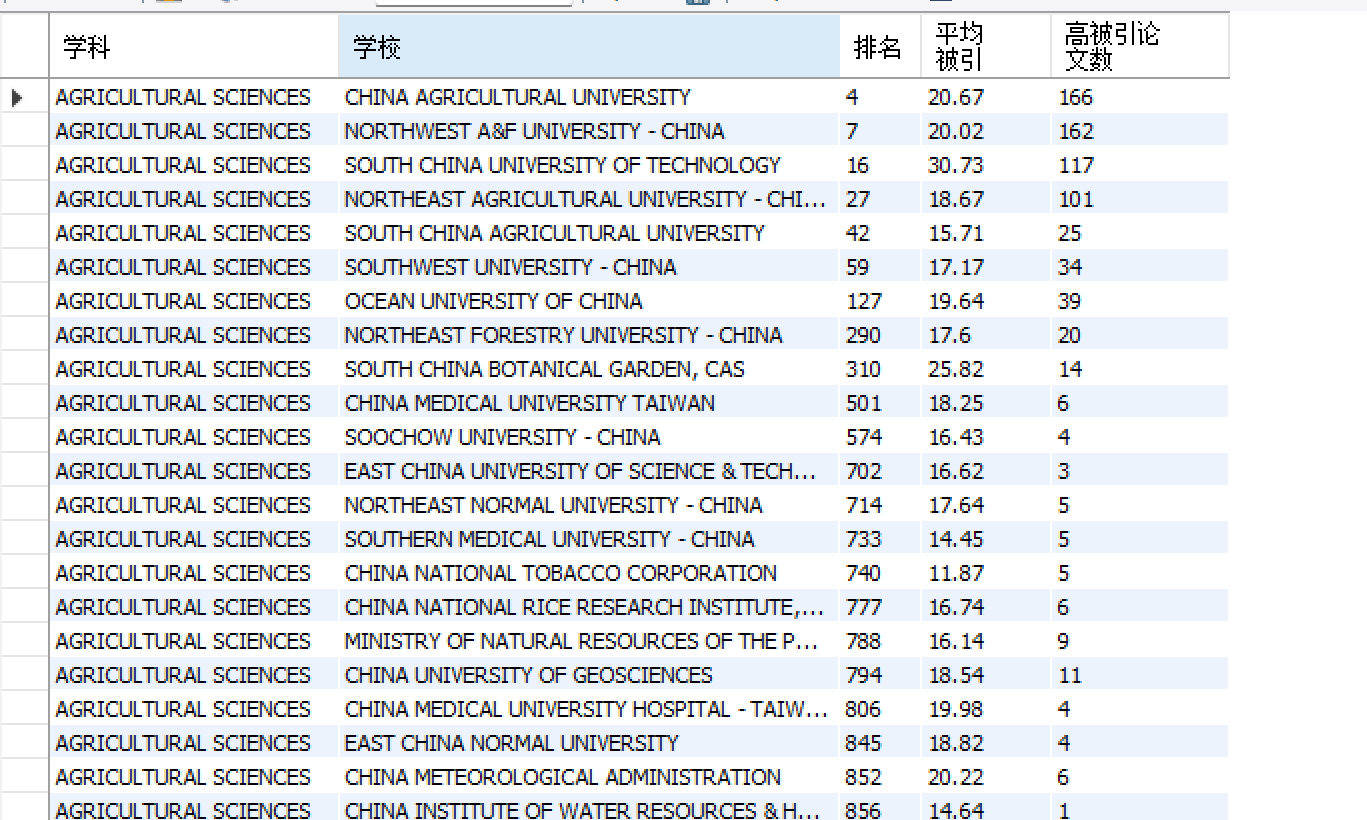
1. 华东师范大学在各学科的排名：

SELECT s.subject\_name, m.rank, m.cites\_per\_paper, m.top\_papers  
FROM metrics m  
JOIN subjects s ON m.subject\_id = s.subject\_id  
JOIN institutions i ON m.institution\_id = i.institution\_id  
WHERE i.institution\_name LIKE '%EAST CHINA NORMAL UNIVERSITY%';



