Updating Database via Scrapy

Yanqiao Chen(12412115), SUSTech, Shenzhen
Instructor: Dr. Shiqi Yu

December 8, 2025

Contents

1 项目基本信息

1.1 使用工具

工具	版本	说明
Python	3.8.10	编程语言
Scrapy	2.5.1	爬虫框架
PostgesSQL	13.3	数据库管理系统

1.2 项目目标

通过爬虫爬取电影数据库 tmdb 电影和演员数据，更新旧有教学电影数据库中信息（自 2019 以后，根据查询可知电影数据库信息截至 2019 年）。限于 API 爬取速度限制，我只爬取了 2019 年以后的部分电影数据和其部分演员数据，如有需要更新之前的数据，只需要调整爬虫代码中的年份范围以及爬取限制即可，但是这可能需要更长的事件或者使用反爬机制，有可能有法律风险，因此此处并不尝试。

限于 API 速率限制，以及为“教学用”（轻量化、小型化）数据库的考量，我们只爬取每个年份的部分电影数据（数量不等）。如果后续需要增量式更新，本爬虫亦可以通过修改 peopleid 起始以及年份范围来进行增量式更新。

1.3 更新内容

更新了三个表单：movies, credits, people, 并且向 countries 中插入了“??”作为数据库未知国家的占位符。同时注意到，现有数据库中的国家代码”sp” 不符合 ISO 3166-1 标准，但是为了统一修改，将读取到的西班牙国家代码”es” 全部替换为”sp”。同时，我们添加了巴勒斯坦”ps” 作为国家代码。

2 项目过程

整体项目过程如下：

1. 向 tmdb 数据库申请 api 密钥
2. 编写 Scrapy 爬虫，爬取 2019 年以后的电影数据
3. 使用 psycopg2 导入数据库

2.1 申请 API

在 tmdb 官网注册账号后，进入设置页面，申请 API 密钥。申请过程中需要填写一些基本信息以及使用目的等。申请成功后，可以在设置页面查看到 API 密钥。

2.2 编写爬虫

我们本次使用 Scrapy 框架编写爬虫。首先创建一个新的 Scrapy 项目，然后创建一个新的爬虫，命名为 tmdb_spider 并存储在 spiders 文件夹下。然后，我们配置 settings.py 文件和 pipeline.py 文件，以便将爬取的数据存储到数据库中。最后，我们编写 tmdb_spider.py 文件，定义爬虫的行为和数据处理逻辑。

一个典型的爬虫框架如下：

- start_requests: 定义初始请求，通常是从某个 URL 开始爬取数据。
- parse: 处理响应数据，提取所需信息，并生成新的请求。
- item_pipeline: 处理提取的数据，进行清洗、存储等操作。

根据主要的几个表格的关系，我们首先从电影本身的信息开始爬取，然后对于每个电影，爬取其演员和工作人员的信息。最后，将所有数据存储到数据库中。

本次项目中，我进行了如下的操作：

1. 发起请求，获取 2019 年以后的电影列表。
2. 将请求转换为 JSON 格式，便于数据处理。
3. 解析电影数据，提取电影 ID、标题、简介、发布日期、评分等信息。
4. 对于每部电影，发起新的请求，获取其演员和工作人员信息。
5. 解析演员和工作人员数据，提取演员 ID、姓名、角色等信息。
6. 将提取的数据存储到数据库中。同时，为了避免重复存入演员信息（保证唯一的 peopleid），我们使用构造 sql 查询语句检查演员是否已经存在于数据库中。如果已经存在，那么 credit 表中应当使用已有的 peopleid，否则插入新的演员信息并获取新的 peopleid。

在实际爬取过程中，因为一些设置上的问题，我们进行了多次不同年份的爬取。由于单个年份的电影数据较多，且电影网站往往将某一年份的电影持续推送，因此我们在以后的爬取中可能需要设置单个年份的爬取上限以满足各年份均匀分布的要求。否则，爬取到的电影数据可能会集中在某些年份，导致数据库更新不均衡；如果需要更新全部 2019 年以后的数据，则需要更长的时间和更复杂的反爬机制。

2.3 数据库更新结果

最后结果显示，我们成功更新了 2019 年电影 4426 条，涉及演员和导演近两万人，并且成功将数据存储到数据库中。数据库的完整性和一致性得到了保证，所有数据均符合预期格式和要求。对于那些国家信息缺失的电影，我们使用了“??”作为占位符，确保数据库的完整性。

