



## Q1. MAE vs. RMSE

- The following table shows the predicted ratings from the RS model (top) and the actual ratings that a user has given to a list of items (below).
- Note that the scores are in range [1, 10].

Items	1	2	3	4	5	6	7	8	9	10
Predicted	4	7	6	5	7	3	5	7	4	3
Actual	3	6	3	6	7	5	4	5	7	5

- Calculate the errors using the following metrics: MAE, NMAE, RMSE, and NRMSE.

giải bài này chi tiết và giải thích từng bước

Chúng ta cần tính toán bốn loại lỗi:

- Mean Absolute Error (MAE)
- Normalized Mean Absolute Error (NMAE)
- Root Mean Squared Error (RMSE)
- Normalized Root Mean Squared Error (NRMS) ↓

### Bước 1: Xác định dữ liệu

Dữ liệu từ bảng:

- Dự đoán (Predicted): [4, 7, 6, 5, 7, 3, 5, 7, 4, 3]
- Thực tế (Actual): [3, 6, 3, 6, 7, 5, 4, 5, 7, 5]

Sai số từng phần tử (lấy giá trị tuyệt đối và bình phương chênh lệch):

- |4 - 3| = 1
- |7 - 6| = 1
- |6 - 3| = 3
- |5 - 6| = 1
- |7 - 7| = 0
- |3 - 5| = 2
- |5 - 4| = 1
- |7 - 5| = 2
- |4 - 7| = 3
- |3 - 5| = 2



Tổng lỗi tuyệt đối:  $1 + 1 + 3 + 1 + 0 + 2 + 1 + 2 + 3 + 2 = 16$

Tổng bình phương sai số:  $1^2 + 1^2 + 3^2 + 1^2 + 0^2 + 2^2 + 1^2 + 2^2 + 3^2 + 2^2 = 1 + 1 + 9 + 1 + 0 + 4 + 1 + 4 + 9 + 4 = 34$

### Bước 2: Tính MAE

$$MAE = \frac{\sum |y_{\text{pred}} - y_{\text{true}}|}{n}$$

$$MAE = \frac{16}{10} = 1.6$$



### Bước 3: Tính NMAE

NMAE chuẩn hóa theo phạm vi  $\max - \min$  (ở đây là  $10 - 1 = 9$ ):

$$NMAE = \frac{MAE}{\max - \min} = \frac{1.6}{9} = 0.1778$$

### Bước 4: Tính RMSE

$$RMSE = \sqrt{\frac{\sum (y_{\text{pred}} - y_{\text{true}})^2}{n}}$$

$$RMSE = \sqrt{\frac{34}{10}} = \sqrt{3.4} \approx 1.84$$

### Bước 5: Tính NRMSE

NRMSE cũng chuẩn hóa theo phạm vi ( $10 - 1 = 9$ ):

$$NRMSE = \frac{RMSE}{\max - \min} = \frac{1.84}{9} = 0.2044$$

Kết luận:

- MAE = 1.6
- NMAE ≈ 0.1778
- RMSE ≈ 1.84
- NRMSE ≈ 0.2044

Vậy sai số lớn nhất là RMSE, và nếu cần so sánh giữa các mô hình trên nhiều bộ dữ liệu, NMAE và NRMSE là những chỉ số quan trọng vì chúng đã được chuẩn hóa theo phạm vi dữ liệu.

chi tiết hơn được không? cả mấy bước tính

Chắc chắn rồi! Chúng ta sẽ tính từng bước thật chi tiết để dễ hiểu.



### Bước 1: Xác định dữ liệu

Dựa vào bảng:

Items	1	2	3	4	5	6	7	8	9	10
Predicted	4	7	6	5	7	3	5	7	4	3
Actual	3	6	3	6	7	5	4	5	7	5

Chúng ta sẽ sử dụng công thức:

$$\text{Error}_i = \text{Predicted}_i - \text{Actual}_i$$

Tính từng giá trị lỗi:

Items	1	2	3	4	5	6	7	8	9	10
Error	4 - 3 = 1	7 - 6 = 1	6 - 3 = 3	5 - 6 = -1	7 - 0 = 7	3 - 5 = -2	5 - 4 = 1	7 - 5 = 2	4 - 7 = -3	3 - 5 = -2

### Bước 2: Tính MAE (Mean Absolute Error)

Công thức:

$$MAE = \frac{1}{n} \sum_{i=1}^n |\text{Error}_i|$$

Tính tổng lỗi tuyệt đối:

$$\begin{aligned} & |1| + |1| + |3| + |-1| + |0| + |-2| + |1| + |2| + |-3| + |-2| \\ & = 1 + 1 + 3 + 1 + 0 + \underset{\downarrow}{1} + 2 + 3 + 2 = 16 \end{aligned}$$

