**BỘ MÔN HỆ THỐNG THÔNG TIN – KHOA CÔNG NGHỆ THÔNG TIN**

**ĐẠI HỌC KHOA HỌC TỰ NHIÊN THÀNH PHỐ HỒ CHÍ MINH, ĐẠI HỌC QUỐC GIA TP HCM**

Sinh viên thực hiện: Nhóm 2122CDCL\_04

GV phụ trách: Hồ Thị Hoàng Vy

Đồ án 2  - CHUYÊN ĐỀ CHỌN LỌC TRONG HỆ THỐNG THÔNG TIN

HỌC KỲ II – NĂM HỌC 2021-2022

**môn chuyên đề chọn lọc trong hệ thống thông tin**

**BẢNG THÔNG TIN CHI TIẾT NHÓM**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Mã nhóm:** | 2122CDCL\_04 | | | |
| **Tên nhóm:** | 2122CDCL\_04 | | | |
| **Số lượng:** | **4** | | | |
| **MSSV** | **Họ tên** | **Email** | **Điện thoại** | **Hình ảnh** |
| 1712486 | Võ Quốc Hưng | 1712486@student.hcmus.edu.vn |  |  |
| 18120418 | Phạm Minh Khoa | 18120418@student.hcmus.edu.vn |  |  |
| 1712850 | Trần Trung | 1712850@student.hcmus.edu.vn |  |  |
| 1712553 | Bùi Tấn Lân | 1712553@student.hcmus.edu.vn |  |  |

**BẢNG PHÂN CÔNG & ĐÁNH GIÁ HOÀN THÀNH CÔNG VIỆC**

|  |  |  |  |
| --- | --- | --- | --- |
| **Bảng phân công & đánh giá hoàn thành công việc** | | | |
| **Công việc thực hiện** | **Người thực hiện** | **Mức độ hoàn thành** | **Đánh giá của nhóm** |
| Lên ý tưởng thuật toán, code, đánh giá, nhận xét | Võ Quốc Hưng | 70% | 10/10 |
| Lên ý tưởng thuật toán, code, lọc dữ liệu, tìm hiểu thông tin dữ liệu | Phạm Minh Khoa | 100% | 10/10 |
| Lên ý tưởng thuật toán, code, đánh giá, nhận xét | Trần Trung | 70% | 10/10 |
| Lên ý tưởng thuật toán, code, thống kê | Bùi Tấn Lân | 100% | 10/10 |

Mục Lục

[**A.** **YÊU CẦU CỦA ĐỒ ÁN** 2](#_Toc102242150)

[1. Yêu cầu đồ án 2](#_Toc102242151)

[2. Yêu cầu nộp 2](#_Toc102242152)

[**B.** **KẾT QUẢ** 3](#_Toc102242153)

[1. Giải thích dataset, ý nghĩa thuộc tính 3](#_Toc102242154)

[2. Ý tưởng thực hiện thuật toán (mã giả) 7](#_Toc102242155)

[2.1. Tạo ma trận tương đồng item-item (giai đoạn offline) 7](#_Toc102242156)

[2.2. Dự đoán rating (giai đoạn online) 8](#_Toc102242157)

[3. Kết quả thực hiện 10](#_Toc102242158)

[4. Đánh giá 19](#_Toc102242159)

[5. Link tham khảo 19](#_Toc102242160)

# **YÊU CẦU CỦA ĐỒ ÁN**

1. Yêu cầu đồ án
2. Dataset: <https://grouplens.org/datasets/movielens/100k/>
3. Hãy cài đặt hệ tư vấn phim theo kỹ thuật items to items Collaborative Filtering  
   (memory based) sử dụng ngôn ngữ python.

