**高等演算法作業二(數據收集與預處理)**

請收集電影評分資料，作資料清理，將結果轉為excel 格式(含編號，欄位名稱)，完成如下面示例等之資料檔案。

繳交日期：10/4前

資料參考：<https://grouplens.org/datasets/movielens/>

<https://www.kaggle.com/datasets/grouplens/movielens-20m-dataset>

<https://tw.alphacamp.co/blog/data-processing-and-data-cleaning>

負責資料集：

|  |  |  |
| --- | --- | --- |
| 姓名 | 資料集 | Paper |
| 家銓 | MovieLens 1M | 5.Doan |
| 承翰 | MovieLens 100K | 3. Forouzandeh |
| 泓緯 | Social media posts | 4. Moreno |
| 家平 | MovieLens 10M | 1. Salur |
| 孟倢 | Liver cancer imaging | 12. Gletsos |
| 書源 | MovieLens 10M | 7.Reece |
| 亦成 | MovieLens 20M | 11. Subramaniam |
| 明樺 | MovieLens 25M | 6. Cocos |

繳交內容：資料檔案(各excel檔，檔案若太大可用雲端連結)及書面報告(取該檔案前20筆作表格並說明之)

資料檔案示例：

1. 電影資料(movie data)

| movieId | title | genres |
| --- | --- | --- |
| 1 | Toy Story (1995) | Animation|Children's|Comedy | |
| 2 | Jumanji (1995) | Adventure|Children's|Fantasy | |
| 3 | Grumpier Old Men (1995) | Comedy|Romance | |
| 4 | Waiting to Exhale (1995) | Comedy|Drama | |
| 5 | Father of the Bride Part II (1995) | Comedy | |

2. 使用者評分資料資料(grading data)

| userId | movieId | rating | timestamp |
| --- | --- | --- | --- |
| 1 | 1193 | 5 | 978300760 |
| 1 | 661 | 3 | 978302109 |
| 1 | 914 | 3 | 978301968 |
| 1 | 3408 | 4 | 978300275 |
| 1 | 2355 | 5 | 978824291 |

3. 使用者資料(user data)

| userId | gender | age | occupation |  | zip-code |
| --- | --- | --- | --- | --- | --- |
| 1 | F | 1 | 10 |  | 48067 |
| 2 | M | 56 | 16 |  | 70072 |
| 3 | M | 25 | 15 |  | 55117 |
| 4 | M | 45 | 7 |  | 2460 |
| 5 | M | 25 | 20 |  | 55455 |

4. 電影對使用者評分矩陣

| userId/  movieId | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | ... |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | 5.0 | 0.0 | 0.0 | 0.0 | 0.0 | 4.0 | 0.0 | 4.0 | 5.0 | 5.0 | ... |
| 2 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 5.0 | ... |
| 3 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... |
| 4 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 3.0 | 0.0 | 0.0 | ... |