Tính MAE:

$$MAE = \frac{16}{10} = 1.6$$

### Bước 3: Tính NMAE (Normalized Mean Absolute Error)

NMAE chuẩn hóa theo phạm vi  $\max - \min$  của dữ liệu, ở đây khoảng giá trị là:

$$\max - \min = 10 - 1 = 9$$

Công thức:

$$NMAE = \frac{MAE}{\max - \min}$$
$$NMAE = \frac{\downarrow}{9} \approx 0.1778$$

### Bước 4: Tính RMSE (Root Mean Squared Error)

Công thức:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (\text{Error}_i)^2}$$

Tính bình phương sai số:

$$1^2 + 1^2 + 3^2 + (-1)^2 + 0^2 + (-2)^2 + 1^2 + 2^2 + (-3)^2 + (-2)^2 \\ = 1 + 1 + 9 + 1 + 0 + 4 + 1 + 4 + 9 + 4 = 34$$

Tính RMSE:

$$RMSE = \sqrt{\frac{34}{10}} = \sqrt{3.4} \approx 1.844$$

### Bước 5: Tính NRMSE (Normalized Root Mean Squared Error)

Chuẩn hóa theo phạm vi  $\max - \min$ :

$$NRMSE = \frac{RMSE}{\max - \min}$$
$$NRMSE = \frac{1.844}{9} \approx 0.2044$$

## Kết luận

Metric	Giá trị
MAE	1.6
NMAE	0.1778
RMSE	1.844
NRMSE	0.2044

- MAE đo lường sai số trung bình tuyệt đối, dễ hiểu nhưng không phản ánh độ lớn của sai số lớn.
- NMAE giúp so sánh trên nhiều tập dữ liệu bằng cách chuẩn hóa giá trị.
- RMSE phạt nặng hơn đối với sai số lớn vì nó dùng bình phương.
- NRMSE giúp chuẩn hóa RMSE, hữu ích khi so sánh giữa nhiều mô hình khác nhau.

Bạn có thể sử dụng NMAE hoặc NRMSE nếu muốn đánh giá tổng thể mô hình trên nhiều tập dữ liệu có quy mô khác nhau.

## Q2. Pearson vs. Spearman

- The following table shows the predicted ratings from the RS model (top) and the actual ratings that a user has given to a list of items (below).
- Note that the scores are in range [1, 10].

Items	1	2	3	4	5	6	7	8	9	10
Predicted	4	7	6	5	7	3	5	7	4	3
Actual	3	6	3	6	7	5	4	5	7	5

- Calculate the values of Pearson correlation coefficient and Spearman rank's correlation.

Chúng ta sẽ tính hệ số tương quan Pearson và Spearman từng bước chi tiết.



### Bước 1: Xác định dữ liệu

Dựa vào bảng:

Items	1	2	3	4	5	6	7	8	9	10
Predicted	4	7	6	5	7	3	5	7	4	3
Actual	3	6	3	6	7	5	4	5	7	5

Ký hiệu:

- $X$  là danh sách giá trị Predicted
- $Y$  là danh sách giá trị Actual

### Bước 2: Tính hệ số tương quan Pearson

Hệ số tương quan Pearson được tính bằng công thức:

$$r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2} \sqrt{\sum(Y_i - \bar{Y})^2}}$$

#### Tính trung bình của X và Y

$$\bar{X} = \frac{4 + 7 + 6 + 5 + 7 + 3 + 5 + 7 + 4 + 3}{10} = \frac{51}{10} = 5.1$$
$$\bar{Y} = \frac{3 + 6 + 3 + 6 + 7 + 5 + 4 + 5 + 7 + 5}{10} = \frac{51}{10} = 5.1$$

#### Tính các phần tử trong công thức Pearson

Tạo bảng các giá trị:

Items	X	Y	$X - \bar{X}$	$Y - \bar{Y}$	$(X - \bar{X})(Y - \bar{Y})$	$(X - \bar{X})^2$	$(Y - \bar{Y})^2$
1	4	3	-1.1	-2.1	2.1	1.21	4.41
2	7	6	1.9	0.9	1.71	3.61	0.81
3	6	3	0.9	-2.1	-1.89	0.81	4.41
4	5	6	-0.1	0.9	-0.09	0.01	0.81
5	7	7	1.9	1.9	3.61	3.61	3.61
6	3	5	-2.1	-0.1	0.21	4.41	0.01
7	5	4	-0.1	-1.1	0.11	0.01	1.21
8	7	5	1.9	-0.1	-0.19	3.61	0.01
9	4	7	-1.1	1.9	-2.09	1.21	3.61
10	3	5	-2.1	-0.1	0.21	4.41	0.01