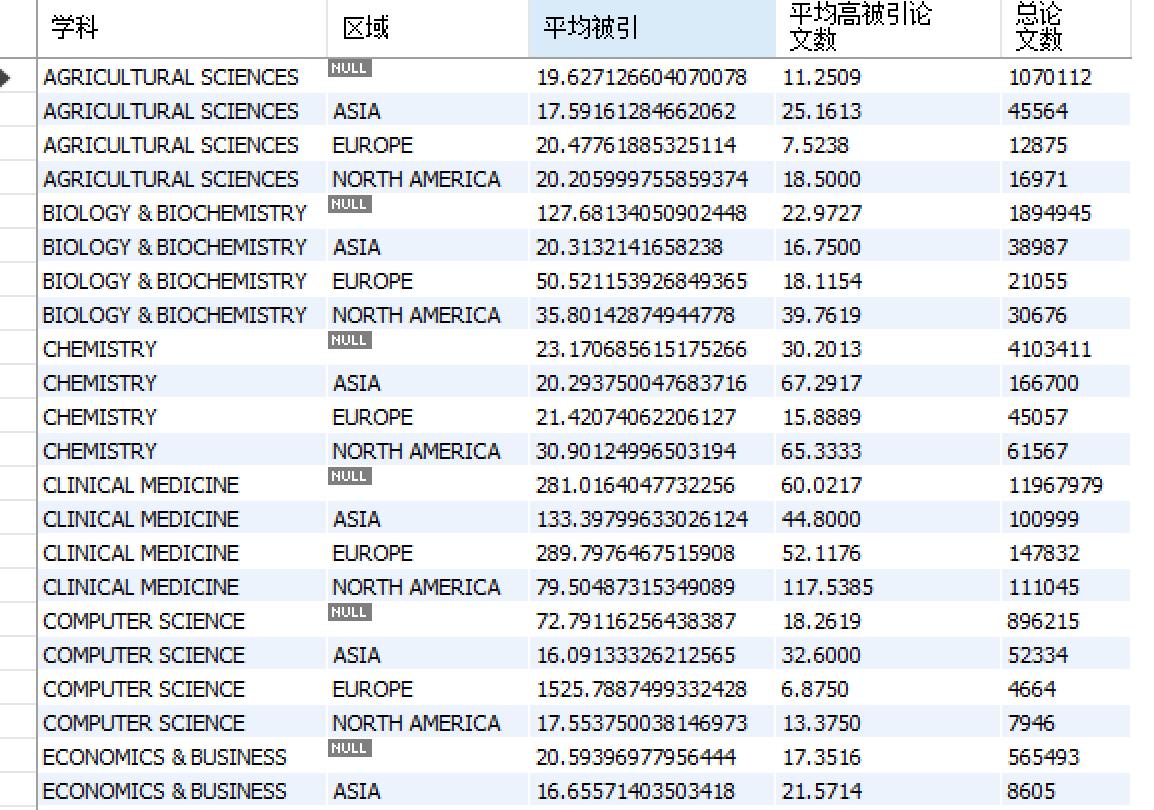
2. 中国（大陆）大学的学科表现：

SELECT s.subject\_name, i.institution\_name, m.rank, m.cites\_per\_paper, m.top\_papers  
FROM metrics m  
JOIN subjects s ON m.subject\_id = s.subject\_id  
JOIN institutions i ON m.institution\_id = i.institution\_id  
WHERE i.country = 'CHINA MAINLAND'  
ORDER BY s.subject\_name, m.rank;



3. 全球不同区域的科研表现：

SELECT s.subject\_name, i.region, AVG(m.cites\_per\_paper), AVG(m.top\_papers), SUM(m.wos\_documents)  
FROM metrics m  
JOIN subjects s ON m.subject\_id = s.subject\_id  
JOIN institutions i ON m.institution\_id = i.institution\_id  
GROUP BY s.subject\_name, i.region;



