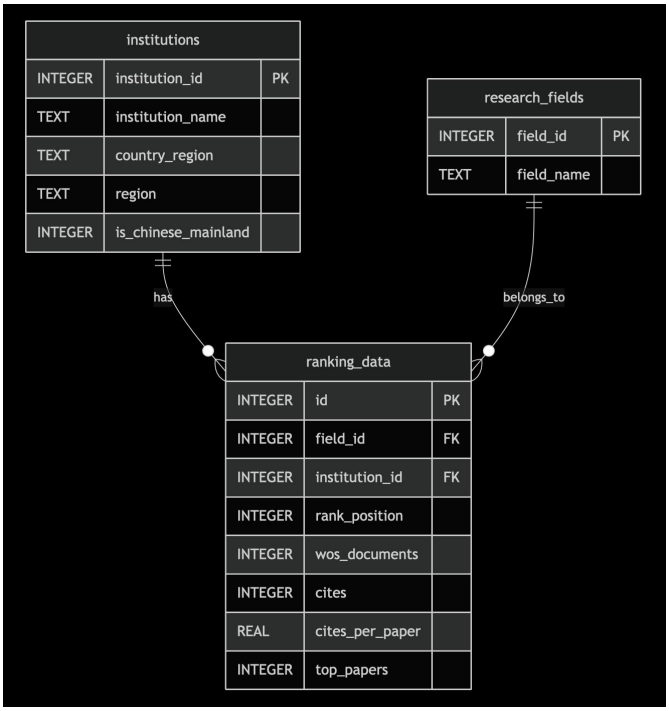
数据科学导论实验 4：基于 SQL 的 ESI 学科数据分析

学号：**10245501405**

姓名：杨云天

一、关系模型 Schema 的构建

1. 关系模型图



2. 关系模型说明

・机构表 (institutions)
主键：institution_id
包含机构的基本信息（名称、国家地区、区域（洲）等）

・学科表 (research_fields)
主键：field_id
存储学科分类信息（学科名称）

・排名数据表 (ranking_data)
主键：id
包含两个外键：field_id 引用 research_fields(field_id)
institution_id 引用 institutions(institution_id)

存储具体的排名指标数据（世界排名、Web of Science 文档数、引用次数、篇均引用、顶尖论文数）

・关系类型：
institutions 与 ranking_data: 一对多
research_fields 与 ranking_data: 一对多

二、基于 Python+SQLite3 的数据库构建

1. 初始化数据库

通过运行 schema_sqlite.sql 文件，自动在 SQLite 数据库中创建机构表（institutions）、学科表（research_fields）、排名数据表（ranking_data）以及相关索引。

2. 读取 CSV 文件

遍历 csv 目录，依次打开 22 个不同学科排名的 CSV 文件。

3. 数据清洗与预处理

对读取的数据进行预处理，包括去除多余的引号和空白字符、将机构名和学科名统一规范（如去重、去空格）。

4. 批量插入数据

针对每一学科或机构的数据，利用 get_or_create 逻辑避免重复插入学科或机构记录。

5. 检查数据库内容

（1） 查看所有学科 fields_df

学科数量：22

	field_id	field_name
0	1	Agricultural Sciences
1	2	Immunology
2	3	Materials Science
3	4	Mathematics
4	5	Microbiology
5	6	Molecular Biology & Genetics
6	7	Multidisciplinary
7	8	Neuroscience & Behavior
8	9	Pharmacology & Toxicology
9	10	Physics
10	11	Plant & Animal Science
11	12	Biology & Biochemistry
12	13	Psychiatry Psychology
13	14	Social Sciences, General
14	15	Space Science
15	16	Chemistry
16	17	Clinical Medicine
17	18	Computer Science
18	19	Economics & Business
19	20	Engineering
20	21	Environment Ecology
21	22	Geosciences

（2） 查看机构统计 institutions

机构统计：

	total	chinese_mainland	countries
0	9990	859	140

三、华东师范大学在各个学科中的排名

1. SQL 核心语法逻辑

SQL 关键字/语法	作用说明	应用原因
SELECT DISTINCT	查询并去除重复记录	避免因数据异常导致同一学科出现多条重复记录，确保结果唯一性
AS 别名	为查询结果的列指定中文别名	提高数据可读性
JOIN ... ON ...	多表连接查询, 关联三张表的数据	数据分散在不同表中，需通过外键关联获取完整信息。 关联逻辑： - ranking_data（排名表） - institutions（存储机构信息，通过 institution_id 关联） - research_fields（存储学科信息，通过 field_id 关联）
WHERE ... LIKE	模糊匹配筛选条件	可确保精准找到目标机构
ORDER BY	按排名位置升序排列	优先展示排名最靠前的学科