* Mô tả dataset cách chi tiết.
* Mô tả bài toán item to item CF với movilen dataset (mô tả ý tưởng thực hiện)
* Đọc và chuyển dữ liệu vào ma trận
* Quan sát mô tả dữ liệu
* Biến đổi, xử lý dữ liệu
* Giai đoạn offline:
  + tính toán độ tương đồng sử dụng KNN
  + tính toán độ tương đồng sử dụng Jaccar
* Giai đoạn online:
  + Phát sinh các gợi ý
* Đánh giá thuật toán:
  + Lựa chọn độ đo đánh giá độ chính xác của dự đoán

1. Yêu cầu nộp

* .doc: báo cáo giải thích về dataset, ý nghĩa thuộc tính, ý tưởng thực hiện thuật  
  toán (mã giả), kết quả thực hiện, đánh giá…
* Video: quay lại quá trình trao đổi và thực hiện của mỗi thành viên có mở camera
* Source code: .pybn và link google colab
* Làm bài trên google colab
* Lần 1:
  + Nộp source + báo cáo
  + Source: .iynb (có hiển thị kết quả thực thi của mỗi đoạn lệnh)
  + File data trước và sau khi xử lý làm sạch dữ liệu
  + Báo cáo: giải thích, mô tả dataset, các bước làm sạch, biến đổi dữ liệu, …
  + Tính toán độ tương đồng sử dụng Jaccard
* Lần 2: nộp hoàn chỉnh
  + Source + báo cáo
  + Source: gồm đầy đủ các bước cả giai đoạn offline + online, file dữ liệu
  + Báo cáo: phân tích, cài đặt, đánh giá
  + Video giải thích bài làm.

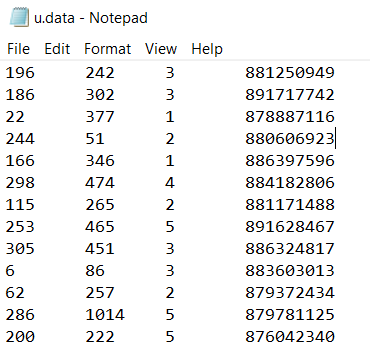
# **Kết quả**

## Giải thích dataset, ý nghĩa thuộc tính

Sử dụng thư viện pandas và file README để tìm hiểu, kiểm tra, lọc data.

Sử dụng hàm read\_csv() để đọc dữ liệu.

* **ml-100k.zip:** chứa tập những file dữ liệu (u.data, u1.base, u1.test…), file chứa script phân data và file thông tin mô tả.
* **u.data:**

****

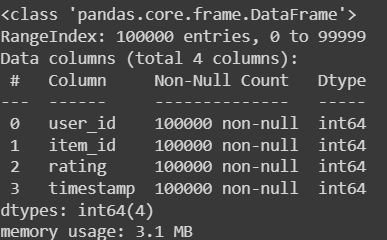
* + Theo mô tả từ file README, file gồm có 100.000 rating (đánh giá từ 1  5) của 943 người dùng về 1682 video. Mỗi user đã đánh giá ít nhất 20 phim. Dữ liệu được sắp xếp ngẫu nhiên và mỗi dòng tuân theo cấu trúc sau: **user id | item id | rating | timestamp**. Các cột cách nhau bởi dấu tab.

***Kiểm tra dữ liệu:***

* + File chỉ chứa dữ liệu, dòng đầu tiên không phải tên của các cột nên sử dụng hàm read\_csv() với tham số names là mảng các tên cột để khỏi nhầm dòng đầu tên là tên các cột. Khi có đối số names, thì tham số chuyển từ mặc định header=0 sang header=non, cho phép lấy mảng names làm tên các cột.
  + Lệnh đọc data từ file u.data trên google drive, sử dụng google colab để thực thi.

pd.read\_csv("/content/drive/MyDrive/Colab Notebooks/moviesLenData/u.data", sep='\t', names=['user\_id','item\_id','rating','timestamp'])

* + Dùng hàm info() để kiểm tra dữ liệu, kết quả:

****

* + - User id, Item id, Rating, timestamp đều là số nguyên.
  + Các **timestamp** là Unix giây kể từ 1/1/1970 UTC, lúc 1/1/1970 là 0 và thời gian sau đó cộng thêm số giây tăng thêm.
  + Kiểm tra giá trị của rating: max=5, min=1

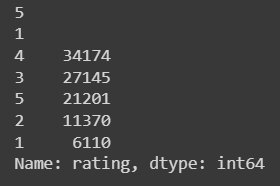
Sử dụng các hàm:

print(df.rating.max())

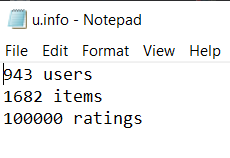
print(df.rating.min())

print(df.rating.value\_counts())