## 六、实验结果与分析

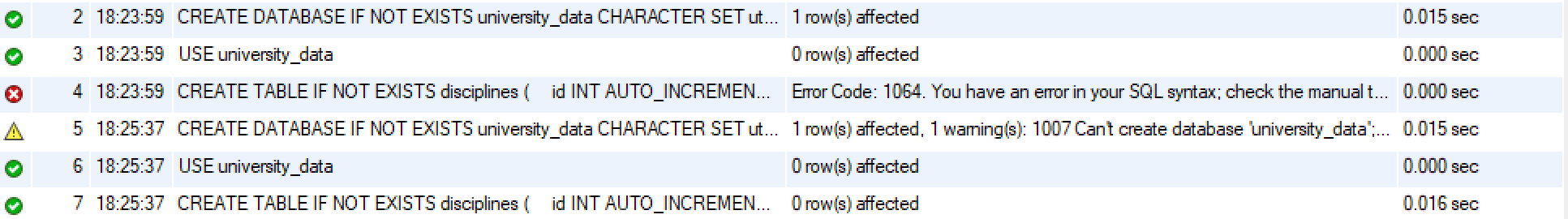
实验成功将22个学科的科研指标数据（共34121条记录）导入MySQL数据库，并通过关系模式优化实现高效查询。实验结果表明，关系型数据库设计显著减少了数据冗余，提升了跨学科、跨地区分析的灵活性。

## 七、实验总结

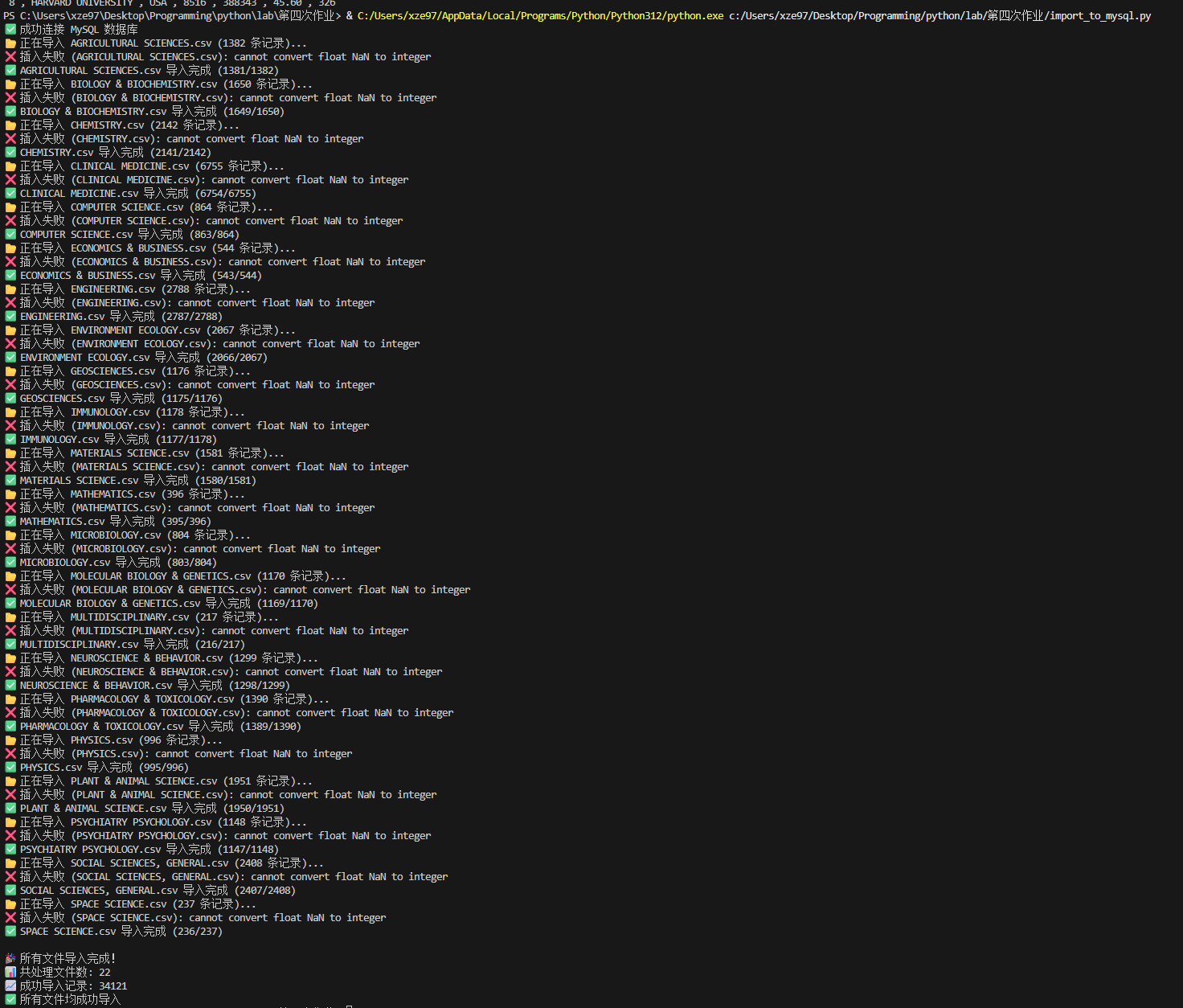
通过本实验，掌握了数据导入、数据库规范化建模与SQL分析查询的完整流程。理解了第三范式的重要性及其在科研数据管理中的实际价值。实验成果为后续学科科研表现分析和可视化奠定了坚实基础。

