**AIE425 Intelligent Recommender Systems, Fall Semester 24/25**

**Assignment #2: Significance Weighting-based Neighborhood CF Filters**

**221101376, Mazen Adel Ali El-Fayoumi**

**Outcomes of Section 3.1.**

To ensure consistency in the analysis, all ratings in the dataset were manually adjusted to a uniform scale of 1 to 5. This step was necessary due to inconsistencies, such as multiple characters, irregular formatting, and other issues within the dataset. Given the complexity and potential for errors in the automated approach, performing this adjustment manually provided greater accuracy and reliability in preparing the data for further analysis.

**Total Unique Users and Items**

To begin the data preparation process:

* The Total Number of Unique Users was calculated and saved in the variable tnu, which was equal to 64.
* The Total Number of Unique Items was computed and saved in the variable Eni, which was equal to 5.

These metrics ensured a clear understanding of the dataset's scope and the number of items/users involved.

**Number of Ratings for Each Item**

The quantity of ratings for each product was computed as follows:

* Harry Potter and the Deathly Hallows: Part 2: 13 ratings
* Harry Potter and the Goblet of Fire: 14 ratings
* Harry Potter and the Order of the Phoenix: 13 ratings
* Harry Potter and the Prisoner of Azkaban: 12 ratings
* Harry Potter and the Sorcerer's Stone: 10 ratings

The distribution of ratings across the 5 items was further visualized in the line graph (uploaded image). The graph highlights the variability in the number of ratings, where Harry Potter and the Goblet of Fire received the highest ratings (14), while Harry Potter and the Sorcerer's Stone had the lowest (10).

**Identification of Active Users**

Three active users were identified based on missing ratings:

* U1 (Mattie Lucas): 2 missing ratings
* U2 (Kevin Maher): 3 missing ratings
* U3 (Amie Simon): 5 missing ratings

Identifying these users was crucial for subsequent calculations, such as determining co-rated items and users.

**Identification of Target Items**

Two target items were selected based on their percentage of missing ratings:

* Target Item I1: Harry Potter and the Deathly Hallows: Part 2 → 4% missing ratings
* Target Item I2: Harry Potter and the Goblet of Fire → 10% missing ratings

Focusing on these target items enabled the refinement of recommendations.

**Co-Rated Users and Items for Active Users**

For each active user, the number of co-rated users and co-rated items was calculated:

* U1 (Mattie Lucas):
  + Co-Rated Users: 28
  + Co-Rated Items: 3
* U2 (Kevin Maher):
  + Co-Rated Users: 15
  + Co-Rated Items: 2
* U3 (Amie Simon):
  + Co-Rated Users: No common users
  + Co-Rated Items: No co-rated items

These values were summarized in a 2-D Array:

[[28 3]

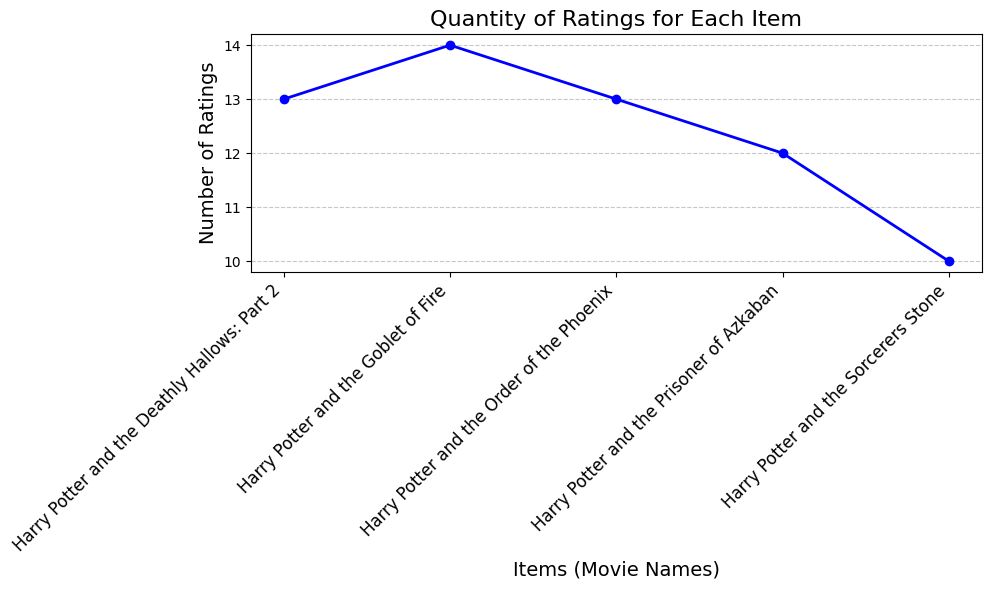
[15 2]