Tính tổng:

$$\sum(X - \bar{X})(Y - \bar{Y}) = 3.69$$

$$\sum(X - \bar{X})^2 = 21.9$$

$$\sum(Y - \bar{Y})^2 = 18.9$$

Tính Pearson:

$$r = \frac{3.69}{\sqrt{21.9} \times \sqrt{18.9}}$$

$$r = \frac{3.69}{\sqrt{414.9}} = \frac{3.69}{20.38} \approx 0.181$$

### Bước 3: Tính hệ số tương quan Spearman

Hệ số tương quan Spearman dựa trên thứ hạng của các giá trị:

1. Xếp hạng các giá trị

Dữ liệu xếp hạng như sau:



Items	X	Rank(X)	Y	Rank(Y)
1	4	7.5	3	9
2	7	2.5	6	3.5
3	6	5	3	9
4	5	6	6	3.5
5	7	2.5	7	1.5
6	3	9	5	6
7	5	6	4	8
8	7	2.5	5	6
9	4	7.5	7	1.5
10	3	9	5	6

2. Tính độ lệch thứ hạng  $d_i$  và bình phương  $d_i^2$

$$d_i = \text{Rank}(X) - \text{Rank}(Y)$$

Items	Rank(X)	Rank(Y)	$d_i$	$d_i^2$
1	7.5	9	-1.5	2.25
2	2.5	3.5	-1	1
3	5	9	-4	16
4	6	3.5	2.5	6.25
5	2.5	1.5	1	1
6	9	6	3	9
7	6	8	-2	4
8	2.5	6	-3.5	12.25
9	7.5	1.5	6	36
10	9	6	3	9

$$\sum d_i^2 = 96.75$$

Tính Spearman:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

$$\rho = 1 - \frac{6 \times 96.75}{10(100 - 1)}$$

$$\rho = 1 - \frac{580.5}{990} = 1 - 0.5864 = 0.4136$$

## Kết quả cuối cùng:

Hệ số tương quan	↓	Giá trị
Pearson		0.181
Spearman		0.414

Pearson thấp do sự khác biệt giá trị tuyệt đối, trong khi Spearman cao hơn vì mô hình bảo toàn thứ hạng tốt hơn.

## Q3. Kendall Tau

- The following table shows the predicted ratings from the RS model (top) and the actual ratings that a user has given to a list of items (below).
- Note that the scores are in range [1, 10].

Items	I1	I2	I3	I4	I5	I6
Predicted	4	7	6	5	3	8
Actual	3	6	4	7	5	9

- Calculate the (original) Kendall Tau correlation.



Chúng ta sẽ tính hệ số tương quan Kendall Tau từng bước.

## Bước 1: Xác định dữ liệu

Dựa vào bảng:

Items	I1	I2	I3	I4	I5	I6
Predicted	4	7	6	5	3	8
Actual	3	6	4	7	5	9

Ký hiệu:

- $X$  là danh sách giá trị Predicted
- $Y$  là danh sách giá trị Actual



$$X = [4, 7, 6, 5, 3, 8]$$

$$Y = [3, 6, 4, 7, 5, 9]$$

## Bước 2: Cặp thứ hạng và số cặp nghịch

Hệ số tương quan Kendall Tau được tính bằng:

$$\tau = \frac{C - D}{\frac{1}{2}n(n - 1)}$$

- $C$  là số cặp đồng bậc (concordant pairs)
- $D$  là số cặp nghịch bậc (discordant pairs)
- $n$  là số lượng phần tử



Ta xét tất cả các cặp  $(X_i, X_j)$  với  $i < j$  và kiểm tra quan hệ tương ứng của  $Y_i, Y_j$ :

### Danh sách các cặp (i, j):

Cặp (i, j)	$X_i, X_j$	Quan hệ của $X$	$Y_i, Y_j$	Quan hệ của $Y$	Loại
(1,2)	(4,7)	$4 < 7$	(3,6)	$3 < 6$	Đồng bậc
(1,3)	(4,6)	$4 < 6$	(3,4)	$3 < 4$	Đồng bậc

(1,4)	(4,5)	4 < 5	(3,7)	3 < 7	Đồng bậc
(1,5)	(4,3)	4 > 3	(3,5)	3 < 5	Nghịch bậc
(1,6)	(4,8)	4 < 8	(3,9)	3 < 9	Đồng bậc
(2,3)	(7,6)	7 > 6	(6,4)	6 > 4	Đồng bậc
(2,4)	(7,5)	7 > 5	(6,7)	6 < 7	Nghịch bậc
(2,5)	(7,3)	7 > 3	(6,5)	6 > 5	Đồng bậc
(2,6)	(7,8)	7 < 8	(5,9) ↓	6 < 9	Đồng bậc
(3,4)	(6,5)	6 > 5	(4,7)	4 < 7	Nghịch bậc
(3,5)	(6,3)	6 > 3	(4,5)	4 < 5	Nghịch bậc
(3,6)	(6,8)	6 < 8	(4,9)	4 < 9	Đồng bậc
(4,5)	(5,3)	5 > 3	(7,5)	7 > 5	Đồng bậc
(4,6)	(5,8)	5 < 8	(7,9)	7 < 9	Đồng bậc
(5,6)	(3,8)	3 < 8	(5,9)	5 < 9	Đồng bậc

