RECOMMENDATION WITH BOOK-CROSSING DATASET

TRẦN QUỐC VIỆT

Table of content

I . Data Overview:

Tập dữ liệu này có tất cả 4 file dữ liệu:

- **users_info.dat:** File này chứa dữ liệu về thông tin user bao gồm 3 trường giá trị:
  - ***User-ID:*** ID của USER.
  - ***Location*:** Thông tin về Location của USER đó.
  - ***Age:*** Thông tin về độ tuổi của USER đó.

	User-ID	Location	Age
0	1	minneapolis, minnesota, usa	24
1	2	san diego, california, usa	20
2	3	novinger, missouri, usa	16
3	4	sonoma, california, usa	34
4	5	berkeley, california, usa	23
5	6	king of prussia, ,	36
6	7	berkeley, ,	22
7	8	rennes, bretagne, france	22
8	9	st. louis, missouri, usa	36
9	10	minneapolis, minnesota, usa	26

Fig 1. *Thông tin 10 dòng dữ liệu đầu tiên của user_info*

→ Bộ dữ liệu thông tin user này có tất cả: **2946** users. **(2946 hàng, 3 cột)**

- **items_info.dat:** File này chứa dữ liệu về thông tin của item (mà cụ thể ở đây là book (cuốn sách)), gồm 9 trường:
  - ***Book-ID:*** ID của cuốn sách.
  - ***ISBN:*** ISBN tương ứng với Book-ID
  - ***Book-Title***: Tiêu đề của cuốn sách.
  - ***Book-Author:*** Tác giả của cuốn sách.
  - ***Year-Of-Publication:*** Năm xuất bản.
  - ***Publisher:*** Nhà xuất bản.
  - ***Image-URL-S:*** Đường dẫn đến ảnh của cuốn sách.
  - ***Image-URL-M:*** Đường dẫn đến ảnh của cuốn sách.

- ***Image-URL-L:*** Đường dẫn đến ảnh của cuốn sách.

	Book_ID	ISBN	Book-Title	Book-Author	Year-Of-Publication	Publisher	Image-URL-S	Image-URL-M	Image-URL-L
0	1	0060973129	Decision in Normandy	Carlo D'Este	1991	HarperPerennial	http://images.amazon.com/images/P/0060973129.0...	http://images.amazon.com/images/P/0060973129.0...	http://images.amazon.com/images/P/0060973129.0...
1	2	0393045218	The Mummies of Urumchi	E. J. W. Barber	1999	W. W. Norton &amp Company	http://images.amazon.com/images/P/0393045218.0...	http://images.amazon.com/images/P/0393045218.0...	http://images.amazon.com/images/P/0393045218.0...
2	3	0425176428	What If?: The World's Foremost Military Histor...	Robert Cowley	2000	Berkley Publishing Group	http://images.amazon.com/images/P/0425176428.0...	http://images.amazon.com/images/P/0425176428.0...	http://images.amazon.com/images/P/0425176428.0...
3	4	0452264464	Beloved (Plume Contemporary Fiction)	Toni Morrison	1994	Plume	http://images.amazon.com/images/P/0452264464.0...	http://images.amazon.com/images/P/0452264464.0...	http://images.amazon.com/images/P/0452264464.0...
4	5	0609804618	Our Dumb Century: The Onion Presents 100 Years...	The Onion	1999	Three Rivers Press	http://images.amazon.com/images/P/0609804618.0...	http://images.amazon.com/images/P/0609804618.0...	http://images.amazon.com/images/P/0609804618.0...

Fig 2. Thông tin 5 dòng đầu tiên của item_info

→ Bộ dữ liệu thông tin sách này gồm có: **17384** cuốn sách **(17384 hàng, 9 cột).**

- **book_ratings.dat:** File này chứa thông tin dữ liệu đánh giá của USER đối với BOOK.
  - ***user:*** ID của user thực hiện đánh giá.
  - ***item:*** ID của cuốn sách được đánh giá.
  - ***rating:*** Mức độ đánh giá giá của user đối với cuốn sách (từ 1-10).

	user	item	rating
0	1	6264	7.0
1	1	4350	7.0
2	1	6252	5.0
3	1	202	9.0
4	1	6266	6.0
5	1	4810	5.0
6	1	6251	9.0
7	1	160	9.0
8	1	161	8.0
9	1	631	10.0

Fig 3. *Thông tin 10 dòng dữ liệu đầu tiên của book-rating.*

→ Bộ dữ liệu về thông tin đánh giá này có **62656** lượt đánh giá **(62656 hàng, 3 cột).**