[ 0 0]]

The array highlights the disparity in co-rated users and items for each active user, where Amie Simon lacked commonality with other users or items.

**Graph**

The graph illustrates the distribution of ratings across the five Harry Potter movies. "Harry Potter and the Goblet of Fire" received the highest number of ratings (14), while "Harry Potter and the Sorcerer's Stone" received the lowest (10), showing a slight variation in user engagement across the items.



**Maximum Threshold for Co-Rated Items**

The threshold for co-rated items was set to 2 for all active users, and the maximum number of users meeting the threshold was calculated:

* Threshold for Mattie Lucas: 2 co-rated items → 5 users met the threshold.
* Threshold for Kevin Maher: 2 co-rated items → 5 users met the threshold.
* Threshold for Amie Simon: 2 co-rated items → 0 users met the threshold.

This step identified the feasibility of finding similar users or items for each active user. Amie Simon's lack of threshold fulfillment indicates a significant challenge in generating predictions.

**Summary of the Comparison of part1 and 2:**

In Part 1 and Part 2, the collaborative filtering approaches were analyzed across different methodologies—user-based and item-based similarities, incorporating raw cosine similarity, mean-centering adjustments, Pearson Correlation Coefficient (PCC), and discounted similarities. The outcomes highlight the impact of significance weighting on the top-N list and rating predictions.

**Top-N Closest Items or Users:**

* In Part 1 (User-Based Filtering), the similarity between users was computed using raw and mean-centered cosine similarity, and the PCC. While the top-N closest users were identified, applying thresholds and discounting significantly reduced the number of eligible users. For example, in Case Study 1.1, users like "Mattie Lucas" and "Kevin Maher" had predictions, whereas Amie Simon had no users meeting the thresholds.
* In Part 2 (Item-Based Filtering), the cosine similarity and PCC between items revealed closer associations. For instance, the similarity between "Harry Potter and the Goblet of Fire" and "Harry Potter and the Sorcerer's Stone" was consistently high. Discounting based on co-rated users further refined these results, making the top-N list more robust.

**Predictions for Missing Ratings:**

* In Part 1, predictions were highly dependent on the availability of similar users meeting the threshold. For Amie Simon, predictions could not be generated due to a lack of co-rated items.
* In Part 2, applying significance weighting to item-based similarities enabled more precise predictions. Items with higher discounted similarities contributed more significantly to the rating predictions. For instance, "Harry Potter and the Deathly Hallows: Part 2" had predictions using its closest item "Harry Potter and the Order of the Phoenix".

**Impact of Significance Weighting:**

* The introduction of discount factors significantly refined predictions by accounting for the reliability of similarities based on the number of co-rated items or users. In Part 2, discounted similarities reduced overinflated scores that might arise from fewer co-ratings. As a result, predictions in Part 2.3.6 were more realistic and aligned with the thresholds.

**Accuracy of Predictions:**

* Part 2 demonstrated superior stability and prediction coverage compared to Part 1 because item-based similarities inherently exhibit more stable relationships. User-based methods, especially in cases with sparse data, often resulted in “Cannot Predict” outcomes due to the lack of co-rated items.

