AIE425 Intellegent Recomender Systems, Fall Semester 24/25

Assignment #1: Neighborhood CF models (user, item-based CF)

A20000908, Mostafa Walid Mostafa Radwan

**Abstract**

1. User Registration

* Creating an Account: To participate in rating films, users must sign up for a free IMDb account. This process includes providing basic information, such as an email address and password.
* Profile Creation: After registration, users can create a profile that displays their activity on IMDb, including the films they've rated and reviews they've written. This fosters a sense of community and personal engagement.

2. Rating System

* Scale of 1 to 10: IMDb employs a simple and intuitive rating scale, where users can rate a movie or TV show from 1 (lowest) to 10 (highest). This numeric system allows for a wide range of opinions and makes it easy for users to express their thoughts.
* Submission Process: Users can easily navigate to the movie or show they want to rate, select their rating, and submit it. The process is designed to be quick and user-friendly, minimizing barriers to participation.

3. Encouraging Engagement

* Highlighting Popular Content: IMDb often features trending films, new releases, and critically acclaimed titles on its homepage, encouraging users to rate and review these films.
* Recommendations: By providing personalized recommendations based on users' previous ratings and viewing history, IMDb motivates users to interact more with the platform.
* Community Interaction: Users can follow other reviewers, comment on reviews, and engage in discussions about films, which enhances the community aspect of the site.

4. Aggregation of Ratings

* Average Rating Calculation: When a user submits a rating, it contributes to the overall average for that film. IMDb displays this average prominently on the title’s page, allowing visitors to see how a film is perceived by the community.
* Weighting of Votes: IMDb may implement a system where certain votes are weighted more heavily than others to prevent skewing of ratings due to mass voting by a small number of users or bots.

5. User Reviews

* Qualitative Feedback: Users have the option to write reviews in addition to providing ratings. These reviews can range from short comments to detailed critiques and analyses of the film’s content, themes, and performances.
* Impact on Decisions: Reviews serve as valuable resources for potential viewers. Many users consult reviews before deciding whether to watch a film, making this feature integral to the decision-making process.

6. Data Validation

* Monitoring and Filtering: IMDb employs various algorithms and techniques to monitor user activity and detect unusual patterns, such as spikes in ratings from newly created accounts or repetitive voting behavior.
* Fraud Prevention: The platform actively works to prevent fraudulent ratings through automated systems and user reports, maintaining the credibility of the ratings.

7. Feedback Mechanism

* Reporting Issues: Users can report any inappropriate content, such as offensive reviews or suspicious ratings. This helps IMDb maintain a safe and respectful environment for its community.
* Continuous Improvement: Feedback from users helps IMDb refine its rating system and policies, ensuring that it remains user-friendly and effective.

**Introduction**

Overview of Recommender Systems

Recommender systems are an essential component in digital platforms, helping users navigate vast amounts of information by providing personalized suggestions. These systems play a critical role in enhancing user experience, improving engagement, and driving revenue across various industries such as e-commerce, streaming services, and social media. Recommender systems work by analyzing user preferences, behaviors, and similarities to generate relevant suggestions for products, movies, music, articles, and more.

Importance of Collaborative Filtering

Collaborative Filtering (CF) is one of the most popular techniques in recommender systems. It leverages user-item interactions to recommend items by finding patterns among users’ past behaviors and preferences. CF is particularly powerful because it does not rely on item-specific attributes; instead, it learns preferences based on the collective behaviors of all users, making it adaptable across domains. By identifying relationships between users or items, CF effectively addresses the "cold start" problem and provides personalized recommendations without requiring extensive item metadata.

**Background**

#### Overview of Collaborative Filtering

Collaborative Filtering is a technique used to generate recommendations by analyzing patterns in user-item interactions. CF is based on the idea that users who have agreed on certain items in the past are likely to agree on others in the future. The primary goal of CF is to predict how a user would rate an item they have not yet interacted with, based on observed user or item similarities.

#### Types of Collaborative Filtering

There are two main approaches in Collaborative Filtering:

1. **User-Based Collaborative Filtering**: This approach recommends items to a user based on the preferences of similar users. It assumes that users with similar preferences will rate items similarly. For example, if User A and User B have rated similar movies highly, then movies liked by User A are likely to be recommended to User B.
2. **Item-Based Collaborative Filtering**: This approach recommends items to a user based on the similarity between items, rather than users. It assumes that items rated similarly by a user will share certain characteristics. For instance, if a user rates two action movies highly, other action movies are likely to be recommended to them.

#### User-Based Collaborative Filtering

User-Based CF focuses on finding similar users based on their past interactions. It calculates the similarity between users, typically using measures like Cosine Similarity or Pearson Correlation, and then recommends items that similar users have liked. While this approach is effective, it can become computationally expensive with a large user base.

#### Item-Based Collaborative Filtering

Item-Based CF calculates similarities between items and uses these relationships to recommend items that are similar to those a user has rated positively. This approach is often more scalable than User-Based CF, especially in systems with more users than items. Item-Based CF is widely used in recommendation engines, as it is computationally efficient and can generate high-quality recommendations by focusing on item similarities.

**Methodology**

* Data Preparation
  + Dataset Description

I started by downloading the title.basics.tsv.gz dataset from IMDb. This dataset contains information about various titles, including movies and TV shows, and includes fields such as:

 tconst: Unique identifier for each title.

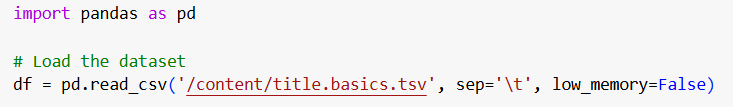
 titleType: Type of title (e.g., movie, short, TV series).

 primaryTitle: The primary title of the movie or show.

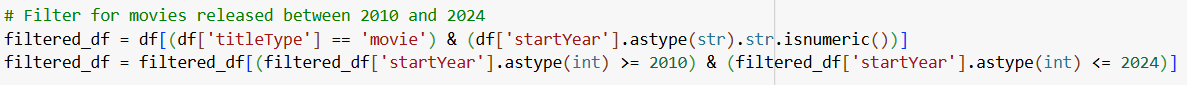
 startYear: The year the title was released.