数据详见附录中的 SQL 文件。

3 课程资源的评估

3.1 Some Comments on the Lecture Notes

Summary Slides 在阅读 Lecture Notes 的时候，我发现在大部分 Slides 中，内容比较分散且不够系统化。在重新阅读复习的过程中，需要花费很大的精力区分例子和关键定义与注意项目。建议在 Lecture Notes 中，根据 Chapter 中不同的概念，增加一页 Summary 进行重点内容的总结和归纳，以此方便在课堂中以及复习过程中方便阅读和理解。

Slides 中的显示错误

- Chapter 2 Slide 13, 14, 15, 29, 61: 文字标题显示重叠；
- Chapter 2 Slide 23: 文字显示重叠；
- Chapter 9 Slide 1, 标题 Trigger 少打了一个 g；

笔误

- Chapter 2 Slide 38, 34: not null 前面错误的添加了逗号，导致语法错误；

一些补充更新

- Chapter 2 Slide 20: 实际上，许多的 SQL 方言可以开启严格模式，从而阻止一些可能不符合逻辑或者可能导致错误的插入，例如 MySQL 的 STRICT_ALL_TABLES 模式；不过，数据库在严格模式下仍然不会符合关系模型中的所有约束；
- Chapter 2 Slide 36: 长段注释语法“/* */”可以补充说明；
- Chapter 3 Slide 30, 第三行在部分数据库的日期转换中可能返回 null，移植性较差；

- Chapter 4 Slide 19，可以在 PPT 中明确附加 where 和 having 的执行顺序（where 对原始行过滤，having 对 group by 后的各组进行过滤）；
- Chapter 4 Slide 27, 可以在 PPT 中附加 count(*) 和 count(col) 对比：count(*) 不会检查数据中是否保存了 null，count(col) 会检查数据行对应的列中是否包含 null；
- Chapter 5 Slide 61, 可以在 PPT 上强调关联子查询和普通子查询效率上存在区别的原因，例如给出一个具体的 SQL 查询例子；
- Chapter 5 Slide 75，可以增加对于此类运算和解释：由于逻辑短路，对于 and 运算，如果第一个是 False，呢么就是 False；如果第一个是 True，还需要检验第二个；
- Chapter 7 Slide 8，介绍全文搜索的时候，可以介绍其优势，例如，全文搜索可以利用倒排索引提高搜索效率，因此其比 like 的搜索效率更高；
- Chapter 7，可以增加对于 MVCC 的介绍；
- Chapter 7 Slide 81, 可以增加对于 with 语法中，指定分隔符等额外选项的语法；

3.2 Some Comments on the Original Filmdb

关于国家编号不符合 ISO-3166-1 的提示 在更新数据库的过程中，我发现现有 filmdb 数据库中的国家编号不符合 ISO-3166-1 标准。例如，西班牙的国家代码应为"es"，但在数据库中使用了"sp"。为了保持一致性，我在爬取数据时将西班牙的国家代码统一替换为"sp"。建议在数据库设计和维护过程中，严格遵守国际标准，以避免混淆和错误。由于 filmdb 中的数据库中原始数据来源不明，因此为了减小对于数据库现有数据的影响，在更新的过程中保持与数据库内部一致。但是建议在后续的数据库设计中，遵循国际标准。

4 一些课程建议

4.1 关于 Lecture

在 Lecture 部分，我认为可以增加一些对于数据库本身架构开发的内容，例如，对于 PostgreSQL 代码的观察和分析，以及这些代码如何实现数据库的基本功能和特性。可以适当地减少对于 SQL 语法本身细节的讲解，而增加在 Lab 和 Project 中增加对于 SQL 语法细节的练习（例如，复杂查询、语法细节等等），如果能够更多地上手练习，对于理解 SQL 语法和数据库设计会有更大的帮助。

另外，从 CMU-15-455 课程中，我发现可以增加一些对于数据库系统内部实现的内容，而不是仅仅停留在 SQL 语法和数据库架构的层面。通过了解数据库系统的内部实现原理，可以更好地理解数据库的性能优化和设计原则。例如，CMU 数据库课程的第一个 Project 是编写一个 Hyperloglog 算法来估计数据的基数，这对于理解数据库中的数据处理和优化有很大的帮助，我们可以将此类算法的实践作为 Lab 的一部分内容。