2. 查询结果（按排名位置升序排列）

华东师范大学共进入17个学科排名

	学科	排名	文献数	引用数	篇均引用	顶级论文数
0	Chemistry	90	5420	164390	30.33	157
1	Mathematics	115	2019	11984	5.94	22
2	Environment Ecology	130	2941	92088	31.31	101
3	Materials Science	196	2720	93969	34.55	57
4	Computer Science	207	1803	22336	12.39	25
5	Geosciences	275	1850	42158	22.79	38
6	Social Sciences, General	314	2176	27524	12.65	51
7	Engineering	317	2567	55450	21.60	86
8	Plant & Animal Science	395	1375	21843	15.89	26
9	Psychiatry Psychology	467	1460	15243	10.44	7
10	Physics	522	3495	50802	14.54	47
11	Biology & Biochemistry	721	897	20837	23.23	18
12	Agricultural Sciences	845	346	6513	18.82	4
13	Neuroscience & Behavior	853	771	14295	18.54	7
14	Molecular Biology & Genetics	867	532	20568	38.66	6
15	Pharmacology & Toxicology	1064	289	5693	19.70	5
16	Clinical Medicine	2852	940	16875	17.95	12

四、中国（大陆地区）大学在各个学科中的表现

1. 按上榜机构数统计

（1） SQL 核心语法逻辑

SQL 关键字/语法		作用说明	应用原因
聚合函数	COUNT( )	统计每个学科的中国大陆机构数量	
	SUM( )	计算总和（文献数、引用数、顶级论文数）	
	AVG( )	计算平均值（篇均引用）	
	ROUND(..., n)	保留 n 位小数	提高可读性
WHERE		`is_chinese_mainland = 1` 精确筛选中国大陆机构	与 LIKE 模糊匹配不同，该字段使用布尔字段进行精确判断，效率更高
GROUP BY		按学科名称分组，将同一学科的所有记录聚合在一起	统计所有机构每个学科的整体表现，而非单个机构
ORDER BY DESC		按上榜机构数从多到少排列	优先展示中国大陆机构数量最多的学科

（2） 查询结果（按上榜机构数降序排列）

中国大陆共在22个学科中有机构上榜

	学科	上榜机构数	文献数	引用数	篇均引用	顶级论文数
0	Engineering	534	1503716	24966306	17.14	24675
1	Chemistry	423	1249235	30165345	21.43	26271
2	Materials Science	375	1119272	32354021	26.40	21769
3	Environment Ecology	337	489374	11317663	22.54	9594
4	Clinical Medicine	259	972240	16535599	26.10	15613
5	Agricultural Sciences	251	274214	5080975	17.75	4615
6	Plant & Animal Science	248	313104	5066265	16.09	7763
7	Biology & Biochemistry	206	310748	6218009	20.90	4173
8	Computer Science	190	325041	5330017	17.07	5882
9	Pharmacology & Toxicology	182	217471	3502876	16.15	2556
10	Social Sciences, General	180	131047	1897518	14.87	4812
11	Geosciences	177	475790	8302420	17.91	7381
12	Molecular Biology & Genetics	115	232132	6896738	31.06	3318
13	Physics	112	551518	9516603	16.61	9832
14	Microbiology	92	81415	1672848	24.36	1347
15	Mathematics	86	124005	942827	8.20	2367
16	Immunology	84	80843	1681473	24.73	1050
17	Neuroscience & Behavior	71	115210	2180233	18.97	1591
18	Psychiatry Psychology	57	55865	711797	12.98	656
19	Economics & Business	55	62617	988422	16.01	1372
20	Multidisciplinary	17	2665	134168	70.39	118
21	Space Science	10	45825	993908	23.75	778