- **book_history.dat:** File này chứa lịch sử truy cập của user với cuốn sách.
  - ***User:*** ID của user truy cập.
  - ***Item:*** ID của cuốn sách được truy cập.
  - ***Accessed:*** default 1

	0	1	2
0	user	item	accessed
1	1	152	1
2	1	153	1
3	1	2176	1
4	1	154	1
5	1	734	1
6	1	6250	1
7	1	457	1
8	1	6264	1
9	1	6222	1

Fig 4. *Thông tin 10 dòng dữ liệu đầu tiên của book_history.*

→ Bộ dữ liệu về lịch dữ truy cập này có **272679** lượt đánh giá **(272679 hàng, 3 cột).**

I. <u>Exploratory Data Analysis:</u>

1. Explore user info:

- Age distribution:

- Thông tin về độ tuổi kéo dài từ 1 - 239 tuổi, độ tuổi phổ biến trong khoảng từ 25 đến 50 tuổi, trong đó độ tuổi có số lượng user nhiều nhất là 29 tuổi, với 124 user đó độ tuổi này.
- Bên cạnh đó vẫn tồn tại các user có độ tuổi nhỏ hơn 5 và lớn hơn 90, nhưng là số ít, tuy nhiện độ tuổi này chúng ta cần xem xét vì mức độ chính xác của dữ liệu, và nếu dữ liệu đúng thì cần xem xét khả nẳng đánh giá sách ở độ tuổi này, vì vậy, trong quá trình xử lý, ta có thể loại bỏ/thay thể chúng bằng giá trị độ tuổi phổ biến nhất.

