功能文档

项目名称：豆瓣电影 Top 250 爬取与分析

项目概述

本项目基于 Python 实现了豆瓣电影 Top 250 的数据爬取、可视化与分析功能，包括获取电影名称、评分、评价人数等核心数据，并对这些数据进行统计与图表展示。

功能模块说明

1. 数据爬取模块

- **模块名称**：`fetch_douban_top250()`
- **功能描述**：从豆瓣电影 Top 250 页面爬取电影数据，包括电影名称、评分、评价人数。
- 输入：
  - 无需输入，内部构造请求 URL。
- 输出：
  - 一个列表，每个元素是包含电影信息的字典（`{"title": ..., "rating": ..., "people_count": ...}`）。
- 实现逻辑：
  - 通过 `requests` 模块发送 HTTP 请求，获取网页 HTML 源码。
  - 使用 `BeautifulSoup` 解析 HTML 内容，提取电影标题、评分和评价人数。
  - 遵守爬虫礼仪，爬取每页后等待 2 秒。
- 关键代码：

```
for page in range(10):  # 每页爬取
    url = f"{base_url}?start={page * 25}"
    response = requests.get(url, headers=headers)
    soup = BeautifulSoup(response.text, "html.parser")
    # 提取信息
    ...
    movies.append({"title": title, "rating": rating, "people_count":
people_count})
```

---

2. 数据可视化模块

- **模块名称**：`visualize_data(movies)`
- **功能描述**：对爬取到的数据进行可视化，生成直观的柱状图、直方图等图表。
- 输入：
  - `movies`：包含电影数据的列表。
- 输出：
  - 图表展示（评分分布图、评价人数 Top 10 图、评分最高 Top 10 图）。
- 实现逻辑：

    - 使用 `matplotlib` 绘制图表，分析评分分布、评价人数和评分最高的电影。
- 关键图表：
    1. 评分分布直方图：
        - 显示评分分布情况，观察评分的集中程度。
    2. 评价人数最多的前 10 部电影：
        - 水平柱状图，展示评价人数最多的电影及其对应人数。
    3. 评分最高的前 10 部电影：
        - 水平柱状图，展示评分最高的电影及其评分。

---

3. 数据分析模块

- **模块名称**：`analyze_data(movies)`
- **功能描述**：对爬取到的数据进行统计分析，包括平均评分、评分区间分布、最低评分电影及相关性分析。
- 输入：
    - `movies`：包含电影数据的列表。
- 输出：
    - 控制台打印统计结果。
    - 图表展示（评分区间分布、评分与评价人数的相关性图）。
- 实现逻辑：
    1. 平均值计算：
        - 计算所有电影的平均评分与平均评价人数。
    2. 评分区间分布：
        - 将评分划分为四个区间 `[0-7, 7-8, 8-9, 9-10]`，统计每个区间的电影数量。
    3. 最低评分电影：
        - 找到评分最低的前 5 部电影，展示其评分与评价人数。
    4. 评分与评价人数的相关性：
        - 使用相关系数分析两者之间的线性关系，并绘制散点图。
- 关键代码：

```
avg_rating = df["rating"].mean()
avg_people_count = df["people_count"].mean()

bins = [0, 7, 8, 9, 10]
df["rating_group"] = pd.cut(df["rating"], bins=bins, labels=labels,
right=False)

correlation = df["rating"].corr(df["people_count"])
```

