豆瓣电影 Top250 数据分析

2020 年 6 月 3 日

1. 数据获取
2. 数据分析
3. 数据展示

数据获取

基本思路

使用 scrapy 框架，通过网络爬虫的方式获取，并将获取到的数据存储到数据库中，方便之后对数据进行分析

数据获取

网页内容分析

通过浏览器查看，各部电影的信息主要分布在电影的详情页，可以通过豆瓣 Top250 的主页跳转

除此之外，分析一部电影的内容还需要使用到他们的评论，各部电影的评论页可以通过电影的详情页的跳转

数据获取

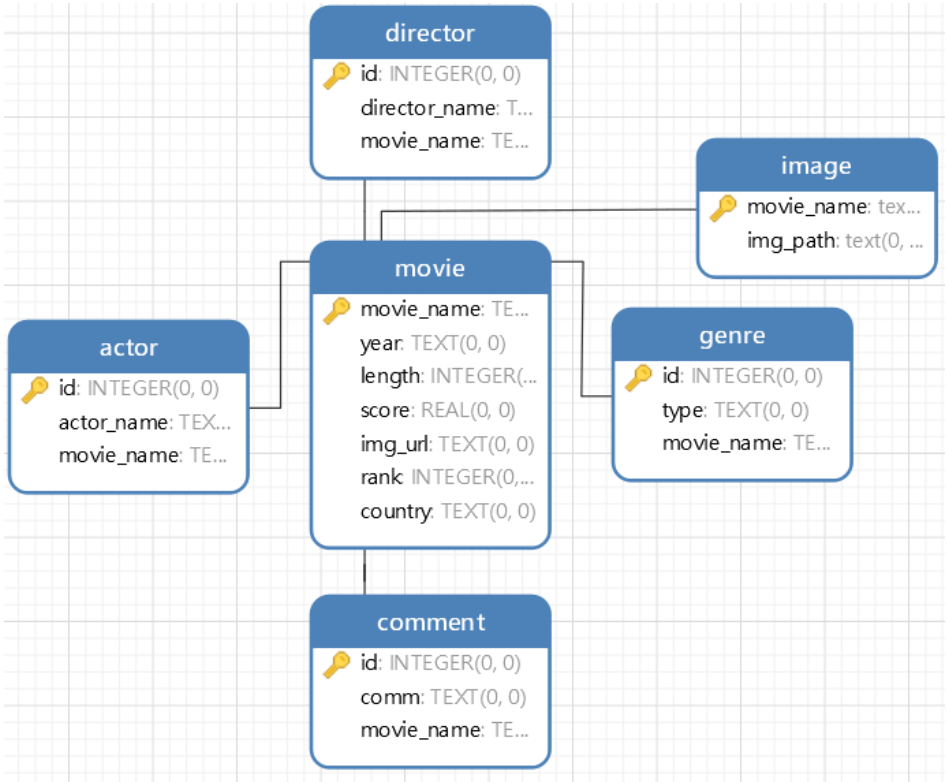
爬虫设计

1. 从豆瓣 Top250 主页开始，爬取各个电影的详情页的超链接
2. 从各个电影的详情页爬取电影的导演，主演，类型等信息，以及评论页的超链接
3. 从各个电影的评论页爬取前 20 条热评（过多的评论会导致数据集过大，并且前 20 条热评已经具有代表性）

数据获取

数据表结构设计

在设计数据表结构中，同时考虑到之后进行数据分析的方便性以及规范化原则，设计了电影，导演，主演，类型，评论，图片共六个数据表



数据分析

基本思路

由于通过爬虫获取数据之后将数据存储在数据库中，因此可以使用 SQL 进行一些简单的初步分析，然后转换为 DataFrame 数据结构再使用 pandas 进行进一步的分析，对于分析的结果通过静态图片或者 web 页面的方式展示

数据展示

豆瓣 Top250 数据分析

数据展示

THANKS