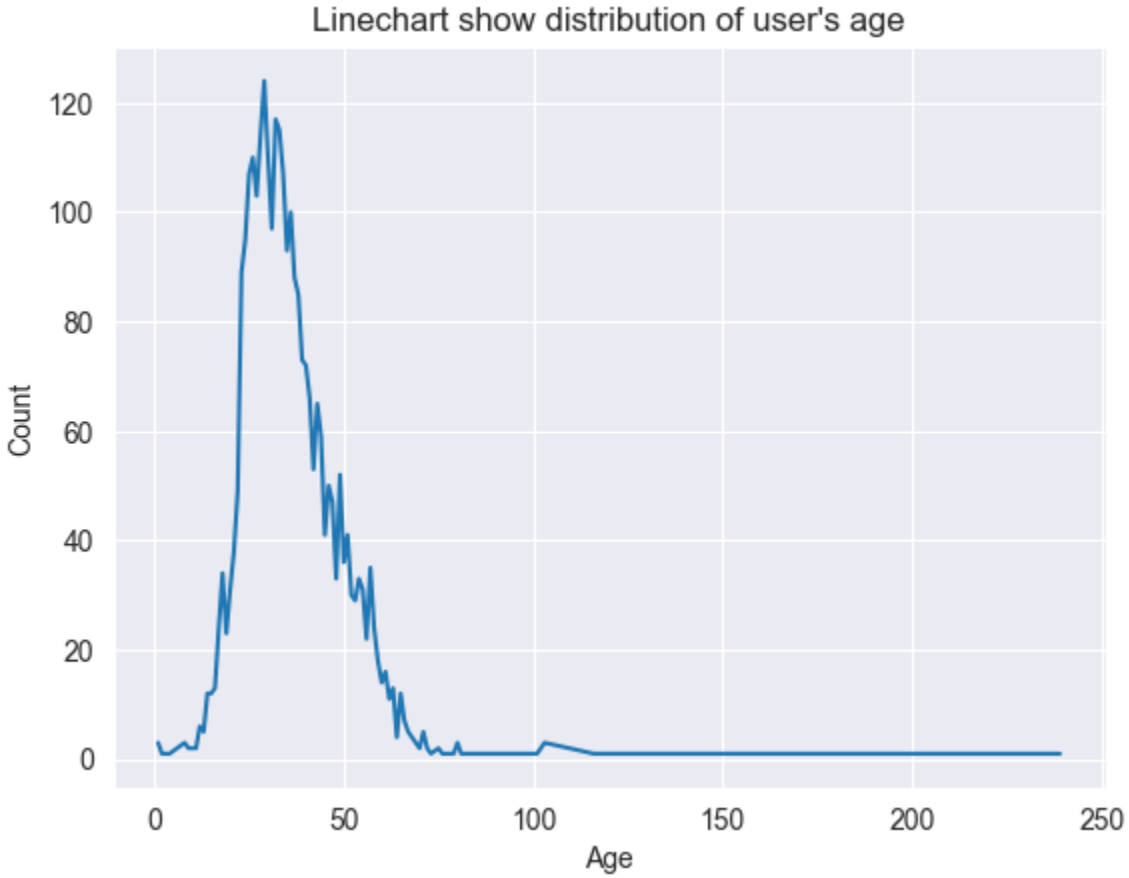

Fig 5. *Biểu đồ về sự phân bố lượng user ở các độ tuổi*

- Explore User's Location:
  - Dựa vào dữ liệu, có tất cả **1799 locations,** từ lượng location này, chúng ta phân tích được **48 countries.**
  - Trong đó, top **5 quốc gia** có lượng user nhiều nhất bao gồm:
    + usa: 2128 users.
    + canada: 284 users.
    + united kingdom: 113 users.
    + australia: 54 users.

+ germany: 49 users.

- Chúng ta có thể xem rõ hơn ở hình bên dưới:

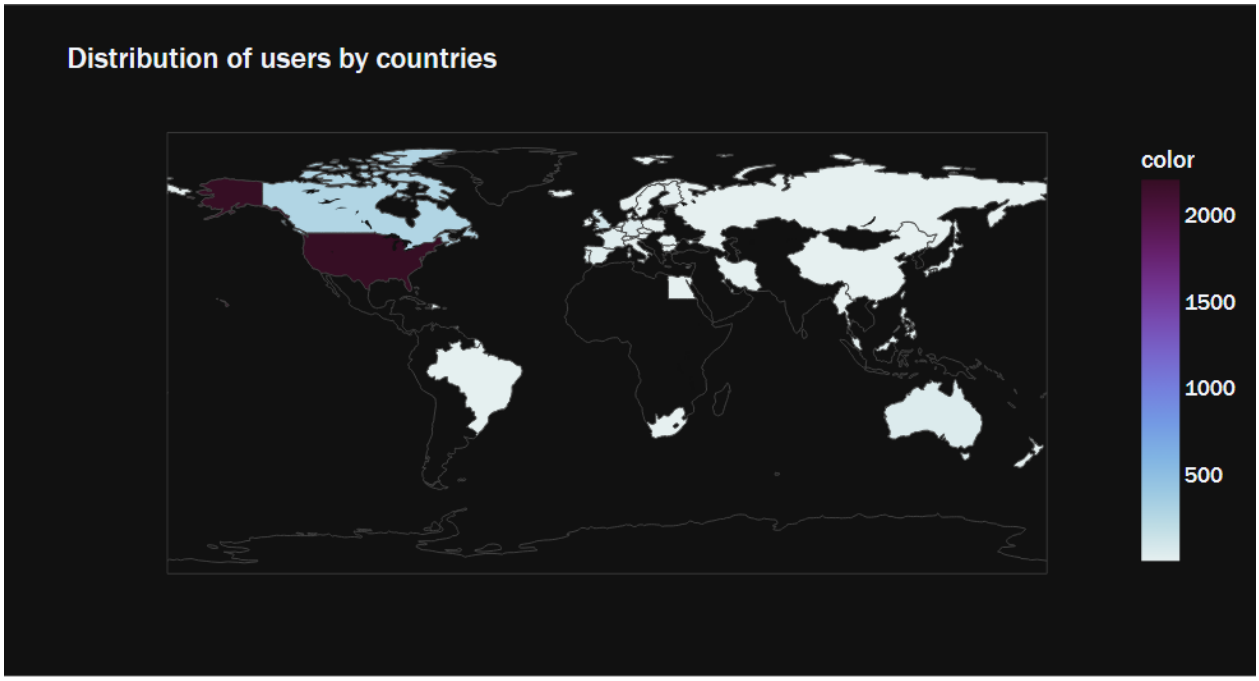

Fig 6. *Biểu đồ về sự phân bố user ở các quốc gia.*

2. Explore item info:

- Đầu tiên, chúng ta sẽ xem xét về những khoảng thời gian mà sách được publish nhiều nhất.