### Tính số lượng cặp

- Số cặp đồng bậc  $C = 11$
- Số cặp nghịch bậc  $D = 4$
- Tổng số cặp có thể có:

$$\frac{n(n-1)}{2} = \frac{6(6-1)}{2} = 15$$

↓

### Tính Kendall Tau

$$\tau = \frac{C - D}{\frac{1}{2}n(n-1)}$$

$$\tau = \frac{11 - 4}{15} = \frac{7}{15} = 0.467$$

### Kết quả cuối cùng:

$$\tau = 0.467$$

Kết luận: Hệ số Kendall Tau 0.467 cho thấy mô hình dự đoán có mức độ tương quan trung bình với thực tế, nghĩa là nó bảo toàn thứ hạng tương đối nhưng vẫn có một số sai lệch đáng kể.

### Q4. Precision, Recall, and F1

- The following table show 9 items and their relevancies (+: relevant or -: irrelevant).

Rank	1	2	3	4	5	6	7	8	9
+/-	+	-	+	+	-	-	-	+	-

- Calculate the values of Precision, Recall, and F1 at every rank position.

Chúng ta sẽ tính Precision, Recall, và F1-score tại mỗi vị trí xếp hạng.

### Bước 1: Xác định dữ liệu

Dữ liệu cho thấy danh sách 9 kết quả với thông tin về mức độ liên quan (+: liên quan, -: không liên quan):

Rank	1	2	3	4	5	6	7	8	9
Relevance	+	-	+	+	-	-	-	+	-

- So lượng tài liệu liên quan (relevant) tổng cộng = 4  
(các tài liệu ở vị trí 1, 3, 4, 8)
- 

## Bước 2: Công thức tính Precision, Recall, và F1-score

- Precision @ k: Tỷ lệ tài liệu liên quan trong top k.

$$P@k = \frac{\text{số tài liệu liên quan trong top } k}{k}$$

- Recall @ k: Tỷ lệ tài liệu liên quan đã tìm thấy  $\downarrow$  tổng số tài liệu liên quan.

$$R@k = \frac{\text{số tài liệu liên quan trong top } k}{\text{tổng số tài liệu liên quan}}$$

- F1-score @ k: Trung bình điều hòa của Precision và Recall.

$$F1@k = \frac{2 \times P@k \times R@k}{P@k + R@k} \quad (\text{nếu } P@k + R@k > 0)$$


---

## Bước 3: Tính toán

Ta đi qua từng vị trí k từ 1 đến 9:

Rank k	Relevant found	Precision P@k	Recall R@k	F1-score F1@k
1	1	1/1 = 1.00	1/4 = 0.25	$2(1.00 \times 0.25) / (1.00 + 0.25) = 0.40$
2	1	1/2 = 0.50	$1/4 \downarrow 0.25$	$2(0.50 \times 0.25) / (0.50 + 0.25) = 0.33$
3	2	2/3 = 0.67	2/4 = 0.50	$2(0.67 \times 0.50) / (0.67 + 0.50) = 0.57$
4	3	3/4 = 0.75	3/4 = 0.75	$2(0.75 \times 0.75) / (0.75 + 0.75) = 0.75$
5	3	3/5 = 0.60	3/4 = 0.75	$2(0.60 \times 0.75) / (0.60 + 0.75) = 0.67$
6	3	3/6 = 0.50	3/4 = 0.75	$2(0.50 \times 0.75) / (0.50 + 0.75) = 0.60$
7	3	3/7 = 0.43	3/4 = 0.75	$2(0.43 \times 0.75) / (0.43 + 0.75) = 0.55$
8	4	4/8 = 0.50	4/4 = 1.00	$2(0.50 \times 1.00) / (0.50 + 1.00) = 0.67$
9	4	4/9 = 0.44	4/4 = 1.00	$2(0.44 \times 1.00) / (0.44 + 1.00) = 0.61$

---

## Kết quả cuối cùng

Rank k	Precision P@k	Recall R@k	F1-score F1@k
1	1.00	0.25	0.40
2	0.50	0.25	0.33
3	0.67	0.50	0.57
4	0.75	0.75	0.75
5	0.60	0.75	0.67
6	0.50	0.75	0.60
7	0.43	0.75	0.55
8	0.50	1.00	0.67
9	0.44	1.00	0.61

## Nhận xét

- Precision giảm dần khi rank tăng.
- Recall tăng dần, đạt 1.0 ở  $k = 8$ .
- F1-score cao nhất ở  $k = 4$  và  $k = 8$  (0.75 và 0.67).  $\downarrow$

## Kết luận

- Hệ thống có độ chính xác cao ban đầu nhưng giảm dần do có nhiều tài liệu không liên quan.
- Hiệu suất tốt nhất đạt ở  $k = 4$  khi Precision và Recall đều đạt 0.75.