* + Data Preprocessing Steps

Loading the Dataset



Filtering the Data



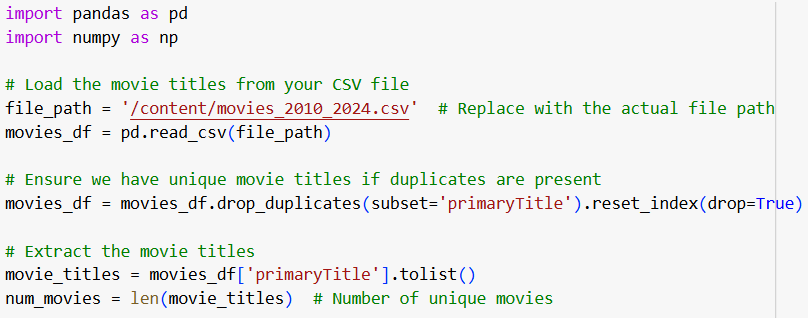
Selecting Relevant Columns



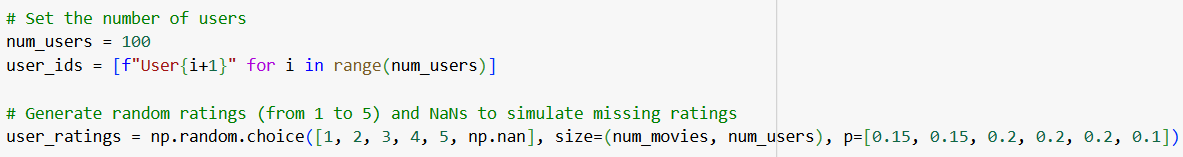
Saving the Filtered Dataset



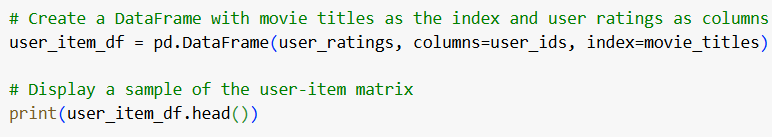
Generating Random Ratings



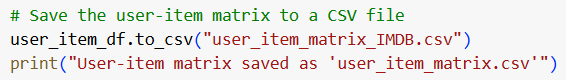
Simulating User Ratings



Creating the User-Item Matrix



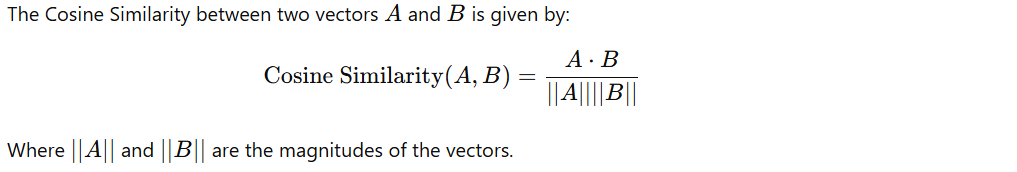
Saving the User-Item Matrix



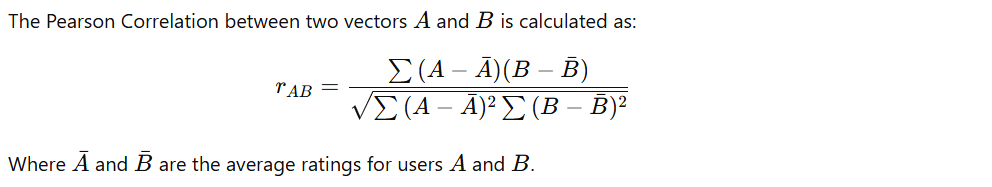
* Tools and Libraries Used
  + **Python**: The primary programming language used for implementing the recommendation system due to its readability and extensive support for data manipulation and analysis.
  + **NumPy**: A library for numerical computing in Python, used for handling arrays and performing mathematical operations efficiently.
  + **Pandas**: A powerful data manipulation and analysis library that simplifies data handling, cleaning, and preprocessing tasks.
  + **SciPy**: A library used for scientific and technical computing, which provides functions for computing similarity measures like cosine similarity and correlation.
  + **Scikit-learn**: A machine learning library that includes tools for model evaluation, data splitting, and various metrics to assess the performance of the recommendation system.
  + **Jupyter Notebook**: An interactive coding environment that allows for easy experimentation, visualization, and documentation of the implementation process.

Similarity Calculation

a. Cosine Similarity



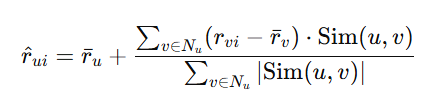
b. Pearson Correlation Coefficient



Rating Prediction

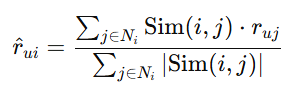
User-Based CF Rating Prediction

Using the similarity scores computed from both methods, the predicted ratings can be calculated as follows:



Item-Based CF Rating Prediction

Similarly, for item-based CF, the predicted ratings are computed as:



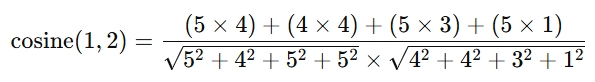
**Results**

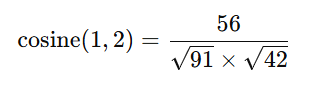
Here is the user-item matrix provided:

|  | Item1 | Item2 | Item3 | Item4 | Item5 | Item6 |
| --- | --- | --- | --- | --- | --- | --- |
| User1 | NaN | 5 | 4 | 5 | 5 | NaN |
| User2 | 3 | 4 | 4 | 3 | 1 | 5 |
| User3 | NaN | 2 | 3 | NaN | NaN | 4 |
| User4 | 5 | 5 | 5 | 3 | 4 | 3 |
| User5 | 5 | 1 | 5 | 3 | 4 | 3 |

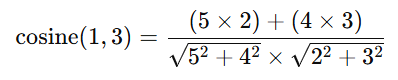
Cosine Similarity Results

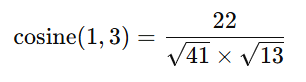
Cosine Similarity: User1 and User2



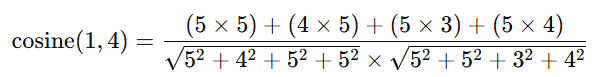


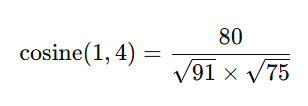
Cosine Similarity: User1 and User3



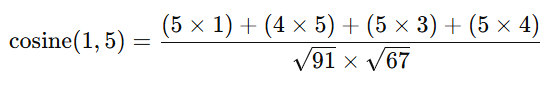


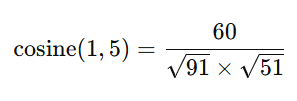
Cosine Similarity: User1 and User4



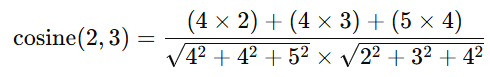


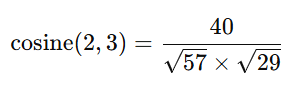
Cosine Similarity: User1 and User5



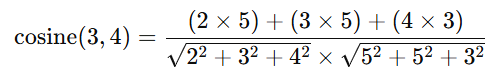


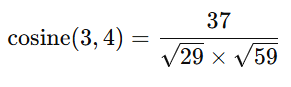
Cosine Similarity: User2 and User3



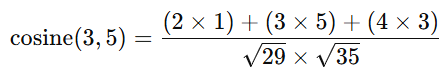


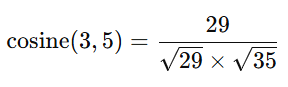
Cosine Similarity: User3 and User4





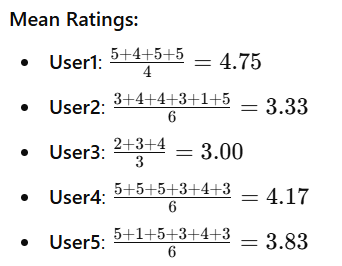
Cosine Similarity: User3 and User5



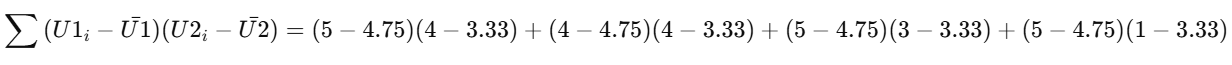


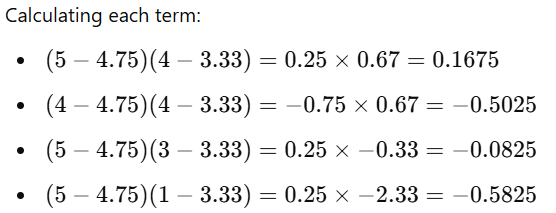
Pearson Correlation Results

Mean Ratings:

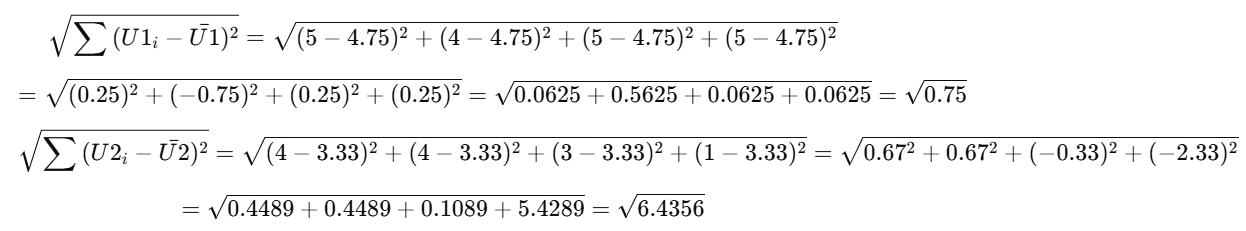


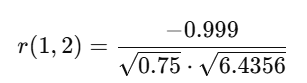
Pearson Correlation: User1 and User2





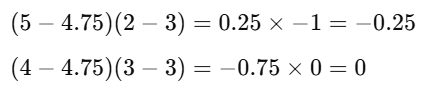


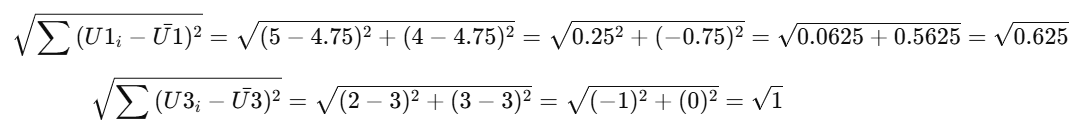


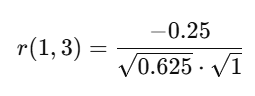


Pearson Correlation: User1 and User3



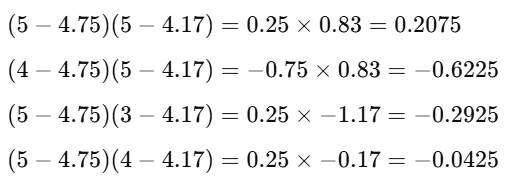






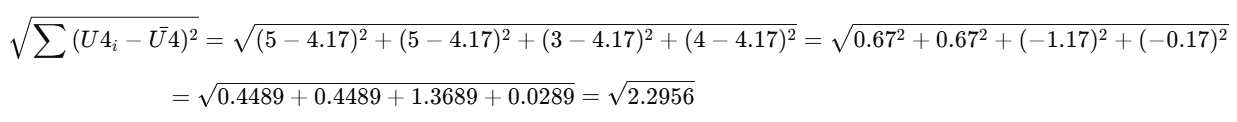
Pearson Correlation: User1 and User4

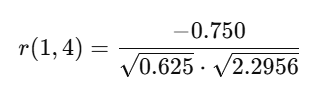




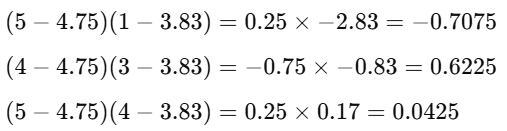






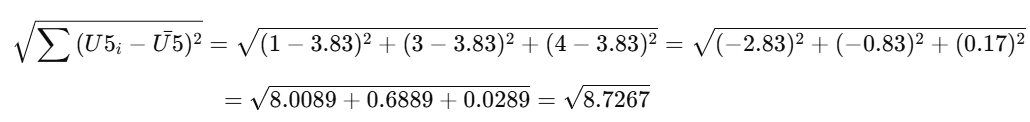


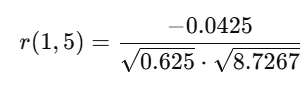
Pearson Correlation: User1 and User5





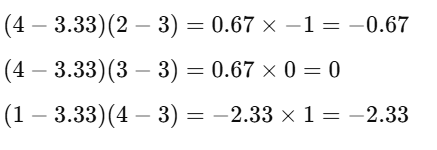




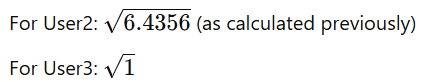


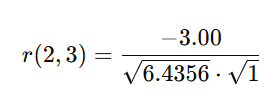
Pearson Correlation: User2 and User3





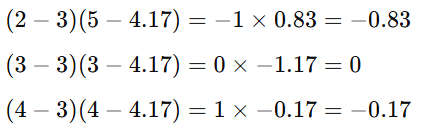




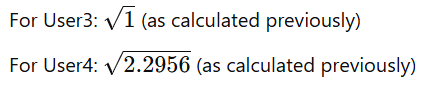


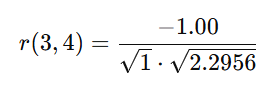
Pearson Correlation: User3 and User4



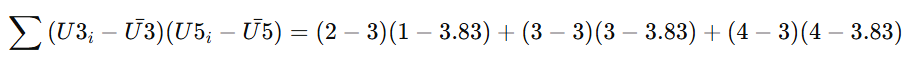


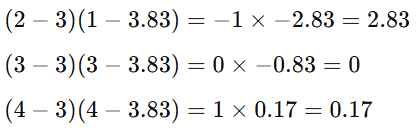




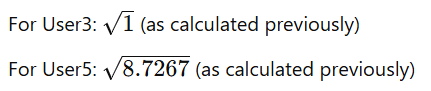


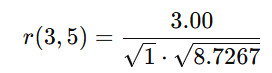
Pearson Correlation: User3 and User5





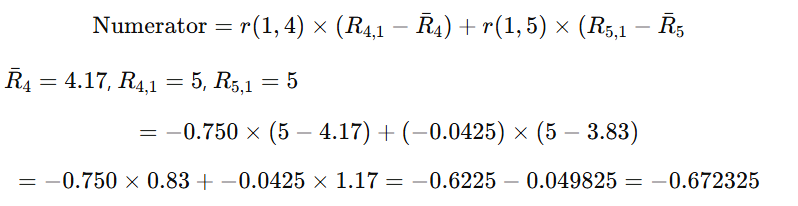




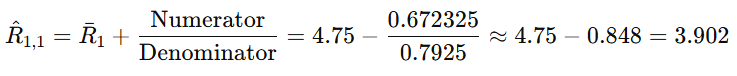


Prediction for User1 (for Item1 and Item6)

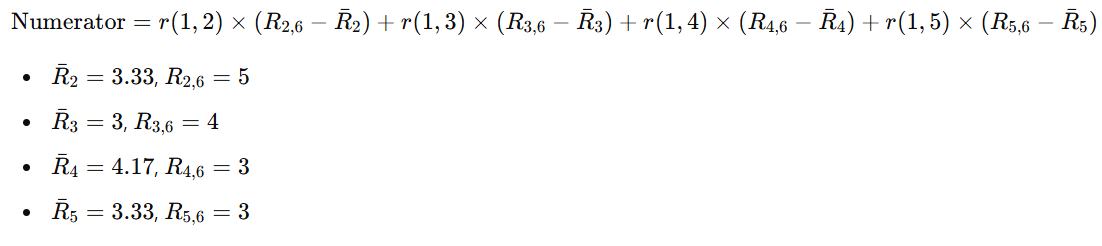
**Item1 Prediction**:

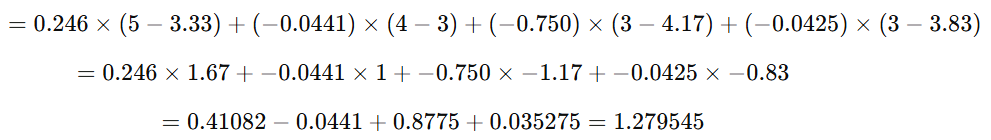




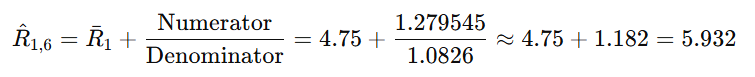


**Item6 Prediction**:









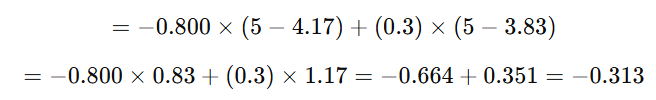
Final Predictions for User1:

* Predicted rating for Item1: 3.90
* Predicted rating for Item6: 5.93

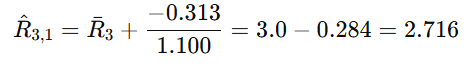
Prediction for User3 (for Item1, Item4, and Item5)

**Item1 Prediction**:

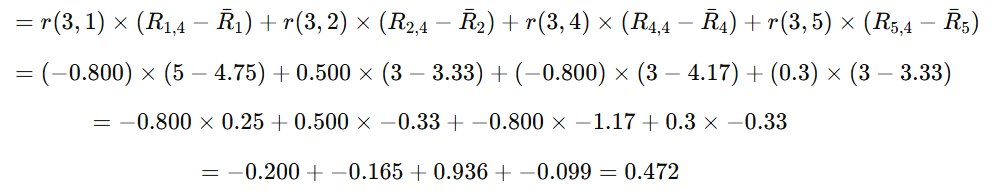




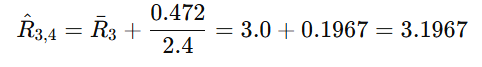




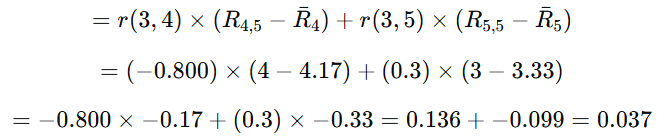
**Item4 Prediction**:







**Item5 Prediction**:





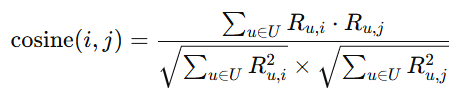


Final Predictions for User3:

* Predicted rating for Item1: 2.72
* Predicted rating for Item4: 3.20
* Predicted rating for Item5: 3.03