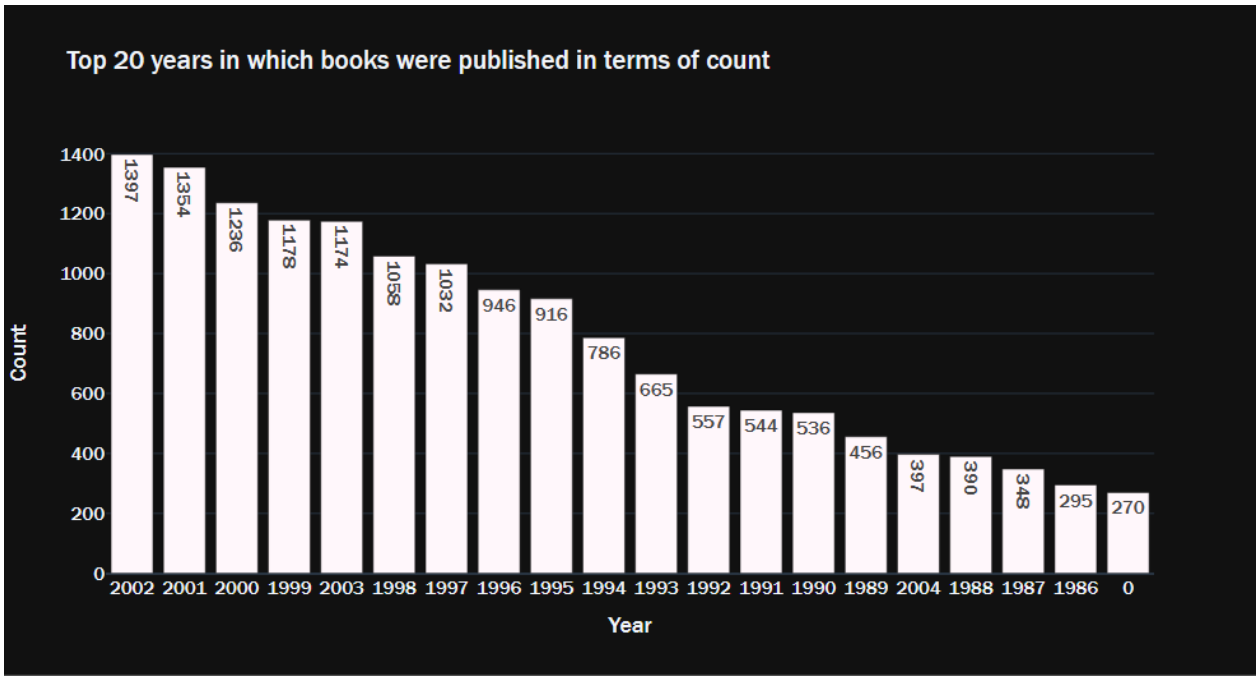

Fig 7. *Biểu đồ thông tin 20 năm mà có số lượng sách được publish nhiều nhất.*

- Năm 0, nguyên nhân có thể là do dữ liệu bị missing, không thu thập được. Còn lại, chúng ta có thể thấy được thông tin 20 năm mà lượng sách được publish nhiều nhất, con số này lớn nhất vào năm 2002 và nhỏ nhất là vào năm 1986.

3. Explore rating data:

Trong phần tiếp theo chúng ta sẽ khám phá về phân phối, số lượng user đánh giá cho mỗi loại rating (1-10).