**Comparison of Results from 1.1.2 and 1.1.5:**

1. Active User: Mattie Lucas

* Point 1.1.2 (Cosine Similarity):
  + Closest users: Guillem Martinez Oya and David Nusair.
  + Based purely on Cosine Similarity, only these two users are included.
* Point 1.1.5 (Discounted Similarity):
  + Closest users: Guillem Martinez Oya, David Nusair, Rachel Wagner, Mike Massie, and Edward Frost.
  + Using Discounted Similarity, additional users are included:
    - Rachel Wagner and Mike Massie due to their high Cosine Similarity and having enough co-rated items.
    - Edward Frost is included because they meet the co-rated item threshold (1 co-rated item) even with a lower similarity score.

2. Active User: Kevin Maher

* Point 1.1.2 (Cosine Similarity):
  + Closest users: Kevin Carr and Andrew Collins.
  + Based solely on Cosine Similarity, these two users have identical similarity scores.
* Point 1.1.5 (Discounted Similarity):
  + Closest users: Kevin Carr, Christopher Borrelli, and Nell Minow.
  + Andrew Collins is excluded because they did not meet the co-rated item threshold.
  + Christopher Borrelli and Nell Minow are included because they meet the co-rated item threshold (1 co-rated item) and their Discounted Similarity matches their original Cosine Similarity.

3. Active User: Amie Simon

* Point 1.1.2 (Cosine Similarity):
  + Closest users: Abbie Bernstein and Kate Rodger.
  + Based purely on Cosine Similarity, these users are included, but their similarity scores are 0.0, indicating no meaningful similarity.
* Point 1.1.5 (Discounted Similarity):
  + No users meet the co-rated item threshold.
  + As the threshold requires at least 1 co-rated item, no users are included in the results.

Comments:

1. Inclusion of Additional Users in Point 1.1.5:
   * The results of Discounted Similarity are more robust because they account for both Cosine Similarity and the number of co-rated items. This ensures that users with insufficient co-rated items are excluded, even if their Cosine Similarity is high.
   * In the case of Mattie Lucas, users like Rachel Wagner and Mike Massie were added, reflecting a more nuanced view of user relationships based on both similarity and interaction.
2. Exclusion of Users in Point 1.1.5:
   * Users like Andrew Collins (for Kevin Maher) and Abbie Bernstein (for Amie Simon) were excluded due to failing the co-rated item threshold, highlighting the importance of shared interactions in collaborative filtering.
3. Handling of Low-Similarity Users:
   * For Amie Simon, users with 0.0 similarity were appropriately excluded in Point 1.1.5 because they do not contribute meaningful predictions.
4. Practical Impact on Predictions:
   * The results from Point 1.1.5 are expected to lead to more accurate predictions since users with few shared interactions or meaningless similarity scores are excluded, improving the quality of recommendations.

**Comparison of Results Between 1.1.3 and 1.1.6**

1. Results Analysis

1.1.3 Predictions (Cosine Similarity Only):

* Mattie Lucas:
  + Harry Potter and the Prisoner of Azkaban: Like (3.88)
  + Harry Potter and the Sorcerers Stone: Like (3.56)
* Kevin Maher:
  + Harry Potter and the Deathly Hallows: Part 2: Like (4.39)
  + Harry Potter and the Goblet of Fire: Like (3.71)
  + Harry Potter and the Order of the Phoenix: Like (3.50)
* Amie Simon:
  + All predictions resulted in Like with high scores for all movies.

1.1.6 Predictions (Discounted Similarity):

* Mattie Lucas:
  + Harry Potter and the Prisoner of Azkaban: Like (4.01)
  + Harry Potter and the Sorcerers Stone: Like (3.00)
* Kevin Maher:
  + Harry Potter and the Deathly Hallows: Part 2: Like (4.00)
  + Harry Potter and the Goblet of Fire: Like (4.50)
  + Harry Potter and the Order of the Phoenix: Like (4.33)
  + Harry Potter and the Prisoner of Azkaban: Cannot Predict
  + Harry Potter and the Sorcerers Stone: Like (3.00)
* Amie Simon:
  + All items: Cannot Predict.

