爬取意义

**用户偏好洞察与推荐系统优化**：

烂番茄是全球知名的影评聚合网站，其包含了影迷和影评人对电影的评分和评论，这些数据非常适合用来分析用户偏好。通过爬取电影的详细信息（如导演、演员、类型等）以及观众和影评人的评分和评论，可以建立一个更精准的个性化推荐系统。例如，分析特定类型电影中的高评分用户评论，有助于理解用户在特定类别中的偏好，进而用于电影推荐。

**影片制作与发行的市场导向分析**：

爬取烂番茄数据可以帮助电影制作公司和发行方了解市场趋势。通过对电影类型、评分以及观众的情感反应进行数据分析，可以判断哪种类型的电影更受欢迎，从而指导未来的电影制作和市场投放。尤其是用户评分和影评人的评论能够直接反映出电影的优缺点，对于电影的制作、改进以及营销策略的制定有重要的参考价值。

**用户情感分析与文本挖掘**：

烂番茄上的评论内容非常适合进行情感分析和文本挖掘，获取影迷对某部电影的正负反馈。可以通过自然语言处理（NLP）技术分析评论文本，提取出电影的受欢迎元素或观众的不满之处。这种信息对于电影改进、宣传方向的调整有重要的作用。

**影评人与观众反应之间的差异分析**：

烂番茄上不仅有普通观众的评分，还有专业影评人的评论。这些数据可以用来对比影评人和普通观众在电影评价上的差异，帮助理解专业评论与大众接受度的关系，从而为电影推广选择合适的媒体策略。

**影评人信任模型的建立**：

通过爬取影评人的评论，可以建立影评人的信任度模型。某些影评人的评论可能与观众的评价更为一致，而另一些影评人的评分则可能偏离大众偏好，通过对这些信息的分析，可以为观众提供推荐信任的影评人，增强影评的参考价值。

爬取过程

爬取任务的第一步是选择烂番茄的目标网页。

使用Scrapy框架中的CSS选择器提取页面上每部电影和每个影评的链接。提取这些链接后，便可以进一步请求对应的详情页，从中获取更详细的信息。

在电影详情页中，需要提取的内容包括电影标题、描述、类型、时长、导演、演员、用户评分和影评人评分等。这些信息能够为电影推荐系统和市场分析提供基础数据。

在影评页中，需要提取影评人姓名、评论内容、日期、评分等信息。这些内容可以用于建立影评人的信任度模型，帮助用户选择值得参考的评论。

烂番茄网站会限制单位时间内来自同一 IP 地址的请求数量。为了避免触发这种速率限制，我设置了合理的下载延迟以及随机化请求时间间隔，以模拟不规则的用户访问模式。

为了防止单一 IP 地址被封禁，我使用了代理 IP 池，以确保每次请求来源于不同的 IP 地址。这样可以有效绕过网站的 IP 封锁机制，同时也可以提高数据采集的稳定性。

烂番茄网站的部分内容是通过 JavaScript 动态加载的，这使得传统的 HTML 解析工具无法直接获取到数据。为了解决这一问题，我使用了 Selenium 自动化工具，模拟浏览器加载网页并提取动态内容。

最后将爬取到的数据需要保存为CSV格式。

数据清洗过程

爬取的数据包含冗余信息或空数据，数据清洗是必不可少的步骤，需要去除无效数据并标准化字段格式。

过程为：

1.分别去除缺失值（去除缺少评论内容的行）。

2.去除重复的评论。

3.填补评分缺失值，例如可以将缺失的评分替换为一个默认值，例如0。