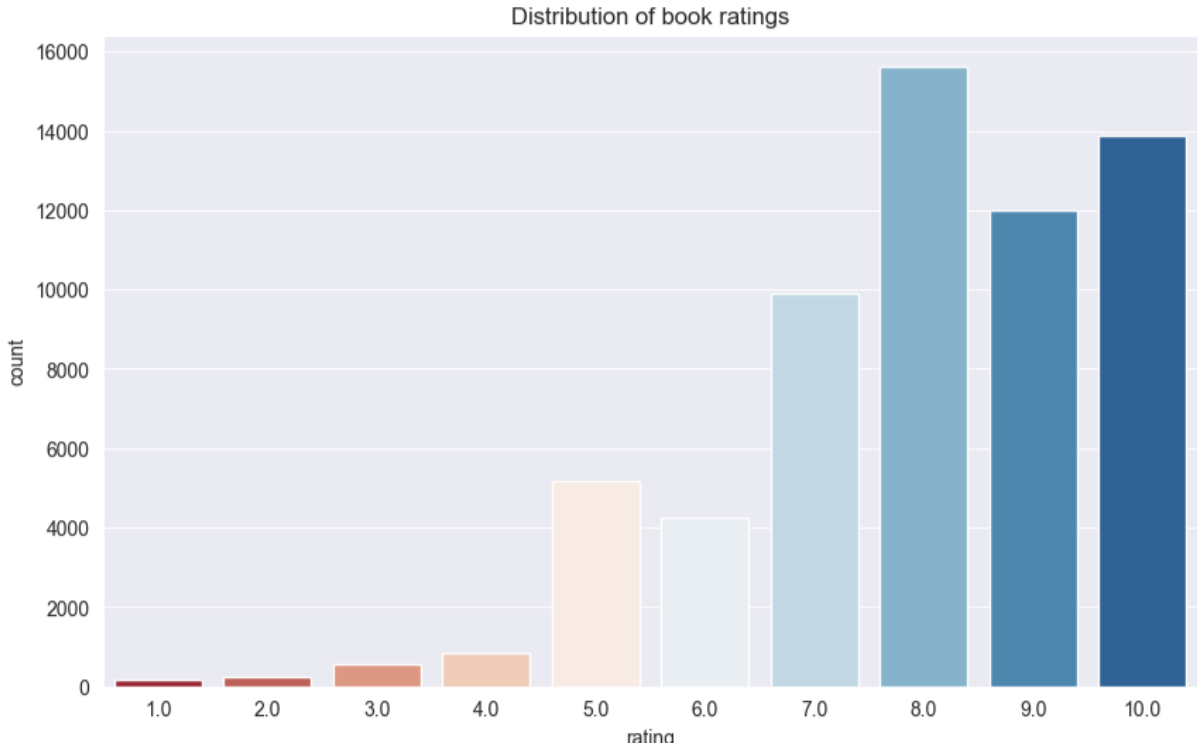

Fig 8. *Biểu đồ về phân phối của từng loại rating được user đánh giá*

- Dựa vào biểu đồ, chúng ta có thể thấy, user thường đưa ra đánh giá khi họ có cảm nhận tốt với cuốn sách (tức là giá trị rating từ 5 trở lên). Con số này hoàn toàn lớn hơn nhất nhiều so với số lượng đánh giá không tốt (tức là từ 5 sao trở xuống).

- Tiếp theo, chúng ta sẽ thử vẽ một biểu đồ, để thể hiện số lượt đánh giá và trung bình giá trị đánh giá của top20 cuốn sách được đánh giá nhiều nhất.