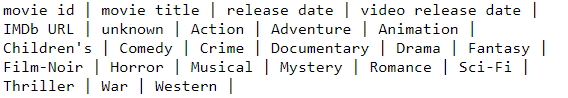
**Kết quả:**

****

* + Tưng tự cho số user và video đều khớp với mô tả trong file README.
  + **Vậy không có dữ liệu sai, null, thiếu sót trong file u.data => không cần lọc dữ liệu.**
* **u.info:** gồm số lượng user, phim, rating trong file **u.data**

****

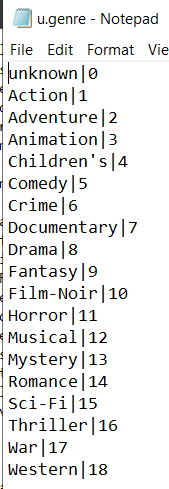
* **u.item:** thông tin về các phim, 19 cột sau cùng là thể loại (từ unknown trở về sau). Giá trị 1 là phim thuộc thể loại, ngược lại, không thuộc là 0. Một phim có thể thuộc nhiều thể loại. movie id dược sử dụng trong file u.data. Cấu trúc như sau:

****

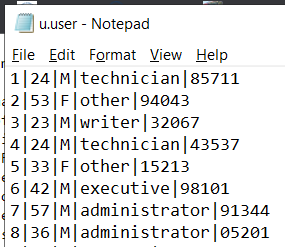
Và dữ liệu được lưu trong file như sau, ngăn cách bằng dấu gạch đứng |, trước cột URL có hai gạch đứng ||:



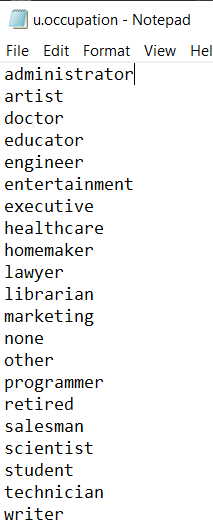
* + Lệnh đọc file u.item: với đối số đầu tiên là đường dẫn tới file u.item, đối số thứ 2 là ký tự ngăn cách (sep), header=None chỉ file dữ liệu không có tên cột cho dữ liệu, tránh việc lấy dòng đầu làm tên cột, names là danh sách tên cột đưa vào, encoding để chỉ loại dữ liệu ký tự bên trong file u.item.
* movieItem = pd.read\_csv('/content/drive/MyDrive/Colab Notebooks/moviesLenData/u.item', sep='|',header=None,names=colItem,encoding='latin-1')
* **u.genre:** danh sách các thể loại, gồm có 19 thể loại. Cột 1 là tên, cột 2 là id, 2 cột được ngăn cách bởi dấu gạch đứng |.

****

* **u.user:** Thông tin nhân khẩu về những user, mỗi cột được ngăn cách bởi dấu gạch đứng | và có cấu trúc như sau: user id | age | gender | occupation | zip code. User id được sử dụng trong u.data.

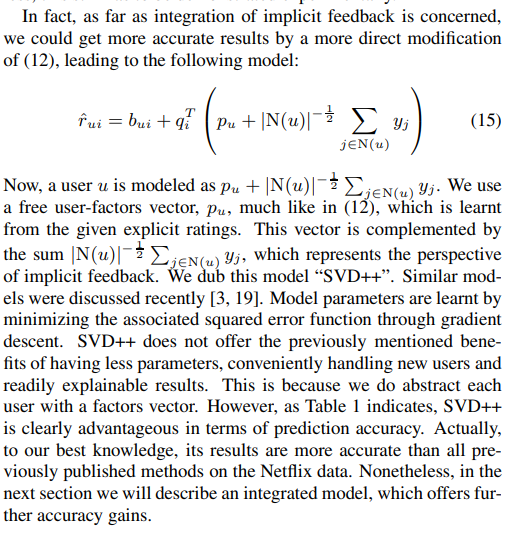
****

* **u.occupation:** danh cách các nghề nghiệp, gồm 21 nghề.