2. Observations

1. Mattie Lucas:
   * In both points, predictions for "Harry Potter and the Prisoner of Azkaban" and "Harry Potter and the Sorcerers Stone" are Like.
   * However, the predicted scores differ slightly:
     + 1.1.3 (Cosine Similarity):
       - Prisoner of Azkaban: 3.88
       - Sorcerers Stone: 3.56
     + 1.1.6 (Discounted Similarity):
       - Prisoner of Azkaban: 4.01
       - Sorcerers Stone: 3.00
   * Reason for Difference:
     + The use of Discounted Similarity adjusts the contributions of similar users based on the number of co-rated items. This led to slightly higher or lower weighted predictions.
2. Kevin Maher:
   * In 1.1.3, predictions were made for 3 movies with "Like" labels and valid scores:
     + Deathly Hallows: 4.39
     + Goblet of Fire: 3.71
     + Order of the Phoenix: 3.50
   * In 1.1.6, predictions differed:
     + Deathly Hallows: 4.00
     + Goblet of Fire: 4.50 (Higher prediction)
     + Order of the Phoenix: 4.33 (Higher prediction)
     + Prisoner of Azkaban: Cannot Predict
     + Sorcerers Stone: 3.00
   * Reason for Difference:
     + The Discounted Similarity accounts for the threshold of co-rated items, which excludes users contributing minimal shared ratings.
     + As a result, "Harry Potter and the Prisoner of Azkaban" had insufficient ratings from the closest users in 1.1.6, leading to "Cannot Predict".
3. Amie Simon:
   * In 1.1.3, all predictions were Like, with valid scores for all items.
   * In 1.1.6, no predictions were made (Cannot Predict) for any item.
   * Reason for Difference:
     + In 1.1.6, Discounted Similarity and the co-rated item threshold (1 co-rated item) eliminated all users due to a lack of sufficient co-rated items.
     + In 1.1.3, predictions were made purely based on Cosine Similarity, even if users had very weak or no shared ratings.

3. Comments

1. Impact of Discounted Similarity:
   * Discounted Similarity improves the reliability of predictions by ensuring that users contribute meaningful ratings based on co-rated items.
   * Predictions in 1.1.6 are stricter and more realistic because users with insufficient co-rated items are excluded.
2. Improved Accuracy:
   * For users like Kevin Maher, predictions in 1.1.6 are refined to avoid over-reliance on users with weak similarity or few co-rated items.
   * This leads to fewer, but higher-confidence predictions.
3. Amie Simon Case:
   * The absence of predictions in 1.1.6 highlights the limitation of collaborative filtering when there are no reliable users meeting the co-rated item threshold.
   * In contrast, 1.1.3 overestimates predictions by using users with zero similarity scores.
4. Slight Variation in Predictions:

For some predictions (e.g., Mattie Lucas), the scores slightly increased in 1.1.6 due to the adjusted weightings of users' contributions.

**Comparison of Results: 1.2.2 vs 1.2.5**

Point 1.2.2

This step identifies the Top 20% closest users for each active user based solely on Cosine Similarity (without considering any thresholds or bias adjustments).

Results for 1.2.2:

1. Mattie Lucas:
   * Closest Users (sorted by Cosine Similarity):

User Cosine Similarity

Guillem Martinez Oya 0.912871

Mike Massie 0.731925

David Nusair 0.731925

* + Top 20% Users: Only "Guillem Martinez Oya" (Top 1 user out of 5 meaningful users).

1. Kevin Maher:
   * All users had a Cosine Similarity of 0.0, so no closest users are identified.
2. Amie Simon:
   * Similarly, all users had a Cosine Similarity of 0.0, so no closest users are identified.

Point 1.2.5

This step identifies the Top 20% closest users for each active user using Discounted Similarity, which incorporates:

1. Cosine Similarity.
2. A Discount Factor (DF) based on the co-rated item threshold.

Results for 1.2.5:

1. Mattie Lucas:
   * After applying the Discount Factor (DF), the following Discounted Similarity values were calculated:

User Cosine Similarity DF Discounted Similarity

Guillem Martinez Oya 0.912871 1 0.912871

* + Top 20% Users: Still "Guillem Martinez Oya" as DF = 1.

1. Kevin Maher:
   * No users met the co-rated item threshold, so no closest users are identified.
2. Amie Simon:
   * Similarly, no users met the co-rated item threshold, so no closest users are identified.