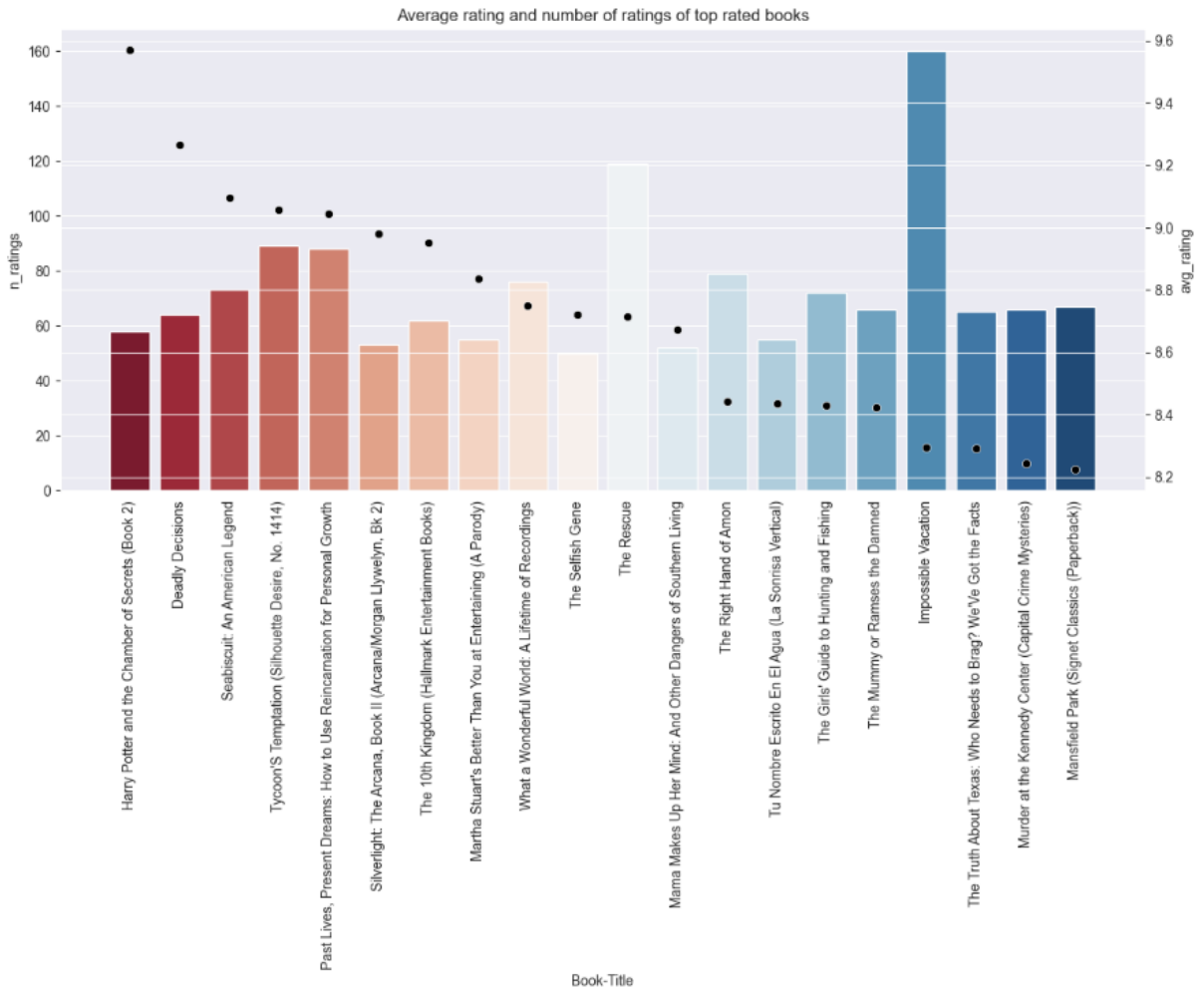

Fig 8. *Biểu đồ mô tả quan hệ giữa số lượt đánh giá và trung bình đánh giá của top 20 cuốn sách nhận được nhiều sự đánh giá nhất*

- Biểu đồ này cho thấy, Một số cuốn sách có giá số lượng đánh giá ít, nhưng trung bình đánh giá lại cao (9.6), tuy nhiên, có một số trường hợp ngược lại, ví dụ như cuốn sách có title là "Impossible Vacation", có số lượng đánh giá rất cao, nhưng trung bình đánh giá lại cực kì thấp, từ đó có thể thấy, cảm nhận, đánh giá không tốt của người dùng đối với cuốn sách này.

- Tiếp theo, chúng ta xem xét về những tác giá nhận được nhiều sự đánh giá nhất. Những tác giải này bao gồm: 'Spalding Gray', 'Nicholas Sparks', 'Katherine Garbera', 'Denise Linn', 'Anne Rice', 'Lauren Haney', 'Wally Lamb', 'Bob Thiele', 'Joe Hutsko', 'LAURA HILLENBRAND', 'J. R. R. Tolkien', 'Melissa Bank', 'Robin McKinley', 'Jane Austen', 'Anne Rice', 'Judith Kelman', 'Margaret Truman', 'Anne Dingus', 'Kathy Reichs', 'Kathryn Wesley'.

