**个人报告-数据获取**

在这个项目中，我主要负责整个数据获取流程，保证能够准确、高效地从The Numbers 网站中获取到 2020 - 2022 年每一年票房前 100 电影的相关信息并存储在文件中，为后续下面数据分析提供数据基础。

主要的工作流程开展介绍如下：

**1.学习爬虫知识**

我先在网站上搜索了一下python爬虫相关知识，再根据需要项目的需求，需要爬取电影数据，就看了相关的视频，更详细的知道如何去设计相关的内容，途中也换了几个电影网站，因为发现有些网页反爬，不仅不好爬而且爬不出来数据，后来探索了一番，确定了现在的电影网站，也需要加设置 User-Agent 请求头来模拟浏览器行为，减少浏览器阻止爬取数据。

**2.确定和获取数据需求**

确定我们需要从 The Numbers 网站获取的电影数据的内容，包括电影排名、电影名称、票房、上映日期和类型等，确保获取的数据能够满足后面的人可以分析。该项目里是近三年的一个数据即2020-2022年的每一年的票房排行榜数据。

选择相关的库比如Requests 库和BeautifulSoup 库来进行数据爬取，用Requests 库进行 HTTP 请求发送，BeautifulSoup 库解析 HTML 文档内容。通过设计相关函数获取指定年份的 Top 100 电影信息，根据年份动态生成请求地址，确保能够准确获取到对应年份的数据。再设置一个User - Agent 请求头，模拟真实浏览器行为，避免被网站反爬虫机制拦截。使用 BeautifulSoup 库，找到包含电影数据的表格内容，并通过遍历表格行和列，通过设置我们需要的行即前100行和只需要电影排名、电影名称、票房、上映日期和类型、和年份的列，提取出各个电影的相关信息。如果表格列数不足，就会设置默认值 “N/A”。

**3.数据整合和储存**

从网站上获取多个年份电影信息即2020 - 2022年的 Top 100 的电影信息，并将各年份的数据整合到一个列表中。每当一年获取成功后都有fetch top 100 movies for year，可以根据这个知道那一年获取数据成功。并且将每一年获取成功的数据放入到excel表格里面，有利于后面分析的人观察。

工作总结：

我成功获取了 The Numbers 网站上 2020 - 2022 年每年票房前 100 电影的详细信息，一共 300 条有效数据记录，并且和网站数据内容的对比，确保了提取的电影排名、票房、上映日期等关键信息的正确。在数据存储中，我利用 pandas 库实现了高效的数据处理和存储，将获取到的数据成功保存为 Excel 文件。

并且我在项目过程中，遇到了网站反爬虫机制，学会利用User - Agent 等方式模拟浏览器行为，防止爬虫被网站阻止。后续还需要进一步去学习更高级的反爬虫技术，比如使用代理ip等。目前项目代码是比较简单的，后面会深入学习相关内容，提高自身能力，还需要学会代码优化，在编写代码的同时及时添加清晰准确的注释，确保代码的质量和可维护性，从而提高整个项目的开发效率和稳定性。