## Q5. NGCG

- The following table shows the predicted ratings from the RS model (top) and the actual ratings that a user has given to a list of items (below).
- Note that the scores are in range [1, 5].

Items	I1	I2	I3	I4
Predicted	3.23	2.13	3.12	4.58
Actual	2	1	4	5

↓ calculate the NDCG value.

### Tính toán NDCG (Normalized Discounted Cumulative Gain)

#### Bước 1: Công thức

NDCG được tính bằng:

$$NDCG = \frac{DCG}{IDCG}$$

trong đó:

- DCG (Discounted Cumulative Gain):

$$DCG = \sum_{i=1}^n \frac{\text{rel}_i}{\log_2(i+1)}$$

- IDCG (Ideal DCG): Là DCG khi các kết quả được ↓

#### Bước 2: Xếp hạng dựa trên dự đoán

Mô hình đề xuất các item với điểm số như sau:

Item	I1	I2	I3	I4
Predicted	3.23	2.13	3.12	4.58
Actual	2	1	4	5

Sắp xếp theo giá trị dự đoán giảm dần:

Rank	Item	Predicted	Actual (Relevance)
1	I4	4.58	5
2	I1	3.23	2
3	I3	3.12	4
4	I2	2.13	1

#### Bước 3: Tính DCG

Áp dụng công thức:

$$\begin{aligned} DCG &= \frac{5}{\log_2(1+1)} + \frac{2}{\log_2(2+1)} + \frac{4}{\log_2(3+1)} + \frac{1}{\log_2(4+1)} \\ &= \frac{5}{1} + \frac{2}{\log_2 3} + \frac{4}{\log_2 4} + \frac{1}{\log_2 5} \\ &= 5 + \frac{2}{1.585} + \frac{4}{2} + \frac{1}{2.322} \end{aligned}$$

$$= 5 + 1.262 + 2 + 0.431 = 8.693$$



#### Bước 4: Tính IDCG

Sắp xếp actual ratings theo thứ tự giảm dần:

Rank	Item	Actual (Relevance)
1	I4	5
2	I3	4
3	I1	2
4	I2	1

Áp dụng công thức:

$$\begin{aligned} IDCG &= \frac{5}{\log_2(1+1)} + \frac{4}{\log_2(2+1)} + \frac{2}{\log_2(3+1)} + \frac{1}{\log_2(4+1)} \\ &= 5 + \frac{4}{\log_2 3} + \frac{\downarrow 2}{\log_2 4} + \frac{1}{\log_2 5} \\ &= 5 + \frac{4}{1.585} + \frac{2}{2} + \frac{1}{2.322} \\ &= 5 + 2.524 + 1 + 0.431 = 8.955 \end{aligned}$$

#### Bước 5: Tính NDCG

$$NDCG = \frac{DCG}{IDCG} = \frac{8.693}{8.955} = 0.971$$

### Kết luận

Giá trị NDCG = 0.971, cho thấy hệ thống xếp hạng khá tốt vì giá trị này rất gần 1.

#### Q1. k-NN collaborative filtering

- This table shows the rating scores (from 1 to 5) of three IT figures, Mark Zuckerberg, Bill Gates and Guido van Rossum, on four programming languages, PHP, Apache Spark, Microsoft .NET and Python.

	PHP	Spark	.NET	Python
Mark Zuckerberg	4.5	4.0	1.5	4.5
Bill Gates	3.0	1.0	4.0	2.0
Guido van Rossum	4.5	2.0	5.0	

- Predict the rating score of Guido van Rossum on Apache Spark using k-Nearest Neighbors (with k = 1) user-based (or item-based) CF.

#### Dự đoán điểm đánh giá của Guido van Rossum trên Apache Spark bằng k-NN (k = 1)

Có hai cách tiếp cận để giải bài toán này:



- User-based CF (Lọc cộng tác dựa trên người dùng): Tìm người dùng có đánh giá tương tự Guido và dự đoán điểm dựa trên đánh giá của họ.
- Item-based CF (Lọc cộng tác dựa trên sản phẩm): Tìm các sản phẩm có điểm đánh giá tương tự với các sản phẩm Guido đã đánh giá, sau đó suy ra điểm.