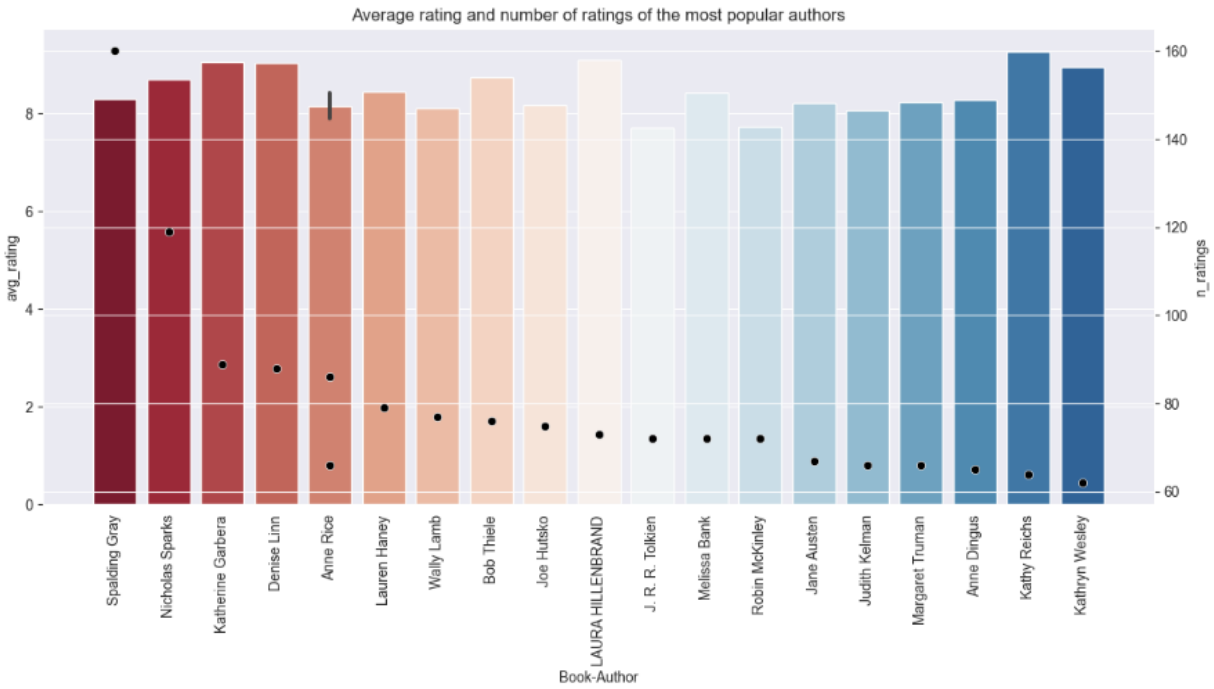

Fig 9. *Biểu đồ mô tả quan hệ giữa số lượt đánh giá và trung bình đánh giá của top 20 tác giả nhận được nhiều sự đánh giá nhất*

- Như vậy, hầu hết các tác giả phổ biết này, đều có kết quả đánh giá khả quan, lớn hơn 7, và đa phần trên 8. Đối với hệ thống gợi ý, chúng ta có thể sử dụng sách của những tác giả này, để gợi ý cho những user mới, lần đầu vào thăm trang web, sản phẩm của chúng ta.

III. **Model Building:**

- **DeepWalk** là một **mô hình nhúng đồ thị (graph embedding model)** được đề xuất bởi Bryan Perozzi, Rami Al-Rfou và Steven Skiena vào năm 2014.
  Mô hình DeepWalk nhúng các đỉnh (nodes) trong đồ thị thành các vectơ một chiều có kích thước cố định, giúp biểu diễn thông tin đồ thị dưới dạng các vectơ số thực.

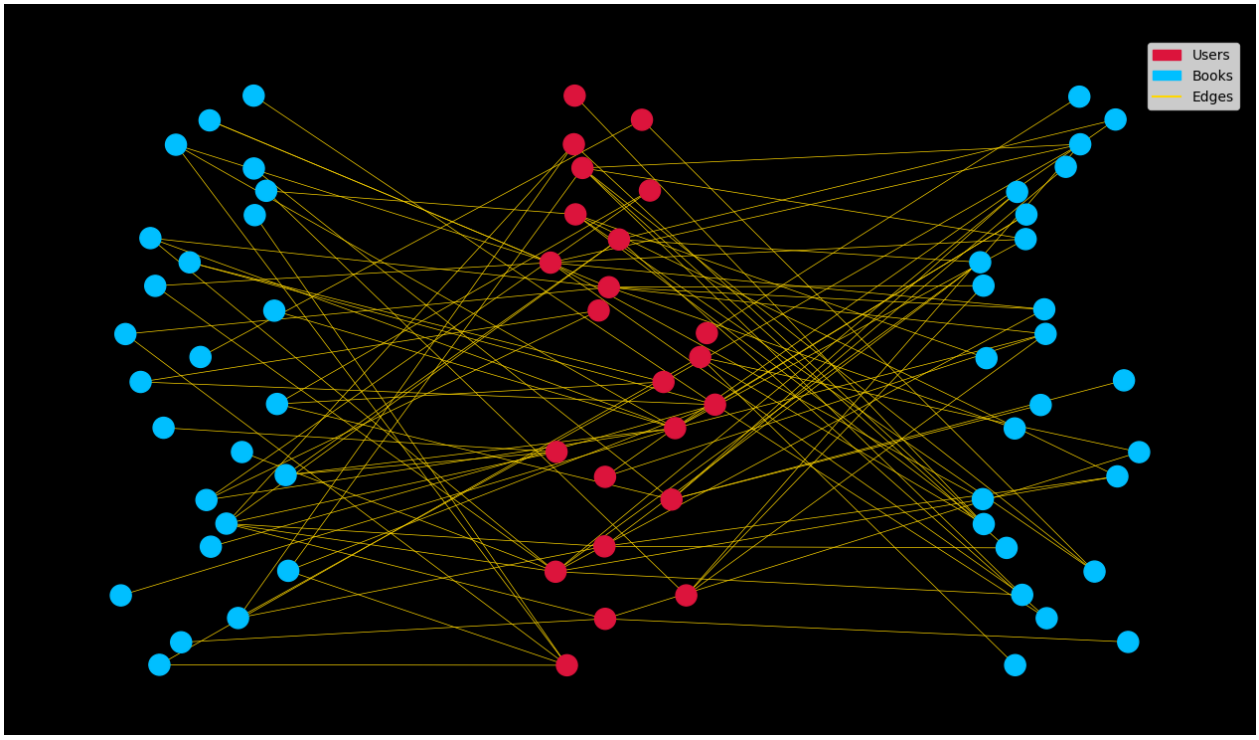

Fig 10. *Đồ thị mô tả mối quan hệ giữa user với book, và đánh giá của user, Dữ liệu đầu vào cho quá trình training thuật toán.*