2. 各学科中国大陆机构进入全球前 10/50/100 排行榜数量

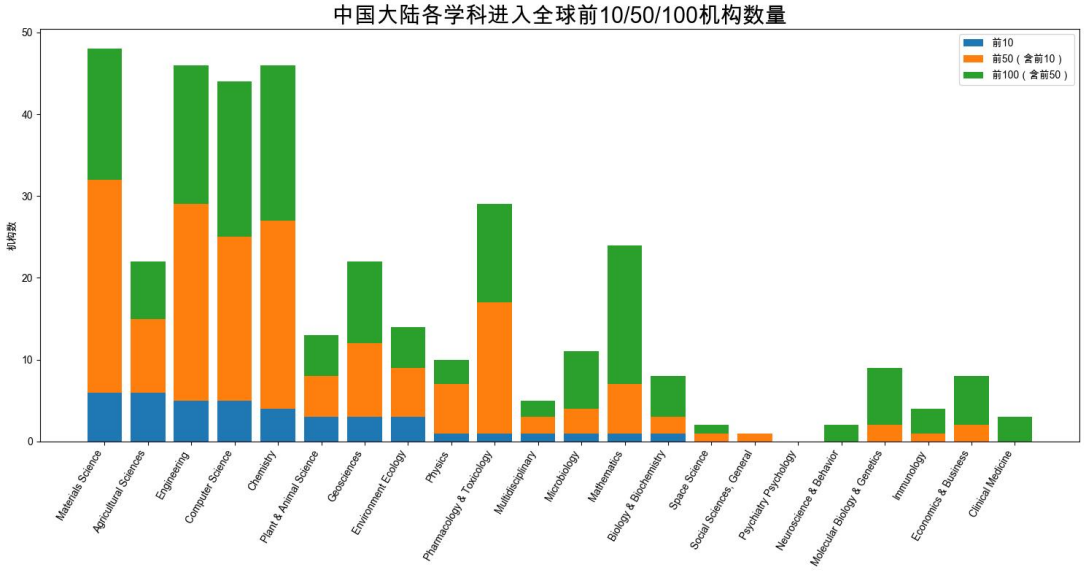
（1） SQL 核心语法逻辑

SQL 关键字/语法	作用说明	应用原因
CASE WHEN 条件分支 CASE WHEN 条件 THEN 返回值 1 ELSE 返回值 2 END	条件判断，根据排名范围返回 0 或 1	实现分段统计，判断机构是否进入某个排名区间

（2） 统计结果与可视化

各学科中国大陆机构进入全球前10/50/100排行榜数量

	学科	前10	前50	前100
0	Materials Science	6	32	48
1	Agricultural Sciences	6	15	22
2	Engineering	5	29	46
3	Computer Science	5	25	44
4	Chemistry	4	27	46
5	Plant & Animal Science	3	8	13
6	Geosciences	3	12	22
7	Environment Ecology	3	9	14
8	Physics	1	7	10
9	Pharmacology & Toxicology	1	17	29
10	Multidisciplinary	1	3	5
11	Microbiology	1	4	11
12	Mathematics	1	7	24
13	Biology & Biochemistry	1	3	8
14	Space Science	0	1	2
15	Social Sciences, General	0	1	1
16	Psychiatry Psychology	0	0	0
17	Neuroscience & Behavior	0	0	2
18	Molecular Biology & Genetics	0	2	9
19	Immunology	0	1	4
20	Economics & Business	0	2	8
21	Clinical Medicine	0	0	3



五、全球不同区域在各个学科中的表现

1. 区域整体表现统计

（1） SQL 核心语法逻辑

SQL 关键字/语法	作用说明	应用原因
复杂 CASE WHEN - `WHEN ... = ...`：单一条件判断（如 = 'CHINA MAINLAND'） - `WHEN ... IN (...)`：多值匹配判断（如北美、欧洲等） - `ELSE`：兜底条件，处理未明确分类的国家	将 140 个国家/地区映射到 7 个区域（按大洲）	原始数据按国家存储，需要聚合到区域级别进行分析
GROUP BY 多字段分组	按学科和区域双重分组： - 第一层：按学科分组（22 个学科） - 第二层：每个学科内按区域分组（7 个区域）	分析每个学科在不同区域的表现差异

（2） 统计结果（部分）

	学科	区域	机构数	文献数	引用数	篇均引用	顶级论文数
0	Agricultural Sciences	欧洲	427	301154	6317941	21.82	4525
1	Agricultural Sciences	中国大陆	251	274214	5080975	17.75	4615
2	Agricultural Sciences	北美	213	206958	4048060	21.76	3065
3	Agricultural Sciences	亚洲其他	201	145518	2174148	16.62	1615
4	Agricultural Sciences	其他	115	57003	1026449	21.69	775
...	...	...	...	...	...	...	...
171	Space Science	中国大陆	10	45825	993908	23.75	778
172	Space Science	大洋洲	9	17648	715001	40.80	685
173	Space Science	其他	7	17844	821890	51.54	773
174	Space Science	南美	4	11871	446086	37.51	340
175	Space Science	非洲	1	1155	58829	50.93	50

176 rows × 7 columns

2. 各区域在前 100 名中的机构数量

（1） SQL 核心语法逻辑

SQL 关键字/语法	作用说明
WHERE 过滤 + GROUP BY 分组 - WHERE：先筛选出前 100 名的记录 - GROUP BY：再对筛选结果进行分组统计	分析顶尖机构的区域分布特征

(2) 统计结果与可视化（颜色越深代表该区域在某学科中进入前 100 名的机构数量越多）