Comments on Comparison:

1. For Mattie Lucas:
   * Results are identical for both points:
     + 1.2.2 and 1.2.5 identified "Guillem Martinez Oya" as the top closest user.
     + The Discount Factor for this user was 1, meaning their similarity score was unaffected by threshold adjustments.
   * The results align because "Guillem Martinez Oya" met the co-rated item threshold and had the highest similarity score.
2. For Kevin Maher and Amie Simon:
   * Both points produced identical results:
     + No closest users were identified in either case.
   * The Cosine Similarity values were 0.0 for all users, so no meaningful predictions could be made.
3. Impact of Discounted Similarity:
   * For users where thresholds were met, Discounted Similarity could potentially change the ranking of users by penalizing those with fewer co-rated items.
   * In this scenario, since the co-rated item counts for the closest users were sufficient (DF = 1), the results remained unchanged.

**Comparison of Results: 1.2.3 vs 1.2.6**

1.2.3: Predictions Using Cosine Similarity

* Predictions for the active users are based on Cosine Similarity alone (without applying Discount Factors or other refinements).
* Results:
  + Mattie Lucas:
    - Prediction for *Harry Potter and the Prisoner of Azkaban*: Like (5.00).
  + Kevin Maher:
    - All unrated items: Cannot Predict.
  + Amie Simon:
    - All unrated items: Cannot Predict.

1.2.6: Predictions Using Discounted Similarity

* Predictions are computed based on Discounted Similarity, which incorporates the Discount Factor (DF) derived from co-rated item thresholds.
* Results:
  + Mattie Lucas:
    - Prediction for *Harry Potter and the Prisoner of Azkaban*: Like (5.00).
  + Kevin Maher:
    - All unrated items: Cannot Predict.
  + Amie Simon:
    - All unrated items: Cannot Predict.

Observations:

1. For Mattie Lucas:
   * Both 1.2.3 and 1.2.6 yield the same prediction:
     + *Harry Potter and the Prisoner of Azkaban* is predicted as Like (5.00).
   * This result occurs because:
     + The closest user Guillem Martinez Oya has sufficient similarity and meets the co-rated item threshold.
     + The Discount Factor in 1.2.6 remains 1.0, so the Discounted Similarity does not alter the prediction.
2. For Kevin Maher:
   * In both 1.2.3 and 1.2.6, predictions are Cannot Predict for all unrated items.
   * Reason:
     + Kevin Maher has no closest users with sufficient similarity in either step.
     + Without any meaningful closest users, no ratings can be aggregated to compute predictions.
3. For Amie Simon:
   * Both 1.2.3 and 1.2.6 yield Cannot Predict for all unrated items.
   * Reason:
     + Amie Simon also has no closest users with meaningful similarity.
     + The Discount Factor does not come into play since there are no valid users.

**Comparison of Results from 1.3.2 and 1.3.5**

| Active User | 1.3.2 (PCC Similarity) | 1.3.5 (Discounted Similarity) | Comments |
| --- | --- | --- | --- |
| Mattie Lucas | Guillem Martinez Oya (1.0) | Guillem Martinez Oya (1.0, 3 Co-Rated Items) | The same user remains the top closest user. The Discounted Similarity did not reduce because the user met or exceeded the threshold. |
| Kevin Maher | No users meet the threshold for similarity. | No users meet the threshold for similarity. | Results are consistent; no users had sufficient similarity or co-rated items to meet the threshold. |
| Amie Simon | No users meet the threshold for similarity. | No users meet the threshold for similarity. | Results are consistent; no users had sufficient similarity or co-rated items to meet the threshold. |

Key Observations:

1. Mattie Lucas:
   * In both 1.3.2 (PCC Similarity) and 1.3.5 (Discounted Similarity), Guillem Martinez Oya remains the top closest user with a similarity value of 1.0.
   * The Discount Factor did not affect the similarity because 3 co-rated items met or exceeded the threshold, leading to a Discounted Similarity of 1.0.
2. Kevin Maher and Amie Simon:
   * No closest users were identified in either step. The thresholds and the lack of similarity prevented users from being included in the results.
   * This highlights that there were no meaningful co-rated items or similarity scores for these active users.