- **DeepWalk** kết hợp hai khái niệm chính là đồ thị ngẫu nhiên (**random walk**) và mô hình không gian vectơ **(word2vec**) để học các nhúng đồ thị.
  Quá trình DeepWalk bao gồm hai giai đoạn chính:
- **Random Walk Generation:** Trong giai đoạn này**, DeepWalk** thực hiện tạo ra các **chuỗi random walk** (đường đi ngẫu nhiên) trên đồ thị. **Random walk** là quá trình di chuyển ngẫu nhiên qua các đỉnh của đồ thị bằng cách đi từ đỉnh hiện tại đến một đỉnh kế tiếp được chọn ngẫu nhiên. Quá trình này tạo ra một tập hợp các chuỗi random walk từ các đỉnh khác nhau trong đồ thị.
- **Skip-gram Model Training:** Sau khi có được tập hợp các chuỗi random walk, DeepWalk sử dụng **mô hình skip-gram (tương tự như mô hình word2vec)** để học các vectơ nhúng đồ thị. **Mô hình skip-gram** học cách dự đoán các đỉnh kế tiếp trong chuỗi random walk dựa trên đỉnh hiện tại. Quá trình này tạo ra các vectơ nhúng (embedding vector) cho các đỉnh trong đồ thị, trong đó mỗi vectơ đại diện cho một đỉnh và chứa thông tin về mối quan hệ giữa đỉnh đó và các đỉnh lân cận.

→ Như vậy, sau quá trình training, ta có thể xây dựng hệ thống gợi ý bằng cách sử dụng các vector tương đồng. Việc sử dụng **Deepwalk** sẽ cho chúng ta những kết quả bất ngờ từ dữ liệu. Các node tương đồng, phản ảnh lên sự tương sự trong hành vi,

những cuốn sách được gợi ý chéo cho những người dùng có sở thích đọc sách gần giống nhau, những cuốn sách nằm trong ngữ cảnh giống nhau, được thích bởi 2 user tương đồng, cũng sẽ được gợi ý cho user.

- **Giải thích một số siêu tham số trong quá trình training:**

- ***min_edge_weight:***
  → Giá trị trọng số tối thiểu của một cạnh, tức là khi khởi tạo đồ thị, chỉ khởi tạo những cạnh nào có trọng số lớn hơn *min_edge_weight*, các giá trị nhỏ hơn giá trị này sẽ được bỏ qua.
- ***percent_select_node:***
  → Phần trăm số đỉnh được chọn để đưa vào quá trình training.
- ***max_walk_length:***
  → số lượng bước đi tối đa, trong một lần khởi tạo bước đi giữa các node.
- ***n_walks_per_node:***
  → số lần 1 node tồn tại trong chuỗi random walk.

** Tham số của mô hình Word2Vec:

- **vector_size: kích thước của vector embedding.**
- **window:** kích thước cửa sổ xác định số lượng từ (node) lân cận sẽ được xem xét khi dự đoán từ hiện tại trong mô hình skip-gram hoặc từ hiện tại sẽ được dự đoán trong mô hình CBOW.
- **epochs**: số lần huấn luyện toàn bộ dữ liệu.

IV. **Đánh giá, thảo luận mô hình:**

- Đối với hệ thống Recommendation, chúng ta có thể đánh giá hiệu suất của mô hình DeepWalk dựa trên phản hồi của người dùng đối với các gợi ý mà chúng ta đưa ra, đồng thời so sánh với kết quả của các phương pháp g ợi ý khác, như collaborative filtering hoặc content-based recommendation, để đánh giá hiệu suất của nó.
- **Vấn đề Coldstart:** Vấn đề này xảy ra khi một user mới bắt đầu vào hệ thống, và chúng ta chưa có nhiều dữ liệu về hành vi để gợi ý cho user này.
  - Đối với vấn đề này, chúng ta có thể lựa chọn nhiều cách xử lý, bao gồm gợi ý cho user này những cuốn sách nổi bật trên hệ thống, hoặc trên quốc gia dựa vào location của họ. Hoặc sử dụng độ tương đồng dựa trên một số thông tin profile hữu dụng của người đó với một người đã có sẵn trong hệ thống, nhằm mục đích đưa ra các gợi ý tốt nhất cho user.

THANK YOU!