#### Bước 1: Tính độ tương đồng (User-based CF)

Ta có 3 người dùng với các điểm đánh giá như sau:

User	PHP	Apache Spark	.NET	Python
Mark	4.5	4.0	1.5	4.5
Bill	3.0	1.0	4.0	2.0
Guido	4.5	?	↓	2.0

Chúng ta sẽ tính **độ tương đồng cosine** giữa Guido và Mark, Guido và Bill bằng công thức:

$$\text{similarity} = \frac{\sum(A_i \cdot B_i)}{\sqrt{\sum A_i^2} \cdot \sqrt{\sum B_i^2}}$$

#### Tính độ tương đồng giữa Guido và Mark

Sử dụng 3 giá trị chung (PHP, .NET, Python):

$$\text{numerator} = (4.5 \times 4.5) + (2.0 \times 1.5) + (5.0 \times 4.5) = 20.25 + 3 + 22.5 = 45.75$$

$$\text{denominator} = \sqrt{(4.5^2 + 2.0^2 + 5.0^2)} \times \sqrt{(4.5^2 + 1.5^2 + 4.5^2)}$$

$$= \sqrt{20.25 + 4 + 25} \times \sqrt{20.25 + 2.25 + 20.25}$$

$$= \sqrt{49.25} \times \sqrt{42.75} = 7.02 \times 6.54 = 45.91$$

$$\text{similarity}(Guido, Mark) = \frac{45.75}{45.91} \approx 0.996$$

#### Tính độ tương đồng giữa Guido và Bill



Dữ liệu chung (PHP, .NET, Python):

$$\text{numerator} = (4.5 \times 3.0) + (2.0 \times 4.0) + (5.0 \times 2.0) = 13.5 + 8 + 10 = 31.5$$

$$\text{denominator} = \sqrt{(4.5^2 + 2.0^2 + 5.0^2)} \times \sqrt{(3.0^2 + 4.0^2 + 2.0^2)}$$

$$= \sqrt{49.25} \times \sqrt{29} = 7.02 \times 5.39 = 37.85$$

$$\text{similarity}(Guido, Bill) = \frac{31.5}{37.85} \approx 0.832$$

#### Chọn người dùng gần nhất

Vì Mark có độ tương đồng cao nhất (0.996), ta sử dụng điểm của Mark để dự đoán:

Dự đoán điểm của Guido trên Apache Spark = 4.0

#### Kết luận



Dự đoán điểm của Guido van Rossum trên Apache Spark = 4.0 (theo k-NN với k = 1, user-based CF).

## Q2. k-NN collaborative filtering

- The following table shows the rating scores (from 1 to 5) of four characters on five book titles. An entry with question marks means no rating yet.

	Book1	Book2	Book3	Book4	Book5
Alice	1	2	5	?	1
George	5	?	1	2	5
Mary	?	3	4	3	4
Tom	1	1	5	4	?

- Predict the rating score of Tom on Book5 using k-Nearest Neighbors (with k = 2) user-based (or item-based) CF.

#### Dự đoán điểm đánh giá của Tom trên Ljk5 bằng k-NN (k = 2)

Chúng ta có bảng đánh giá của các nhân vật đối với 5 cuốn sách:

User	Book1	Book2	Book3	Book4	Book5
Alice	1	2	5	?	1
George	5	?	1	2	5
Mary	?	3	4	3	4

Cần dự đoán điểm của Tom trên Book5 bằng phương pháp k-NN ( $k = 2$ ), sử dụng User-based CF.

### Bước 1: Tính độ tương đồng Cosine giữa Tom và những người khác

Sử dụng 4 giá trị chung (Book1, Book2, Book3, Book4) để tính toán độ tương đồng:

#### Tương đồng giữa Tom và Alice

$$\text{numerator} = (1 \times 1) + (1 \times 2) + (5 \times 5) + (4 \times ?)$$

Vì Alice chưa đánh giá Book4, ta bỏ qua:

$$\text{numerator} = 1 + 2 + 25 = 28$$

$$\begin{aligned}\text{denominator} &= \sqrt{(1^2 + 1^2 + 5^2 + 4^2)} \times \sqrt{(1^2 + 2^2 + 5^2)} \\ &= \sqrt{1 + 1 + 25 + 16} \times \sqrt{1 + 4 + 25} \\ &= \sqrt{43} \times \sqrt{30} \approx 6.56 \times 5.48 = 35.96\end{aligned}$$

$$\text{similarity}(Tom, Alice) = \frac{28}{35.96} \approx 0.779$$

#### Tương đồng giữa Tom và George

$$\text{numerator} = (1 \times 5) + (1 \times ?) + (5 \times 1) + (4 \times 2)$$

Vì George chưa đánh giá Book2, ta bỏ qua:

$$\text{numerator} = 5 + 5 + 8 = 18$$

$$\begin{aligned}\text{denominator} &= \sqrt{(1^2 + 1^2 + 5^2 + 4^2)} \times \sqrt{(5^2 + 1^2 + 2^2)} \\ &= \sqrt{43} \times \sqrt{30} = 35.96\end{aligned}$$

$$\text{similarity}(Tom, George) = \frac{18}{35.96} \approx 0.501$$

#### Tương đồng giữa Tom và Mary

$$\text{numerator} = (1 \times ?) + (1 \times 3) + (5 \times 4) + (4 \times 3)$$

Vì Mary chưa đánh giá Book1, ta bỏ qua:

$$\text{numerator} = 3 + 20 + 12 = 35$$

$$\begin{aligned}\text{denominator} &= \sqrt{43} \times \sqrt{(3^2 + 4^2 + 3^2)} \\ &= \sqrt{43} \times \sqrt{34} = 38.27\end{aligned}$$

$$\text{similarity}(Tom, Mary) = \frac{35}{38.27} \approx 0.915$$

### Bước 2: Dự đoán điểm cho Book5

Chọn  $k = 2$  người dùng gần nhất, tức là Mary (0.915) và Alice (0.779).

Sử dụng công thức:

$$\begin{aligned}\text{Predicted Rating} &= \frac{\sum(\text{similarity} \times \text{rating})}{\sum \text{similarity}} \\ &= \frac{(0.915 \times 4) + (0.779 \times 1)}{0.915 + 0.779} \\ &= \frac{3.66 + 0.779}{1.694} \\ &= \frac{4.43}{1.694} \approx 2.62\end{aligned}$$

## Kết luận

Dự đoán điểm đánh giá của Tom trên Book5 là 2.62 theo phương pháp k-NN (k = 2, user-based CF).

## Q2. k-NN collaborative filtering

- The following table show the rating scores (from 1 to 5) of four characters on five book titles. An entry with question marks means no rating yet.

	Book1	Book2	Book3	Book4	Book5
Alice	1	2	5	?	1
George	5	?	1	2	5
Mary	?	3	4	3	4
Tom	1	1	5	4	?

- Predict the rating score of Tom on Book5 using k-Nearest Neighbors (with k = 2) user-based (or item-based) CF.

Bạn có muốn mình tiếp tục tính toán lại dự đoán điểm của Tom trên Book5 theo k-NN hay bạn cần giúp đỡ với nội dung khác trong hình?

## Q2. k-NN collaborative filtering

- The following table show the rating scores (from 1 to 5) of four characters on five book titles. An entry with question marks means no rating yet.

	Book1	Book2	Book3	Book4	Book5
Alice	1	2	5	?	1
George	5	?	1	2	5
Mary	?	3	4	3	4
Tom	1	1	5	4	?

- Predict the rating score of Tom on Book5 using k-Nearest Neighbors (with k = 2) user-based (or item-based) CF.

giải bài

Chúng ta sẽ sử dụng phương pháp k-NN (k-Nearest Neighbors) để dự đoán điểm của Tom trên Book5 với  $k = 2$ .

### Bước 1: Tính độ tương đồng giữa Tom và các người dùng khác

Sử dụng công thức cosine similarity để tính độ tương đồng giữa Tom và các người dùng khác.

Dữ liệu có sẵn:

	Book1	Book2	Book3	Book4	Book5
Alice	1	2	5	?	1
George	5	?	1	2	5
Mary	?	3	4	3	4
Tom	1	1	5	4	?

Chúng ta cần tính toán độ tương đồng giữa Tom và Alice, George, Mary (bỏ qua các ô không có dữ liệu).

Sau đó chọn ra hai người dùng gần nhất (có giá trị tương đồng cao nhất).

### Bước 2: Dự đoán điểm của Tom trên Book5

Sau khi tìm được 2 người có độ tương đồng cao nhất, dùng công thức trung bình trọng số để dự đoán điểm.

Bạn có muốn mình thực hiện chi tiết các bước tính toán này không?



tính toán chi tiết, giải thích từng bước

Chúng ta sẽ giải bài toán sử dụng phương pháp k-NN dựa trên độ tương đồng Cosine (Cosine Similarity).

## Bước 1: Tính độ tương đồng giữa Tom và các người dùng khác

Sử dụng công thức cosine similarity:

$$\text{similarity}(A, B) = \frac{\sum(A_i \times B_i)}{\sqrt{\sum A_i^2} \times \sqrt{\sum B_i^2}}$$

### Tập dữ liệu

	Book1	Book2	Book3	Book4	Book5
Alice	1	2	5	?	1
George	5	?	1	2	5
Mary	?	3	4	3	4
Tom	1	1	5	4	?

Chúng ta tính độ tương đồng giữa Tom và từng người dùng khác.