****

* **u1.base  u5.base và u1.test  u5.test:** .base là dữ liệu train, .test là dữ liệu test. Dữ liệu được lấy ra từ u.data chia 80% cho u1.base và 20% cho u1.test. u1  u5 được tạo khi chạy file thực thi mku.sh.
* **ua.base, ua.test, ub.base, ub.test:** .base là tập dữ liệu train, .test là tập dữ liệu test với mỗi user lấy 10 đánh giá của user (tức lấy 10 phim được đánh giá bởi mỗi người dùng). File ua.test bị lỗi mở bằng notepad, lên toàn tiếng trung, nhưng mở bằng visual code thì được. Dữ liệu được lấy từ file u.data khi chạy file mku.sh.
* **allbut.pl:**  script tạo ra các tập training và test trong đó tất cả trừ n rating người dùng nằm trong dữ liệu training.

## Ý tưởng thực hiện thuật toán (mã giả)



*ảnh tử bài báo của Koren*

The other one, denoted by N(u), contains all items for which u provided an implicit preference.

Yj is the (implicit) item factors : numpy array of size (n\_items, n\_factors) tương đương với qi nhưng radom giá trị khác nhau.

bias cua user u

bias cua item i

trung bình rating

**[***m x s***]:** ma trận nhân tố ẩn biểu diễn người dùng (user-factor matrix)

**[***n x s***]:** ma trận nhân tố ẩn biểu diễn sản phẩm (item-factor matrix)

* Model: *u=1…m; i=1…n;*  **[***m x s***]; [***n x s***];**

Model estimation:

Tikhonov regularization: regularization weight

Tối ưu bằng stochastic gradient descent:

(Từ công thức sau chúng ta sẽ tách tường phần để tích ở giai đoàn offline.)

Xét hàm mục tiêu tại 1 điểm dữ liệu:

Lấy đạo hàm theo biến (dưới từ min trong công thức):

### Giai đoạn offline

α có giá trị mặc định là 0.005, λ là 0.02, số factor mặc định là 20.

Step 1: Initialize randomly , **,** in (-1,1).

Step 2: { | }

Step 3:   
Step 4: Repeat until convergence (sẽ cho 1 số mặc định là 20 “epoches”):

Step 5: Randomly shuffle the elements in 𝕐.

Step 6: For each pair in :

Step 7:

Step 8:

Step 9:

Step 10:

Step 11:

Step 12:

Step 13:

### Giai đoạn online

<https://github.com/NicolasHug/Surprise/blob/master/surprise/prediction_algorithms/matrix_factorization.pyx#:~:text=%22%20processing%20epoch%20%7B%7D%22.format(current_epoch))-,for,-u%2C%20i%2C%20r%20in%20trainset.all_ratings()%3A>

<https://surprise.readthedocs.io/en/stable/matrix_factorization.html#surprise.prediction_algorithms.matrix_factorization.SVDpp>

<https://github.com/Murtazali05/matrix-factorization/blob/master/svdpp.py>

## Kết quả thực hiện

# **Lấy data từ website grouplens và giải nén, không cần kết nối drive**

**- Giải nén file**

[ ]

DATASET\_LINK='http://files.grouplens.org/datasets/movielens/ml-100k.zip'  
!wget -nc http://files.grouplens.org/datasets/movielens/ml-100k.zip  
!unzip -n ml-100k.zip

--2022-04-30 13:31:09-- <http://files.grouplens.org/datasets/movielens/ml-100k.zip>

Resolving files.grouplens.org (files.grouplens.org)... 128.101.65.152

Connecting to files.grouplens.org (files.grouplens.org)|128.101.65.152|:80... connected.

HTTP request sent, awaiting response... 200 OK

Length: 4924029 (4.7M) [application/zip]

Saving to: ‘ml-100k.zip’

ml-100k.zip 100%[===================>] 4.70M 28.5MB/s in 0.2s

2022-04-30 13:31:10 (28.5 MB/s) - ‘ml-100k.zip’ saved [4924029/4924029]

