**1. 项目简介**

该项目用于从 The Numbers 网站获取指定年份的前100部电影的相关信息，包括排名、电影标题、票房、上映日期和电影类型。爬取的数据将保存到一个 Excel 文件 (top\_movies.xlsx) 中。

**2. 环境要求**

操作系统: Windows/Linux/macOS

Python版本: 3.6 或更高版本

依赖库: 以下 Python 库需安装

requests：用于发送 HTTP 请求。

beautifulsoup4：用于解析 HTML 网页内容。

pandas：用于数据处理和将数据保存为 Excel 文件。

**3. 安装步骤**

3.1 安装 Python 依赖库

在命令行中运行以下命令以安装必要的库：

openpyxl 是用于将 Pandas DataFrame 保存为 Excel 文件所需的库。

3.2 下载项目代码

将代码文件保存为 fetch\_top\_movies.py。

3.3 配置环境

确保 Python 环境已正确安装，并且 requests、beautifulsoup4、pandas 库已安装。

3.4 运行项目

在命令行或终端中运行以下命令以执行脚本：

3.5 输出文件

运行脚本后，电影数据将被保存到当前目录下的 top\_movies.xlsx 文件。该文件将包含以下列：

Rank: 电影排名

Title: 电影标题

Box Office: 票房

Release Date: 上映日期

Genres: 电影类型

Year: 电影年份

**代码部分：包括中文注释**

import requests

from bs4 import BeautifulSoup

import pandas as pd

# 获取指定年份的Top 100电影信息

def fetch\_top\_movies(year):

    url = f'https://www.the-numbers.com/market/{year}/top-grossing-movies'

    headers = {

        'User-Agent': 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/85.0.4183.121 Safari/537.36'

    }

    try:

        response = requests.get(url, headers=headers)

        response.raise\_for\_status()  # 检查请求是否成功

        soup = BeautifulSoup(response.text, 'html.parser')# 解析网页内容

        movies = []# 用于存储电影信息的列表

        # 找到包含电影数据的表格

        table = soup.find('table')

        if table:

            rows = table.find\_all('tr')

            for row in rows[1:101]:  # 获取前100行数据

                columns = row.find\_all('td')# 获取每行中的所有单元格

                if columns:

                    rank = columns[0].get\_text(strip=True)# 获取排名

                    movie\_title = columns[1].get\_text(strip=True)# 获取电影标题

                    box\_office = columns[6].get\_text(strip=True)  # 获取票房

                    release\_date = columns[2].get\_text(strip=True) if len(columns) > 3 else 'N/A' # 获取上映日期

                    Genres = columns[4].get\_text(strip=True) if len(columns) > 4 else 'N/A' # 获取电影类型

                    movies.append({

                        'Rank': rank,

                        'Title': movie\_title,

                        'Box Office': box\_office,

                        'Release Date': release\_date,

                        'Genres': Genres,

                        'Year': year

                    })

        return movies

    except requests.exceptions.RequestException as e:

        print(f"Error fetching data for {year}: {e}")

        return []

# 获取多个年份的Top 100电影信息

def fetch\_recent\_top\_movies(years):

    all\_movies = []

    for year in years:

        print(f"Fetching top 100 movies for {year}...")

        movies = fetch\_top\_movies(year)

        all\_movies.extend(movies)

    return all\_movies

def main():

    recent\_years = range(2020, 2023)  # 获取多个年份的Top 100电影信息

    top\_movies = fetch\_recent\_top\_movies(recent\_years)

    if top\_movies:

        # 使用 pandas 将数据写入 Excel

        df = pd.DataFrame(top\_movies)

        df.to\_excel('top\_movies.xlsx', index=False)

        print("数据已成功保存到 top\_movies.xlsx 文件中。")

    else:

        print("未能获取电影数据。")

if \_\_name\_\_ == "\_\_main\_\_":

    main()