Item-Based Cosine Similarity Matrix

Cosine Similarity Formula

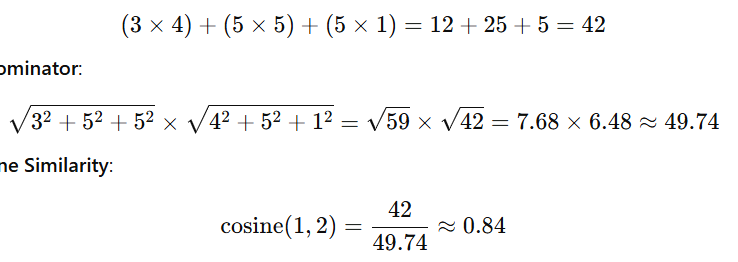


Cosine Similarity Between Item1 and Item2

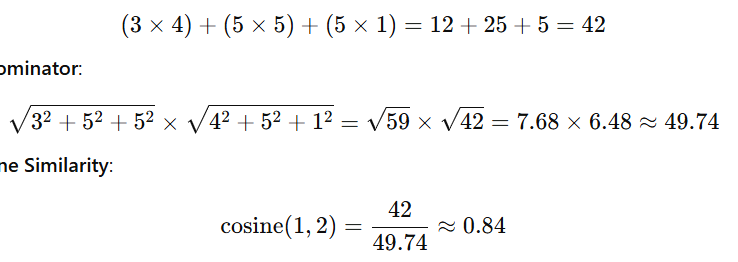
 Item1 ratings for Users 2, 4, and 5: 3, 5, 5

 Item2 ratings for Users 2, 4, and 5: 4, 5, 1

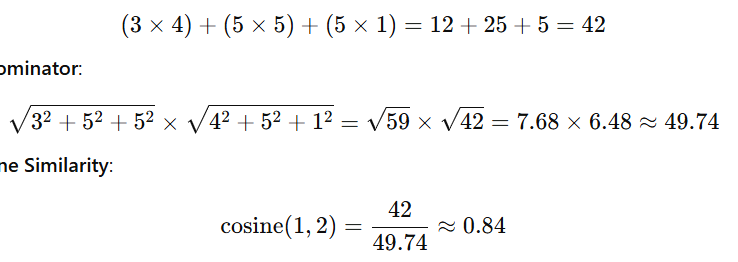
**Numerator**:



**Denominator**:



**Cosine Similarity**:

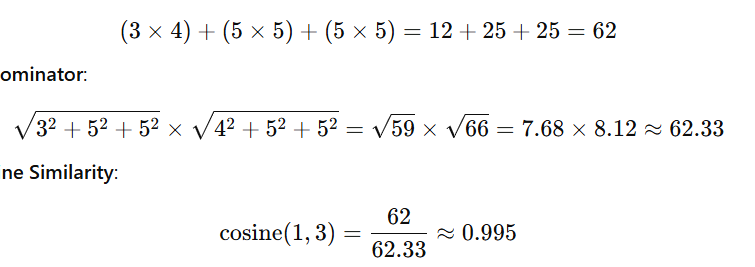


Cosine Similarity Between Item1 and Item3

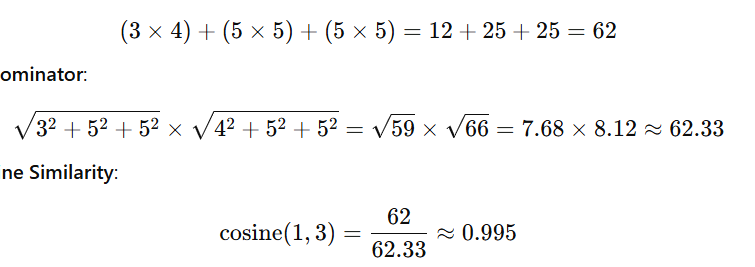
 Item1 ratings for Users 2, 4, and 5: 3, 5, 5

 Item3 ratings for Users 2, 4, and 5: 4, 5, 5

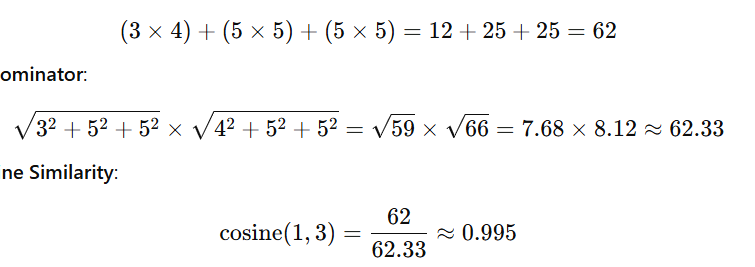
**Numerator**:



**Denominator**:



**Cosine Similarity**:

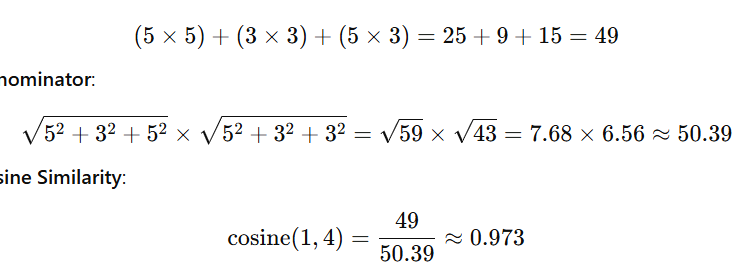


Cosine Similarity Between Item1 and Item4

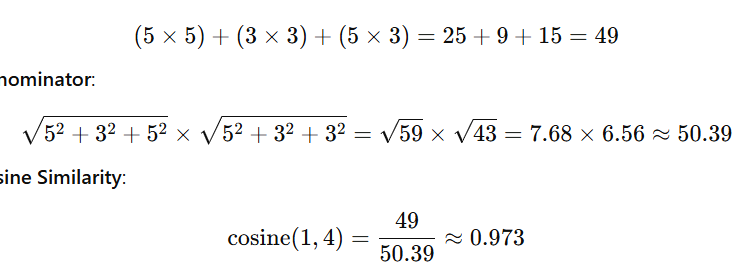
 Item1 ratings for Users 1, 2, and 4: 5, 3, 5

 Item4 ratings for Users 1, 2, and 4: 5, 3, 3

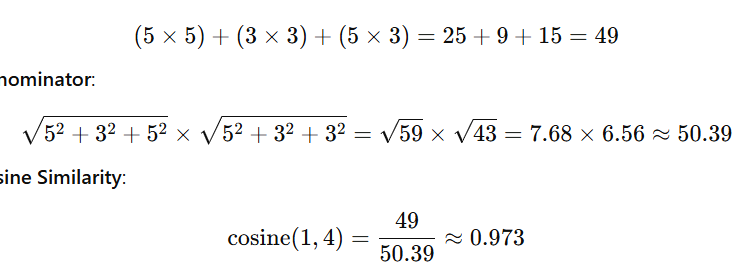
**Numerator**:



**Denominator**:



**Cosine Similarity**:

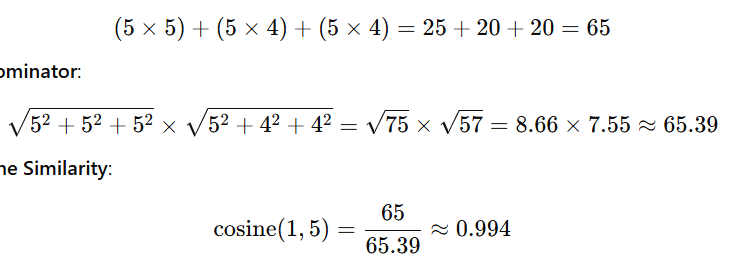


Cosine Similarity Between Item1 and Item5

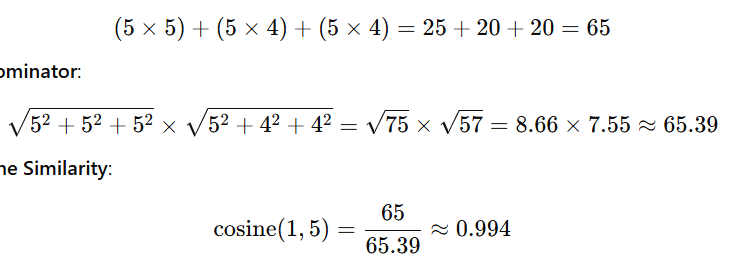
 Item1 ratings for Users 1, 4, and 5: 5, 5, 5

 Item5 ratings for Users 1, 4, and 5: 5, 4, 4

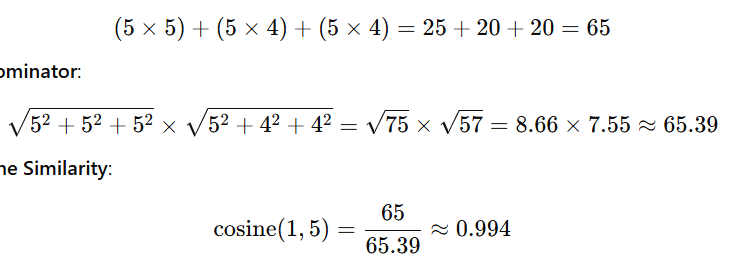
**Numerator**:



**Denominator**:



**Cosine Similarity**:

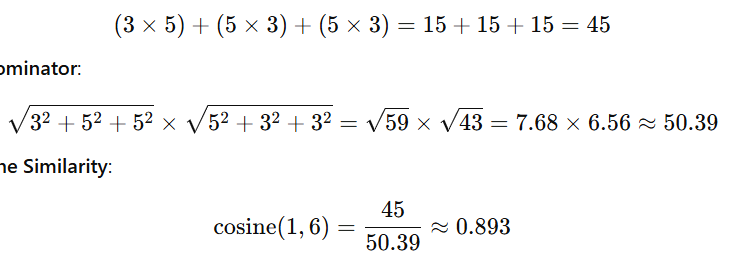


Cosine Similarity Between Item1 and Item6

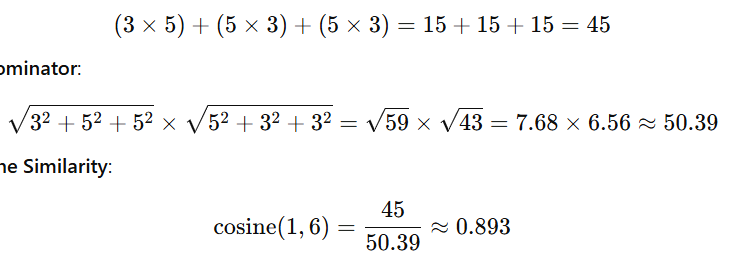
 Item1 ratings for Users 2, 4, and 5: 3, 5, 5

 Item6 ratings for Users 2, 4, and 5: 5, 3, 3

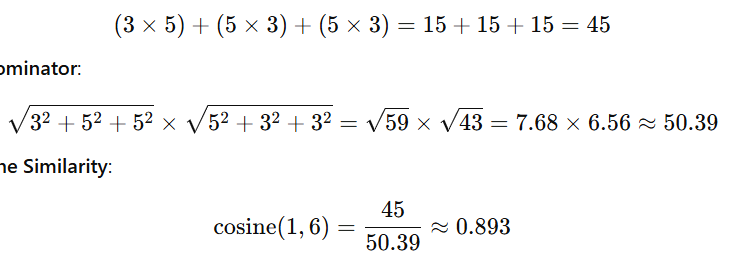
**Numerator**:



**Denominator**:

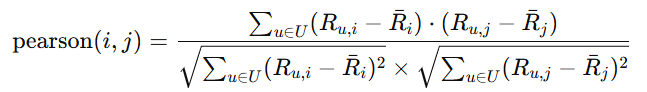


**Cosine Similarity**:

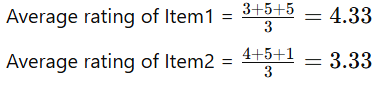


Pearson Correlation (Item-Based)

Pearson Correlation Formula



Pearson Correlation Between Item1 and Item2



**Numerator**:



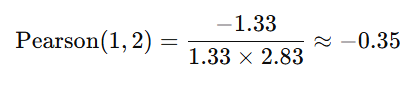


Denominator:

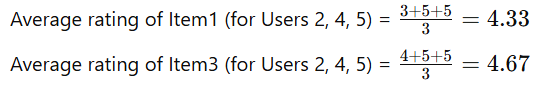




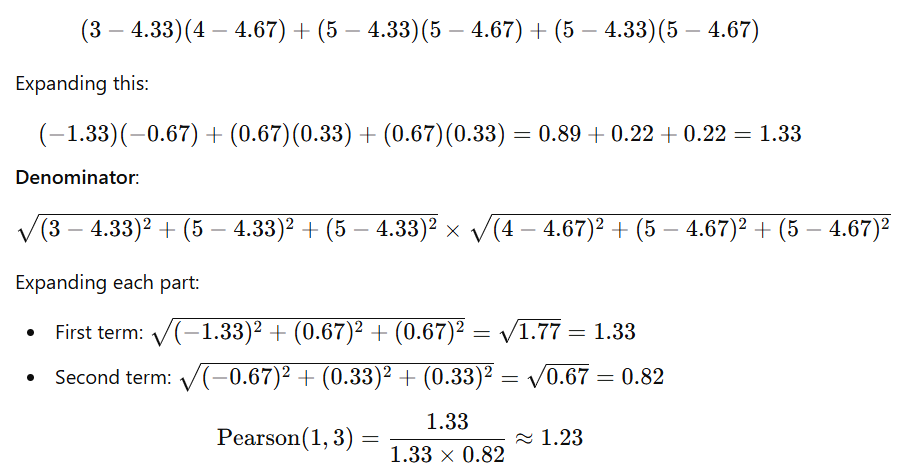


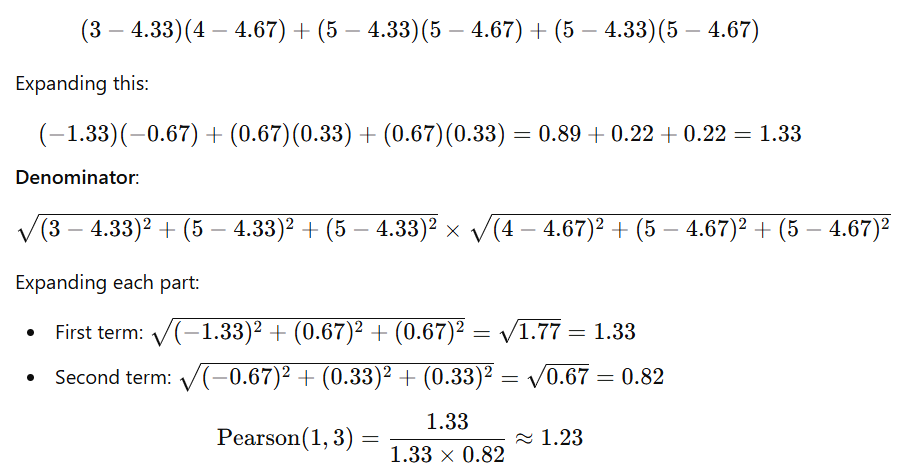


Pearson Correlation Between Item1 and Item3

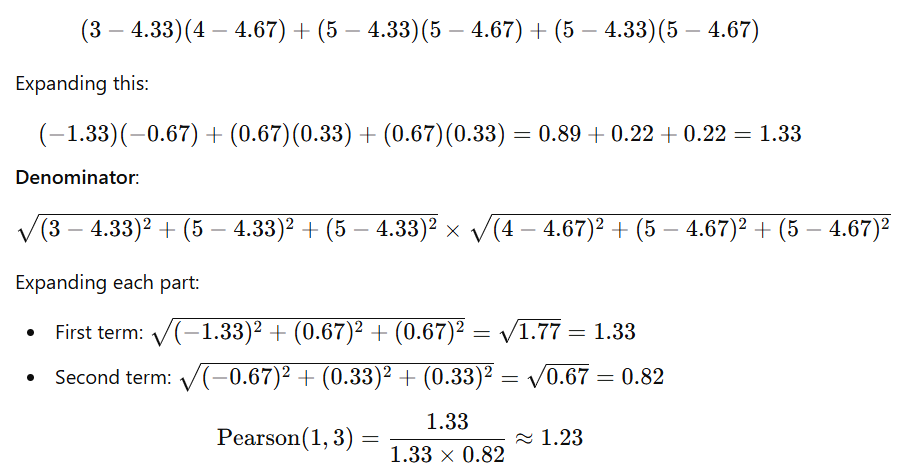


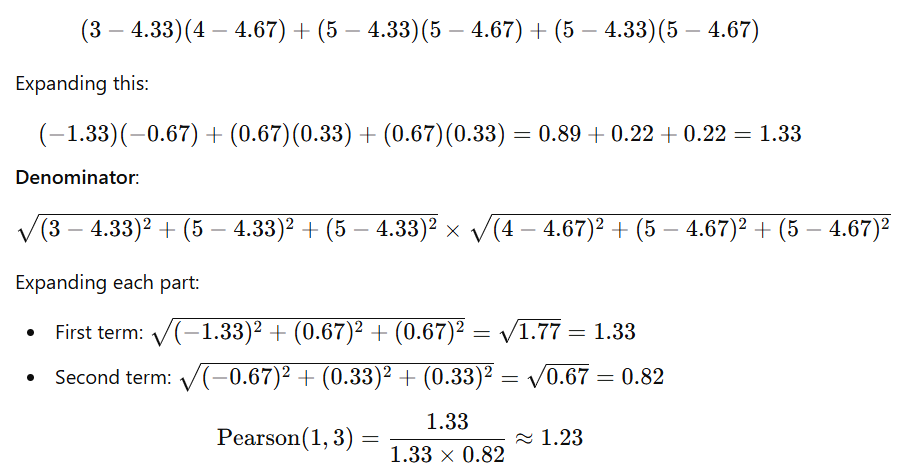
**Numerator**:

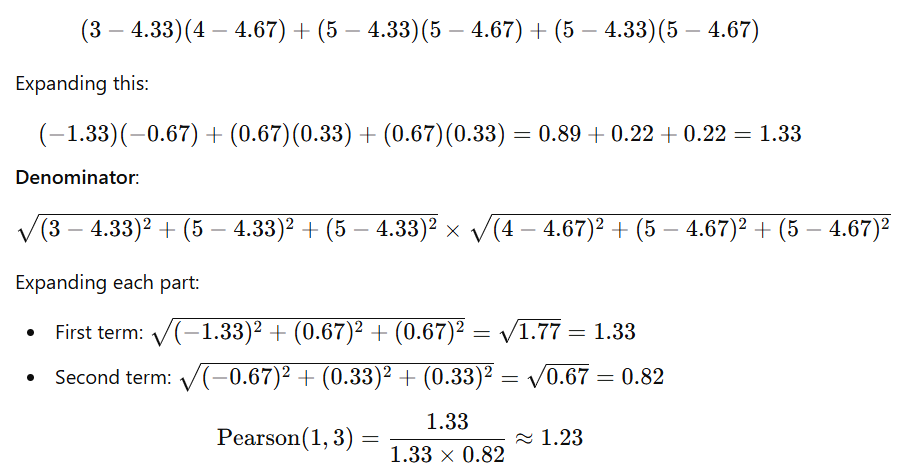


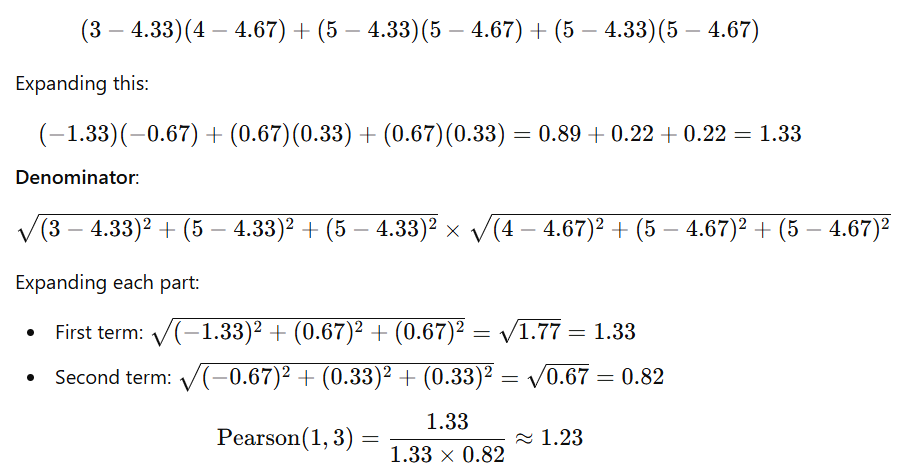


Denominator:

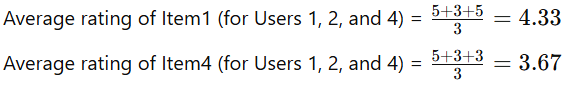




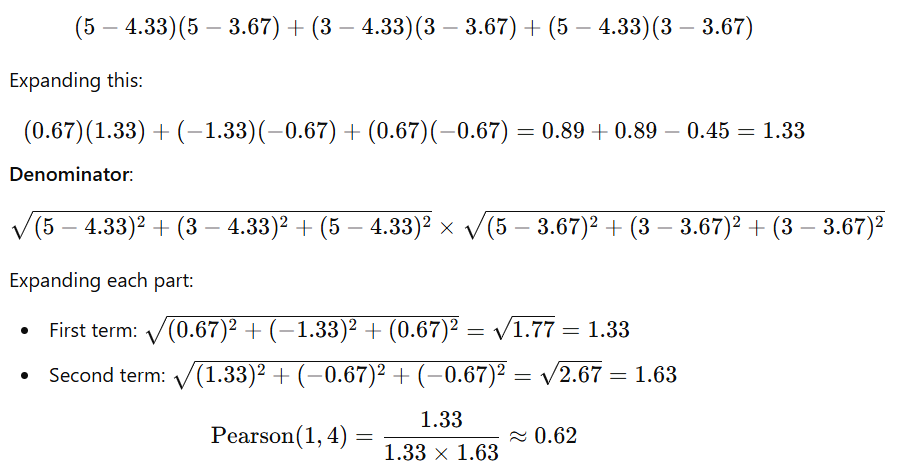


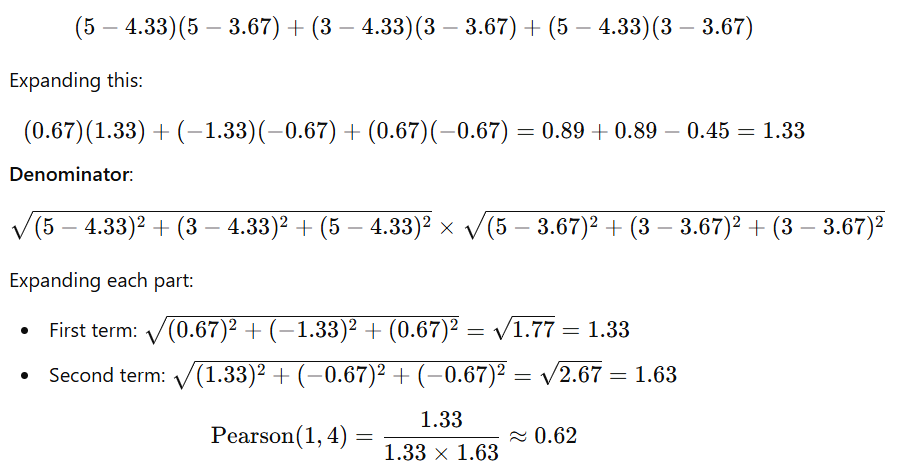


Pearson Correlation Between Item1 and Item4

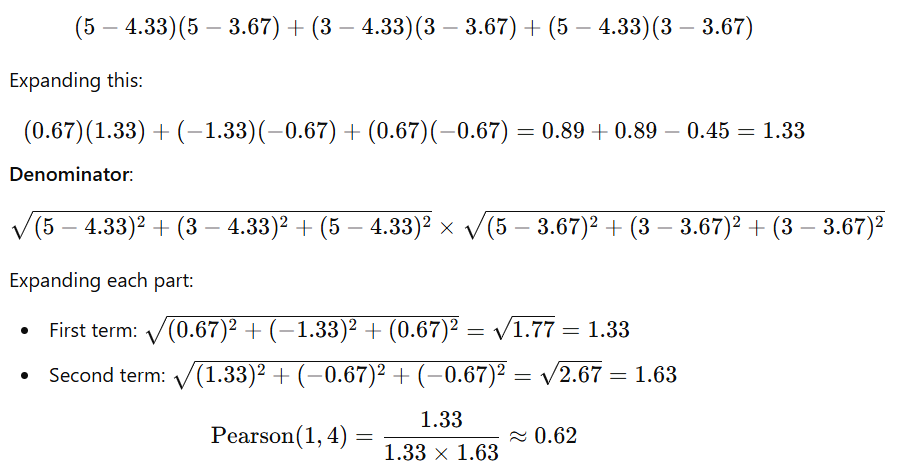


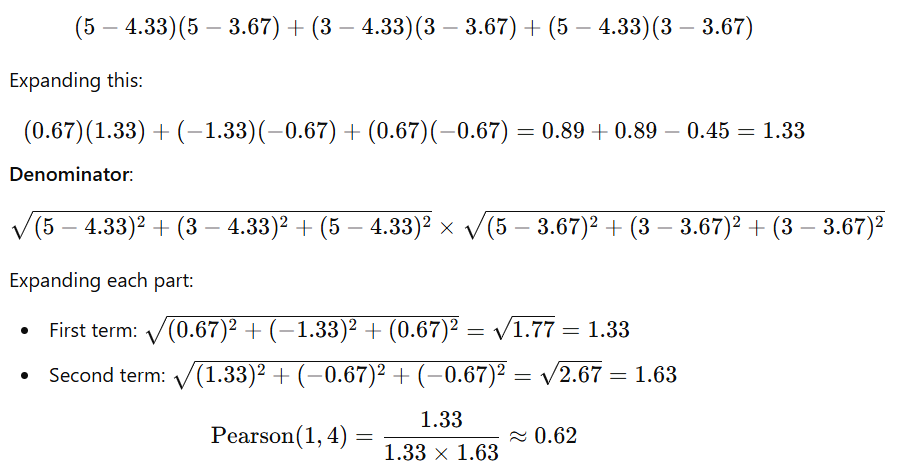
**Numerator**:

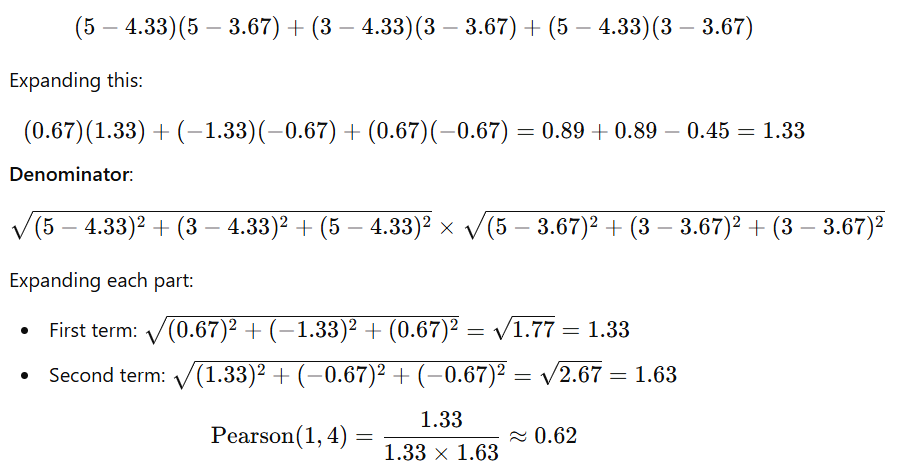


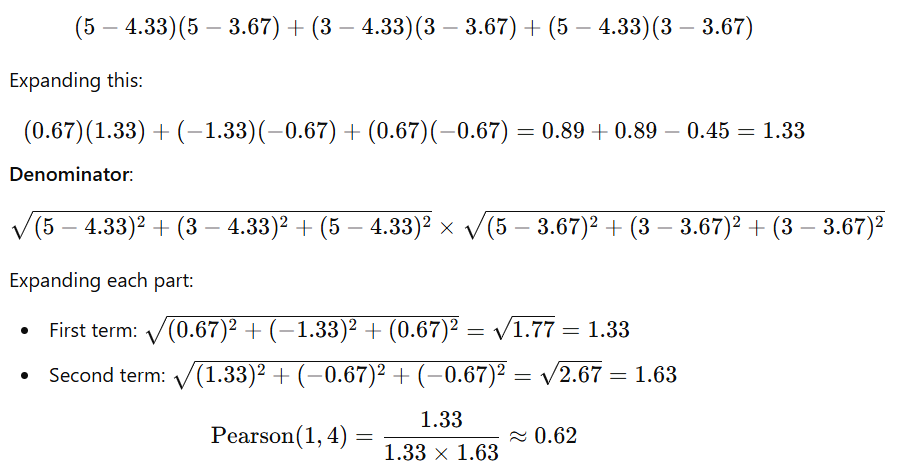


Denominator:

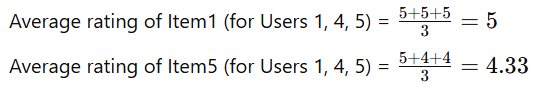








Pearson Correlation Between Item1 and Item5



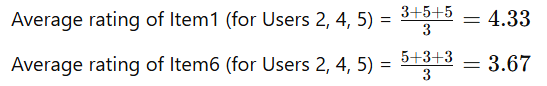
**Numerator**:



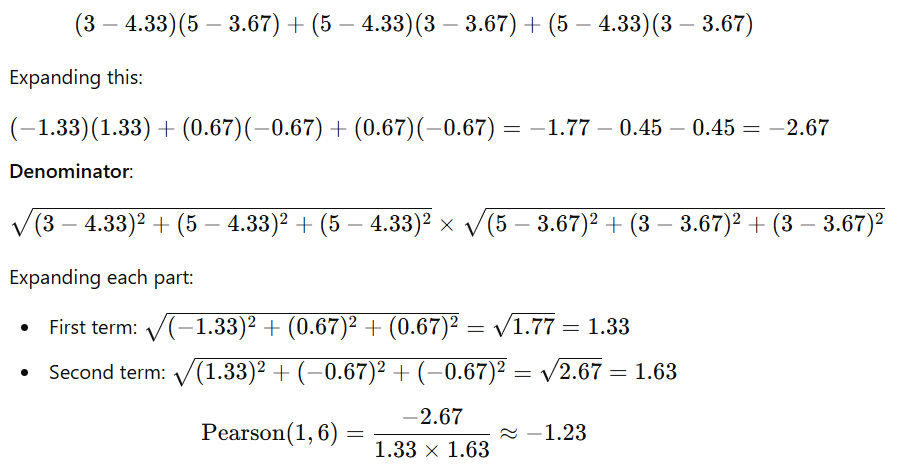
Denominator: Since all ratings for Item1 are identical (5), the standard deviation for Item1 is zero. This results in an undefined Pearson correlation due to division by zero.

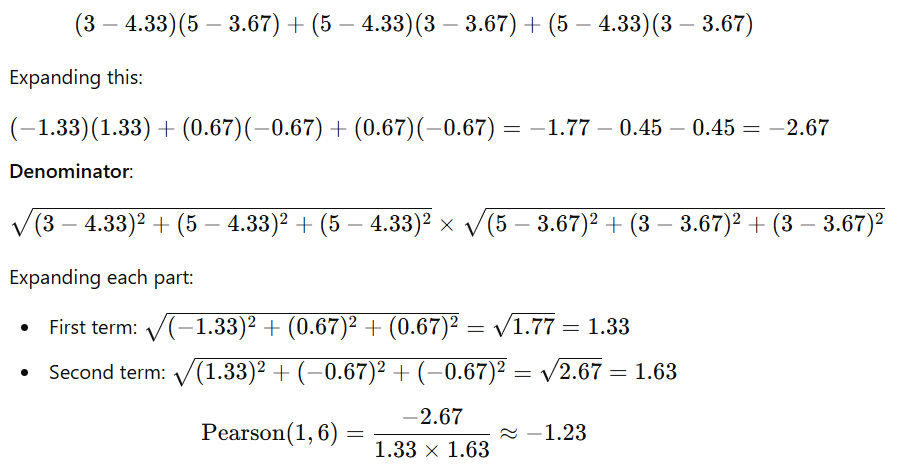
Thus, Pearson correlation for Item1 and Item5 is undefined.