Archive: ml-100k.zip

creating: ml-100k/

inflating: ml-100k/allbut.pl

inflating: ml-100k/mku.sh

inflating: ml-100k/README

inflating: ml-100k/u.data

inflating: ml-100k/u.genre

inflating: ml-100k/u.info

inflating: ml-100k/u.item

inflating: ml-100k/u.occupation

inflating: ml-100k/u.user

inflating: ml-100k/u1.base

inflating: ml-100k/u1.test

inflating: ml-100k/u2.base

inflating: ml-100k/u2.test

inflating: ml-100k/u3.base

inflating: ml-100k/u3.test

inflating: ml-100k/u4.base

inflating: ml-100k/u4.test

inflating: ml-100k/u5.base

inflating: ml-100k/u5.test

inflating: ml-100k/ua.base

inflating: ml-100k/ua.test

inflating: ml-100k/ub.base

inflating: ml-100k/ub.test

[ ]

#adding library  
import numpy as np  
  
import pandas as pd

**- Xem file u.data**

[ ]

column\_names1 = ['user id','movie id','rating','timestamp']  
dataset = pd.read\_csv('ml-100k/u.data', sep='\t',header=None,names=column\_names1)  
dataset.head()

# **Check data infomation, datatype, file formation, seperate character**

[ ]

#adding library  
import numpy as np  
  
import pandas as pd

[ ]

#add data from drive  
#from google.colab import drive  
#drive.mount('/content/drive')

Drive already mounted at /content/drive; to attempt to forcibly remount, call drive.mount("/content/drive", force\_remount=True).

[ ]

#get data into data frame, then check data (u.data)  
df  = pd.read\_csv("ml-100k/u.data", sep='\t', names=['user\_id','item\_id','rating','timestamp'])  
df.info()

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 100000 entries, 0 to 99999

Data columns (total 4 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 user\_id 100000 non-null int64

1 item\_id 100000 non-null int64

2 rating 100000 non-null int64

3 timestamp 100000 non-null int64

dtypes: int64(4)

memory usage: 3.1 MB

[ ]

#check min, max rating and count how many the distincts value  
print("rating:")  
print(df.rating.max())  
print(df.rating.min())  
print(df.rating.value\_counts())  
  
#check how many the distincts value of users id  
print("user:")  
print(df.user\_id.value\_counts())  
#check how many the distincts value of video  
print("video:")  
print(df.item\_id.value\_counts())

rating:

5

1

4 34174

3 27145

5 21201

2 11370

1 6110

Name: rating, dtype: int64

user:

405 737

655 685

13 636

450 540

276 518

...

441 20

36 20

812 20

895 20

93 20

Name: user\_id, Length: 943, dtype: int64

video:

50 583

258 509

100 508

181 507

294 485

...

852 1

1505 1

1653 1

1452 1

1641 1

Name: item\_id, Length: 1682, dtype: int64

[ ]

df['rating'].describe()

count 100000.000000

mean 3.529860

std 1.125674

min 1.000000

25% 3.000000

50% 4.000000

75% 4.000000

max 5.000000

Name: rating, dtype: float64

[ ]

[ ]

plt.hist(df['rating'],log=True);

plt.plot();

plt.show();

## Đánh giá

**Đánh giá sơ bộ**

* Những phim ít đánh giá sẽ cho kết quả dự đoán rating thấp
* Nhưng phim không có đánh giá nào (phim mới vào hệ thống) sẽ không tính toán được.

## Link tham khảo

1. Chuyển dataframe thành item matrix: <https://stackoverflow.com/questions/37576594/rearrange-a-pandas-data-frame-to-create-a-2d-ratings-matrix>
2. Sparse matrix (sparse dataframe) to dense dataframe, or compressed sparse row matrix: <https://discuss.dizzycoding.com/pandas-sparse-dataframe-to-sparse-matrix-without-generating-a-dense-matrix-in-memory/>

<https://docs.scipy.org/doc/scipy/reference/generated/scipy.sparse.csr_matrix.html>

1. Tham khảo tham số svd++ trang surprise: <https://surprise.readthedocs.io/en/stable/matrix_factorization.html#surprise.prediction_algorithms.matrix_factorization.SVDpp>
2. Tham khảo code <https://github.com/NicolasHug/Surprise/blob/master/surprise/prediction_algorithms/matrix_factorization.pyx>
3. Tài liệu pdf của Yehuda Koren, nộp chung với đồ án.
4. Link google colab: <https://colab.research.google.com/drive/16UqHiFQB1BHt3kJdWZns7BUBFVfk6O47?usp=sharing>