此外，增加一些对于现代数据库前沿的内容，例如分布式数据库、并发控制等内容，可以帮助学生了解数据库领域的最新发展和趋势。

另外，我觉得这门课开在图灵班大二上实际上不太好，因为大二的计算机学生通常对于计算机软件开发的知识尚不成熟，语言熟练度尚且不足，对于数据库系统的理解也比较浅显。正因为此，许多同学对于数据库的理解止步于 SQL 语言本身，对于数据库的重要性和实际应用缺乏深入了解。在 CMU-15-455 课程中，该课程要求学生先修计算机系统课程以及 C++ 语言课程，并且在课程初期有一个 Pass or Fail Project （只有拿到满分，才能够继续课程，否则将被劝退）。我认为我们在课程的初期也可以设置类似的门槛，充分利用南科大前四周的退课时间窗口，确保选课的同学对于数据库系统有足够的兴趣和基础，从而提高课程的整体水平和学习效果。；同时，如果学生学习过计算机组成原理和计算机系统，他会知道为什么把需要查询的数据缓存在内存上会更快，从而更好理解数据库设计的动机；如果学生学习过操作系统，他会更好地理解并发控制和事务管理的原理，从而更好地理解数据库系统的设计和实现；如果学生学习过编译原理，他会更好地理解 SQL 查询优化器的工作原理，从而更好地理解数据库系统的性能优化。

如上，可以表明数据库本身实际上是一个需要大量前置课程的课程，因此我建议将数据库课程安排在大三上学期，这样，按照图灵班的进度，学生在大三上学期之前已经学习过计算机系统、操作系统、编译原理等课程，从而为数据库课程打下坚实的基础。

我相信这样的调整会显著提升学生对数据库系统的理解和兴趣，从而提高课程的整体质量和学习效果。

4.2 关于 Lab

在 Lab 部分，我认为可以分为两个部分：SQL 语法练习和数据库编程练习。这些练习不必上交评分，而是作为学生自我练习和巩固知识的机会。通过实际操作，学生可以更好地理解 SQL 语法和数据库设计的原理。同时，我认为这一门课需要几个助教来协助完成 Lab 的设计和实现工作。

同时，让助教制定

4.3 关于 Project

实际上，我们可能需要设计一个课程数据库，让学生在 Lab 中进行实际的数据库设计和实现。通过实际操作，学生可以更好地理解数据库的设计原则和实现方法。例如，CMU-15-455 课程中，课程项目组提供了一个名为 Bustub 的教学数据库系统，学生需要在此基础上实现各种数据库功能和特性。这种实践性的学习方式可以帮助学生更好地理解数据库系统的工作原理和设计思路；至于国内，华中科技大学在数据库课程中提供了一个基于 oceanbase 开发的 miniob 系统，学生需要在此基础上实现各种数据库功能和特性。我认为我们也可以设计一个类似的教学数据库系统，让学生在 Lab 中进行实际的数据库设计和实现。这一份工作可以作为课程助教的长期接续工作。

另外，CMU-15-455 课程中还提供了很多需要阅读的 Paper，我们可以效仿这一点，在 Project 中增加“综述”这一部分内容，让学生阅读一些数据库前沿的论文，并撰写综述报告。这不仅可以帮助学生了解数据库领域的最新发展和趋势，还可以提高他们的学术阅读和写作能力。

如上，我们可以列出设计出的若干 Project 供学生任选:

- 实现数据库的关键优化功能，例如 Hyperloglog 等比较冷门的算法；
- 让学生阅读指定及自行查找指定的论文，撰写综述报告；
- 对数据库感兴趣的同学，可以根据自己的兴趣，进行设计和探索，并且尝试在课程结束后发表论文，这可以作为 Bonus；
- 协助教师设计和实现一个教学数据库系统并开源，为学弟学妹们提供一个更好的学习平台。
- 实现爬虫，爬取指定网站的数据并存储到数据库中。
- 使用小型框架构建一个网站，连接数据库，并且通过模拟攻击测试数据库的安全性。

5 总结

A Python Scrapy 爬虫代码

Listing 1: tmdb_spider.py 爬虫代码

Listing 2: pipeline.py 数据库存储代码

Listing 3: items.py Item 定义代码

Listing 4: settings.py 数据库配置代码

B 更新后的数据库导出的 SQL 文件