## 八、运行截图

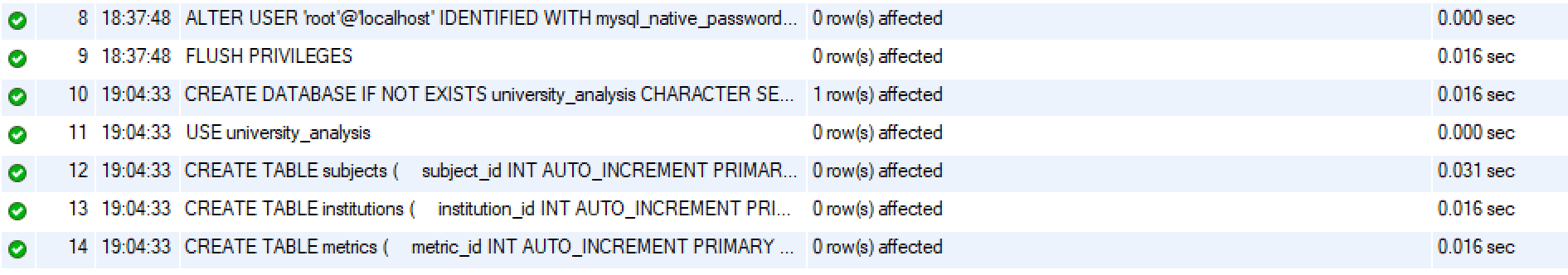
1.创建数据库



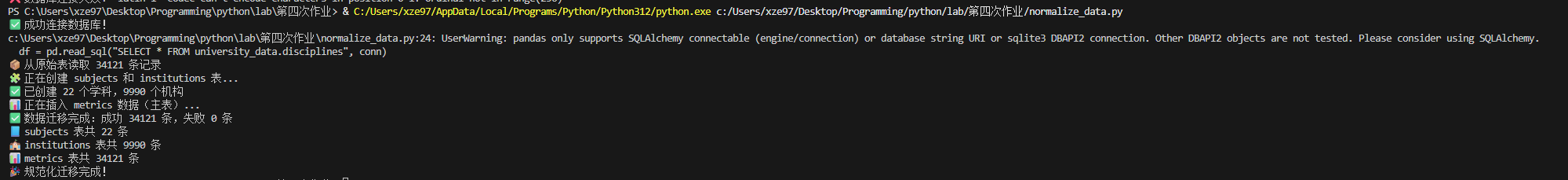
1. 初步导入数据



1. 创建表格



1. 规范化迁移



1. test1、2、3

4e991135e3da539b9338b438d9714447

ef0663fc876b8a3849a8708a39b70ac3

1504986cd2f8534cd349dc301038302c