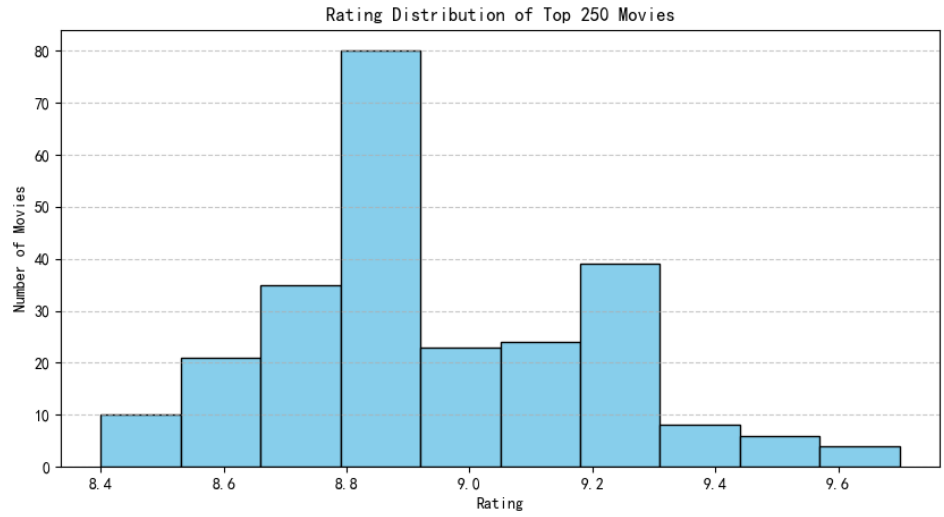
输出

控制台打印

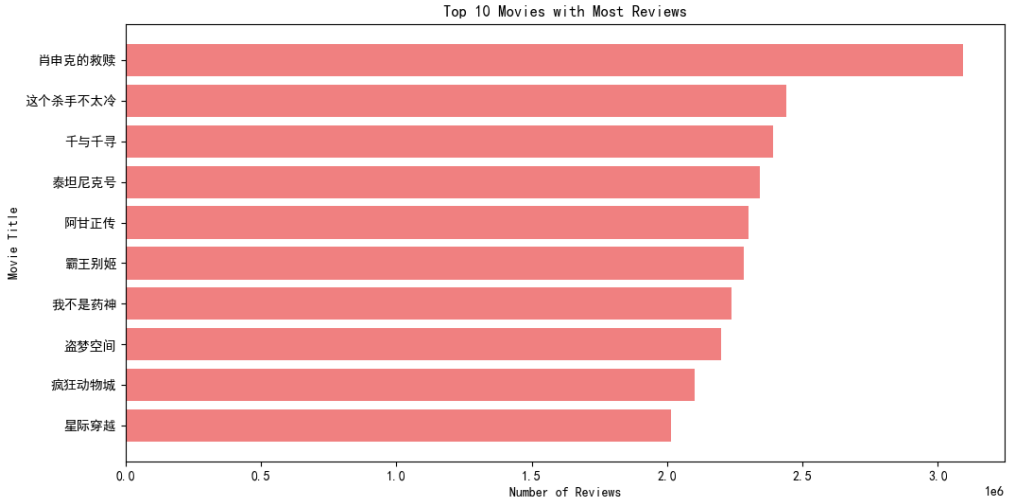
```
Fetching data, please wait...
Page 1 data fetched.
Page 2 data fetched.
Page 3 data fetched.
Page 4 data fetched.
Page 5 data fetched.
Page 6 data fetched.
Page 7 data fetched.
Page 8 data fetched.
Page 9 data fetched.
Page 10 data fetched.
Data fetching completed!
Starting visualization and analysis...
Average rating: 8.94
Average number of reviews: 809795

Rating interval distribution:
rating_group
0-7       0
7-8       0
8-9     146
9-10    104
Name: count, dtype: int64

Lowest Rated 5 Movies:
     title  rating  people_count
211   爱乐之城     8.4       1023484
161  你的名字。      8.5       1515784
192   恐怖游轮     8.5        934919
203    源代码     8.5        884387
212   真爱至上     8.5        773398

Correlation between rating and number of reviews: 0.32
```

图表

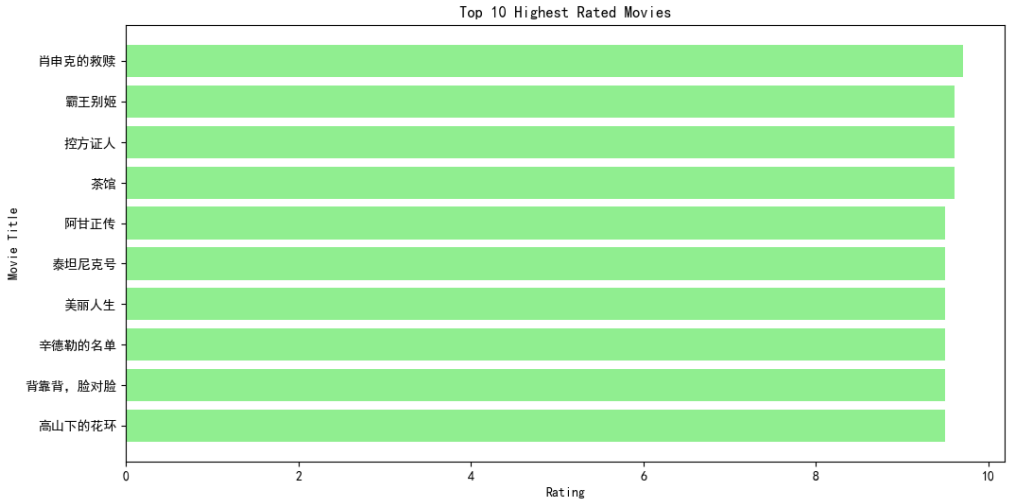
1. **评分分布直方图**:



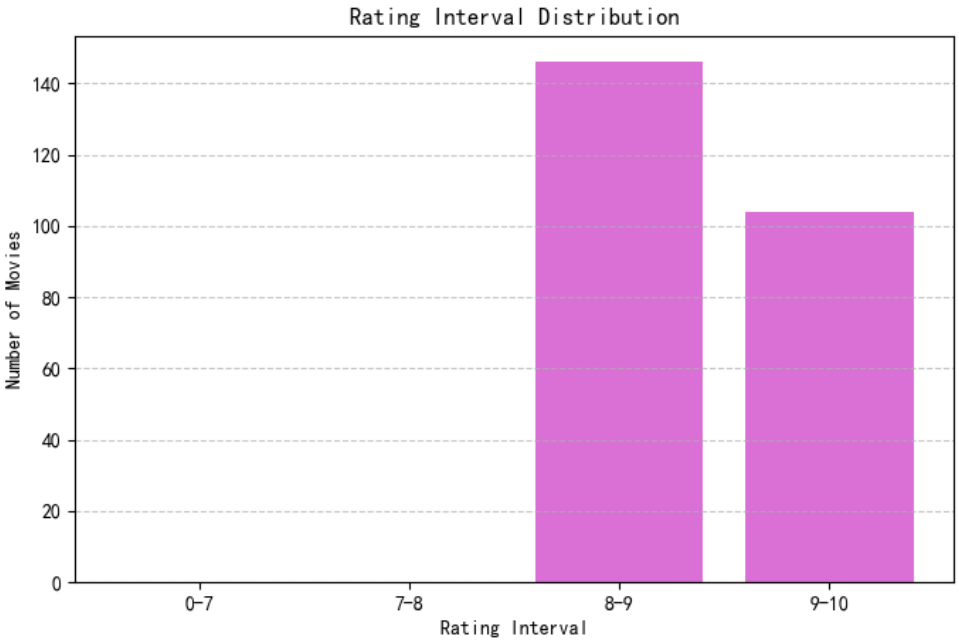
2. **评价人数 Top 10 水平柱状图**:



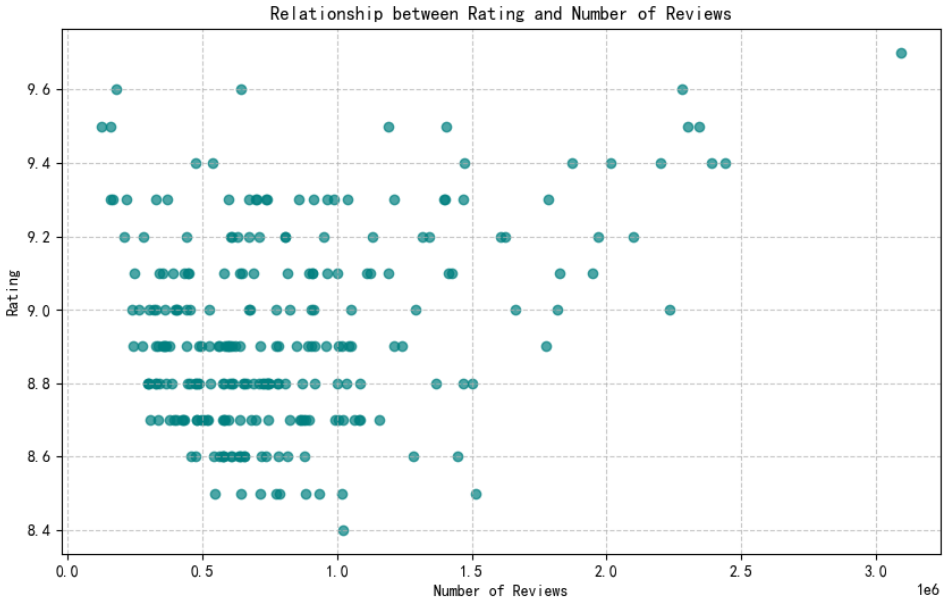
3. **评分最高 Top 10 水平柱状图**:



4. **评分区间分布柱状图**:



5. **相关性散点图**:



环境需求

- **Python 版本**：3.7 及以上。
- 依赖库：
  - `requests`：用于 HTTP 请求。
  - `BeautifulSoup`（`bs4`）：用于 HTML 解析。
  - `pandas`：用于数据处理。
  - `matplotlib`：用于数据可视化。

注意事项

1. 爬虫：
    - 爬取时设置 `User-Agent` 和 `time.sleep` 避免被服务器封禁。
2. 网络连接：
    - 确保网络连接畅通，能够访问豆瓣电影页面。
3. 字体设置：
    - 如出现中文显示问题，需确保 Matplotlib 字体支持中文，推荐使用 `SimHei` 或其他支持中文的字体。