## Tính độ tương đồng giữa Tom và Alice

### Bước 1: Chỉ xét các cột có giá trị

Các giá trị chung giữa Tom và Alice:

- Book1: 1, 1
- Book2: 1, 2
- Book3: 5, 5
- Book5: ?, 1 (bỏ qua vì Tom chưa có giá trị)

Dữ liệu cần tính:

$$\text{Tom} = (1, 1, 5)$$

$$\text{Alice} = (1, 2, 5)$$

### Bước 2: Áp dụng công thức

$$\sum(A_i \times B_i) = (1 \times 1) + (1 \times 2) + (5 \times 5) = 1 + 2 + 25 = 28$$

$$\sqrt{\sum A_i^2} = \sqrt{1^2 + 1^2 + 5^2} = \sqrt{1 + 1 + 25} = \sqrt{27}$$

$$\sqrt{\sum B_i^2} = \sqrt{1^2 + 2^2 + 5^2} = \sqrt{1 + 4 + 25} = \sqrt{30}$$

$$\text{similarity}(\text{Tom}, \text{Alice}) = \frac{28}{\sqrt{27} \times \sqrt{30}} = \frac{28}{\sqrt{810}} \approx 0.99$$

## Tính độ tương đồng giữa Tom và George

### Bước 1: Chỉ xét các cột có giá trị

Các giá trị chung giữa Tom và George:

- Book1: 1, 5
- Book3: 5, 1
- Book4: 4, 2
- Book5: ?, 5 (bỏ qua)

Dữ liệu cần tính:

$$\downarrow$$

$$\text{Tom} = (1, 5, 4)$$

$$\text{George} = (5, 1, 2)$$

**Bước 2: Áp dụng công thức**

$$\sum(A_i \times B_i) = (1 \times 5) + (5 \times 1) + (4 \times 2) = 5 + 5 + 8 = 18$$

$$\sqrt{\sum A_i^2} = \sqrt{1^2 + 5^2 + 4^2} = \sqrt{1 + 25 + 16} = \sqrt{42}$$

$$\sqrt{\sum B_i^2} = \sqrt{5^2 + 1^2 + 2^2} = \sqrt{25 + 1 + 4} = \sqrt{30}$$

$$\text{similarity}(\text{Tom}, \text{George}) = \frac{18}{\sqrt{42} \times \sqrt{30}} = \frac{18}{\sqrt{1260}} \approx 0.45$$

**Tính độ tương đồng giữa Tom và Mary****Bước 1: Chỉ xét các cột có giá trị**

Các giá trị chung giữa Tom và Mary:

- Book2: 1, 3
- Book3: 5, 4
- Book4: 4, 3
- Book5: ?, 4 (bỏ qua)

Dữ liệu cần tính:

$$\text{Tom} = (1, 5, 4)$$

$$\text{Mary} = (3, 4, 3)$$

**Bước 2: Áp dụng công thức**

$$\sum(A_i \times B_i) = (1 \times 3) + (5 \times \downarrow) + (4 \times 3) = 3 + 20 + 12 = 35$$

$$\sqrt{\sum A_i^2} = \sqrt{1^2 + 5^2 + 4^2} = \sqrt{1 + 25 + 16} = \sqrt{42}$$

$$\sqrt{\sum B_i^2} = \sqrt{3^2 + 4^2 + 3^2} = \sqrt{9 + 16 + 9} = \sqrt{34}$$

$$\text{similarity}(\text{Tom}, \text{Mary}) = \frac{35}{\sqrt{42} \times \sqrt{34}} = \frac{35}{\sqrt{1428}} \approx 0.95$$

**Bước 2: Chọn 2 người có độ tương đồng cao nhất**

- Alice: 0.99
- Mary: 0.95

Do đó, ta chọn Alice và Mary để dự đoán điểm của Tom.

**Bước 3: Dự đoán điểm của Tom cho Book5**

Dùng công thức dự đoán dựa trên trung bình trọng số:

$$\hat{r}_{\text{Tom}, \text{Book5}} = \frac{\sum(\text{similarity} \times \text{rating})}{\sum \text{similarity}}$$

$$\hat{r}_{\text{Tom}, \text{Book5}} = \frac{(0.99 \times 1) + (0.95 \times 4)}{0.99 + 0.95}$$

$$= \frac{0.99 + 3.8}{1.94} = \frac{4.79}{1.94} \approx 2.47$$

---

## Kết luận

Dự đoán điểm số của Tom trên Book5 là khoảng 2.47.

Nếu cần làm tròn, ta có thể dự đoán 2 hoặc 3.

Bạn có muốn mình giải thích thêm không? 😊

↪ ⌂ 📋 🔍 🔍 ↵

Hỏi bất kỳ điều gì



Tim kiếm

Suy luận



ChatGPT có thể mắc lỗi. Hãy kiểm tra các thông tin quan trọng.

?