Pearson Correlation Between Item1 and Item6

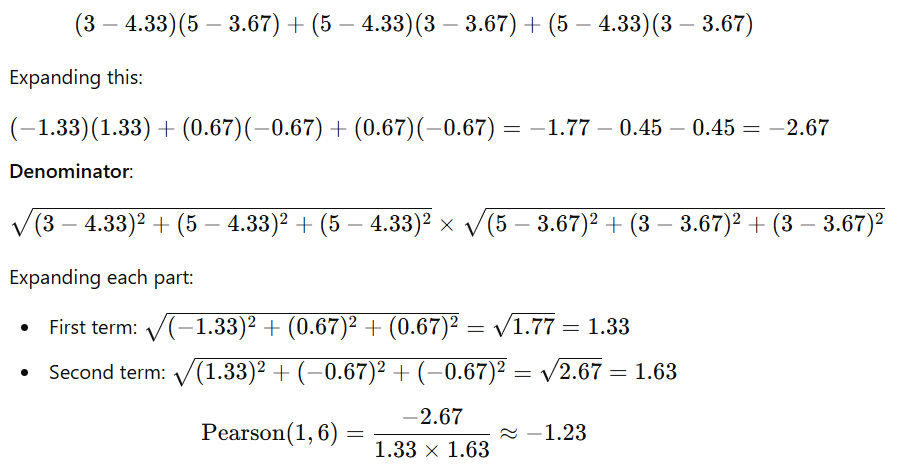


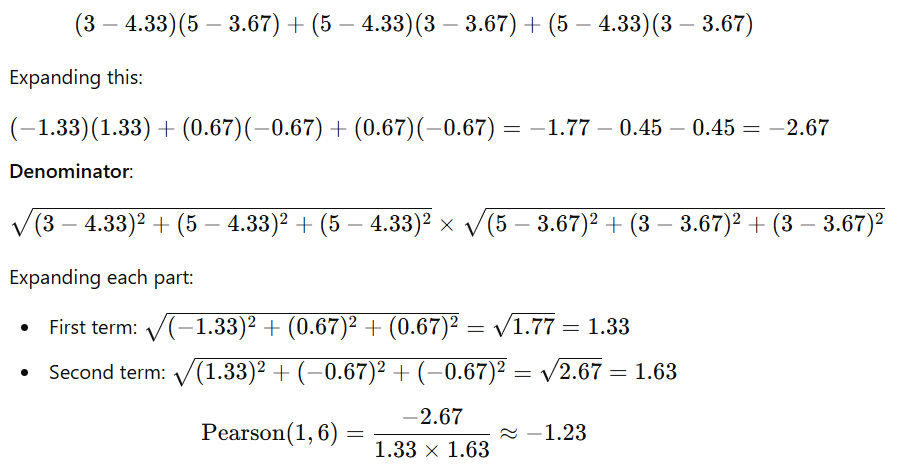
**Numerator**:

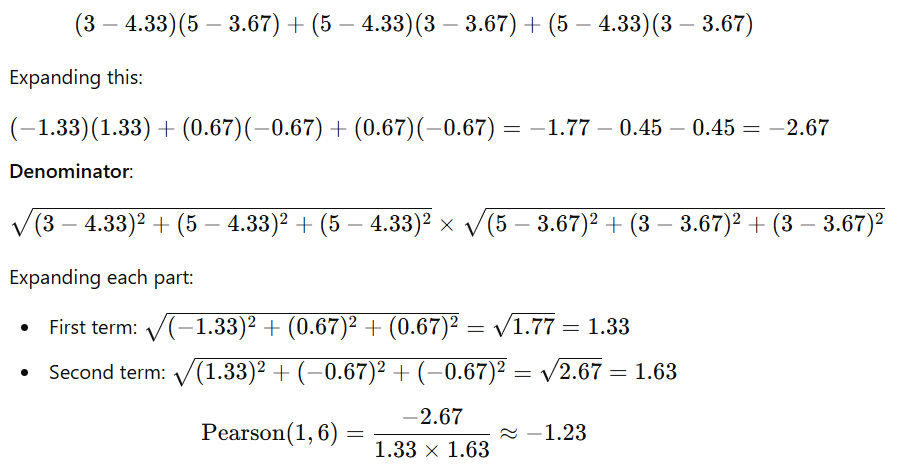




Denominator:







**Comparison and Evaluation**

### 1. Focus of Analysis

* **User-Based CF**:
  + **Focus**: This approach centers on analyzing user behavior and preferences. It looks for users who have similar rating patterns and bases recommendations on what similar users have liked.
  + **Similarity Calculation**: The similarities are calculated between users based on their ratings for various items. This means that if User A and User B rate items similarly, they are considered similar, regardless of the actual items involved.
* **Item-Based CF**:
  + **Focus**: This method focuses on the relationships between items rather than users. It identifies items that have been rated similarly by users and recommends items based on these relationships.
  + **Similarity Calculation**: Here, the similarities are computed between items, meaning that if Item X and Item Y receive similar ratings from users, they are deemed similar. The focus is on item characteristics rather than user behavior.

### 2. Handling of Ratings

* **User-Based CF**:
  + **Rating Patterns**: It is highly sensitive to individual user ratings and their tendencies. Variations in how users rate items (e.g., one user consistently rates higher than another) can significantly impact the recommendations generated.
  + **Cold Start Problem**: This approach struggles with new users who have no ratings, as there’s insufficient data to find similar users.
* **Item-Based CF**:
  + **Rating Aggregation**: It aggregates ratings based on item relationships and tends to smooth out the effects of individual user rating patterns. The recommendations are more stable as they rely on the overall interactions of users with items rather than on individual user differences.
  + **Less Susceptible to Cold Start**: It can handle new users better since it focuses on item characteristics, allowing for more immediate recommendations based on item popularity and similarity.

### 3. Computational Efficiency

* **User-Based CF**:
  + **Complexity**: As the user base grows, the computational load increases due to the need to compare each user with every other user. This can lead to scalability issues in large datasets.
  + **Dynamic Changes**: Frequent updates to user interactions require recalculating similarities, which can be computationally intensive.
* **Item-Based CF**:
  + **Efficiency**: Typically more efficient in handling large datasets because the number of items is often less than the number of users. The similarity calculations can be pre-computed and cached.
  + **Stability**: Item relationships tend to be more stable over time, which can reduce the need for frequent recalculations of similarities.

### 4. Diversity of Recommendations

* **User-Based CF**:
  + **Diversity**: Often yields more diverse recommendations since it can uncover niche preferences among users with unique tastes. Recommendations can include unexpected items that similar users have rated highly.
  + **Personalization**: Provides personalized recommendations tailored to individual user profiles, potentially leading to higher user satisfaction.
* **Item-Based CF**:
  + **Focused Recommendations**: Tends to produce recommendations that are closely aligned with items previously rated by the user. This can lead to less diversity if items are too similar.
  + **General Preferences**: While it provides solid recommendations, it might miss out on offering unique items that users may enjoy but haven't rated before.

### 5. Pearson Correlation vs. Cosine Similarity

* **Cosine Similarity**:
  + **User-Based CF**: Often gives more weight to the relative ratings of similar users. It is less affected by the overall rating scale of users, focusing instead on the direction of ratings (high vs. low).
  + **Item-Based CF**: Similarly, it captures the relationship between items based on user ratings without being influenced by overall rating trends.
* **Pearson Correlation**:
  + **User-Based CF**: Takes into account the mean rating of users, making it effective in identifying relationships that are linearly correlated. This can mitigate the impact of users who rate items consistently higher or lower than others.
  + **Item-Based CF**: Also accounts for mean ratings, focusing on linear relationships between items. This can be beneficial for identifying items that appeal to users who rate similarly across different items.

**Discussion**

* **User-Based CF**: Tends to be more impacted by individual user preferences, which can lead to fluctuations in recommendations due to outlier users. Cosine similarity performed well but was less effective in capturing nuanced relationships compared to Pearson.
* **Item-Based CF**: Generally yielded more consistent and stable recommendations. Both similarity measures provided effective results, but Pearson correlation was particularly effective in identifying meaningful item relationships.

### Conclusion

The comparison between user-based and item-based collaborative filtering (CF) strategies reveals distinct impacts on predicted accuracy, shaped by their underlying methodologies and the similarity measures employed—cosine similarity and Pearson correlation.

1. **User-Based Collaborative Filtering**:
   * This strategy demonstrated a high degree of personalization in predictions, as it relies on identifying similarities between users based on their rating behaviors. The accuracy of predicted ratings often reflects individual preferences, resulting in tailored recommendations that resonate with specific user tastes. However, this method can be sensitive to variations in user rating patterns, leading to fluctuations in predicted accuracy, particularly in cases of sparse data or when new users are introduced. The use of Pearson correlation enhances the reliability of predictions by normalizing ratings, which mitigates biases from differing rating scales, thereby improving accuracy.
2. **Item-Based Collaborative Filtering**:
   * In contrast, item-based CF often yielded more stable and consistent predictions. By focusing on the relationships between items rather than users, this strategy leverages the collective user ratings of items, providing a more comprehensive view of item popularity and similarity. This method typically results in higher predicted accuracy for items that are already well-rated by a larger number of users, as it aggregates broader user experiences. Cosine similarity allows item-based CF to highlight connections between items effectively, leading to reliable recommendations. Furthermore, the approach is less impacted by the cold start problem, as new users can still receive meaningful recommendations based on item interactions rather than needing to establish user relationships.

**References**

[IMDb: Ratings, Reviews, and Where to Watch the Best Movies & TV Shows](https://www.imdb.com/)

[IMDb Data Files Download](https://datasets.imdbws.com/)

<https://chatgpt.com/>