

# 1 引言

## 2 数据集和知识图谱构建

在本项目中,我们选择了 TMDB 作为我们的数据源. 现有的主要的电影数据源包括 TMDB, OMDb, IMDb, MovieLens 等, 其中 TMDB 在电影数据 API 中以其丰富的数据、开放的社区驱动模式和高质量的 API 而著称。它不仅适用于电影和电视节目的信息查询和展示, 还提供了强大的推荐和搜索功能, 是开发电影相关应用的理想选择。相比其他 API, 如 IMDb 和 OMDb, TMDB 在数据的实时更新、社区参与和免费使用方面具有明显的优势。

数据集的爬取主要分为搜索和获取详细信息两步. 首先, 通过在 "https://api.themoviedb.org/3/movie/popular?api\_key={api\_key}&page={page}" 指定 api\_key 和 page (取值于 1 - 100), 我们可以获得 2000 部电影的简要信息; 然后, 在 "https://api.themoviedb.org/3/movie/{movie\_id}?api\_key={api\_key}&append\_to\_response=credits" 中指定 api\_key 和 movie\_id (来源于上一步的简要信息), 可以获得这些电影的详细信息。

获取数据之后, 我们进行了一系列探索性分析。

通过以下两个链接

- 搜索: "https://api.themoviedb.org/3/movie/popular?api\_key={api\_key}&page={page}";
- 电影详情: "https://api.themoviedb.org/3/movie/{movie\_id}?api\_key={api\_key}&append\_to\_response=credits"