

**1. 目标（第一阶段）**

* 每人从不同角度独立对同一份数据做分析。
* 每个人输出一份组件得分表（component scores），格式统一。
* 第二阶段由合并脚本把所有组件得分统一归一化并按权重合成最终推荐分数，输出 Top-N 推荐列表。
* 原始数据只读：data/raw/movies.xlsx（不能改或重命名）。  
  若做预处理，请把结果放在 data/processed/。

**2. 四人分工（每人输出一个组件分数 CSV）**

（每人只需负责输出 movie\_id,title,component\_score 三列）

**01 — Content（内容与相似度）**

* 负责列：genres, keywords, overview, title
* 目标输出：reports/tables/01\_content\_scores.csv（列：movie\_id,title,component\_score）
* 思路举例：TF-IDF 或关键词权重、genre 相关性、overview 信息量等。

**02 — Ratings & Popularity（评分与受欢迎度）**

* 负责列：vote\_average, vote\_count, popularity
* 目标输出：reports/tables/02\_rating\_scores.csv
* 思路举例：贝叶斯平均/加权投票数、popularity 合成得分。

**03 — Business（票房/出品/制作方）**

* 负责列：budget, revenue, production\_companies
* 目标输出：reports/tables/03\_business\_scores.csv
* 思路举例：ROI、票房规模、知名制片公司加权等。

**04 — Time & Locale（时间/语言/地域/时长）**

* 负责列：release\_date, runtime, spoken\_languages, original\_language
* 目标输出：reports/tables/04\_time\_scores.csv
* 思路举例：新片优先（时间衰减）、语言是否匹配目标用户、时长适配度。

每人内部实现细节可以自由发挥（可以生成一些图表，第一次弄不需要太复杂），  
但输出格式必须完全一致（见下面第 3 条）。

**3. 输出与评分规范（必须严格遵守）**

**文件与位置**

* 每人输出文件命名（必须）：
  + reports/tables/01\_content\_scores.csv
  + reports/tables/02\_rating\_scores.csv
  + reports/tables/03\_business\_scores.csv
  + reports/tables/04\_time\_scores.csv

**文件内容要求**  
每个文件必须包含且仅需这三列（顺序最好相同）：

movie\_id,title,component\_score

* movie\_id：与原数据中唯一标识一致（字符串或数字均可）
* title：电影名（便于人工核对）
* component\_score：分数（数值）

**统一分数尺度与校验（关键）**

* 输出分数可先按你自己的算法产生原始分数。
* **建议最高分 100，最低分 0**（平均值大约在 50 左右为宜）。
* 第二阶段合并时会自动统一归一化到相同尺度（防止不同人分数区间不一致）。

**4. 每人必须记录的内容（务必执行）**

每个人都要在自己编号对应的脚本文件旁边创建一个简单的记录文件，写明：

**文件位置**

reports/logs/01\_content\_log.txt （每个人按编号命名）

**内容模板（复制填好即可）**

# 01\_content\_log.txt

负责成员：XXX

分析主题：内容相似度（Content）

主要使用的字段：genres, keywords, overview, title

主要思路：

- 对关键词进行分词统计，计算出现频率

- 对 genre 做独热编码统计，算出 genre 热度得分

- 对 overview 计算字数和平均词长作为信息量指标

评分范围：0–100（中位数≈50）

代码文件：scripts/01\_content\_scores.py

运行方式：

python scripts/01\_content\_scores.py

输出文件：

reports/tables/01\_content\_scores.csv

每个人都要记录这份说明文件。  
这样在第二阶段合并时可以追踪来源、解释算法。