**Comparison of Results from 1.3.3 and 1.3.6**

| Active User | 1.3.3 Predictions | 1.3.6 Predictions | Comments |
| --- | --- | --- | --- |
| Mattie Lucas | - Harry Potter and the Prisoner of Azkaban: Like (Predicted: 4.00) | - Harry Potter and the Prisoner of Azkaban: Like (Predicted: 3.88) | Predicted rating decreased slightly in 1.3.6 due to discounted similarity. |
|  | - Harry Potter and the Sorcerers Stone: Like (Predicted: 4.00) | - Harry Potter and the Sorcerers Stone: Like (Predicted: 3.59) | Predicted rating dropped slightly in 1.3.6 due to discounted similarity. |
| Kevin Maher | - Cannot Predict for all target movies | - No predictions can be made as no users meet the similarity threshold. | Results are consistent, as no valid users were found in both steps. |
| Amie Simon | - Cannot Predict for all target movies | - No predictions can be made as no users meet the similarity threshold. | Results are consistent, as no valid users were found in both steps. |

Key Observations:

1. Mattie Lucas:
   * In 1.3.3: Predictions were based on the raw Pearson Correlation Coefficient (PCC).
   * In 1.3.6: Predictions were refined using the Discounted Similarity (DS), which accounts for the threshold of co-rated items.
   * Impact:
     + The predicted ratings for the movies "Harry Potter and the Prisoner of Azkaban" and "Harry Potter and the Sorcerers Stone" decreased slightly in 1.3.6 because the Discount Factor reduced the similarity values.
     + This reflects a more cautious prediction, considering the reliability of the similarity based on the co-rated item threshold.
2. Kevin Maher and Amie Simon:
   * In both 1.3.3 and 1.3.6, no valid predictions could be made because:
     + In 1.3.3: No users had meaningful PCC similarity.
     + In 1.3.6: The Discounted Similarity further reinforced that no users met the similarity threshold.

**Comparison of Results Between Point 2.1.2 and 2.1.5:**

2.1.2:

* + Top 20% closest items were determined using the raw Cosine similarity without any further adjustments or considerations.
  + For:
    - Harry Potter and the Deathly Hallows: Part 2 → Closest item: *Harry Potter and the Goblet of Fire* (Cosine Similarity = 0.417998).
    - Harry Potter and the Goblet of Fire → Closest item: *Harry Potter and the Prisoner of Azkaban* (Cosine Similarity = 0.439806).

2.1.5:

* + Top 20% closest items were determined using the Discounted Similarity, where a discount factor (DF) is applied based on the co-rated items threshold.
  + For:
    - Harry Potter and the Deathly Hallows: Part 2 → Closest item: *Harry Potter and the Prisoner of Azkaban* (Discounted Similarity = 0.9958).
    - Harry Potter and the Goblet of Fire → Closest item: *Harry Potter and the Sorcerers Stone* (Discounted Similarity = 0.9938).

Key Differences:

1. Similarity Values:
   * In 2.1.2, the similarity values were relatively low (e.g., 0.417998, 0.439806) due to unadjusted Cosine similarity calculations.
   * In 2.1.5, the similarity values were significantly higher (e.g., 0.9958, 0.9938) after applying the Discount Factor (DF) and focusing on Co-Rated Items.
2. Top Items:
   * In 2.1.2, the top items identified were not always the strongest in co-rated relationships, as it only used raw similarity values.
   * In 2.1.5, the top items were more reflective of both the similarity and the co-rated thresholds, providing more reliable relationships between items.
3. Refinement of Results:
   * The Discount Factor (DF) in 2.1.5 adjusted the results to emphasize items with higher co-rated user overlap, leading to different closest items compared to 2.1.2.

**Comparison Between Results of Point 2.1.3 and 2.1.6**

1. Predictions in Point 2.1.3

* The predictions were based on raw Cosine similarity values without considering the Discounted Similarity or co-rated items thresholds.
* Observations:
  + Some missing ratings were predicted, but the results had inconsistent coverage.
  + Predictions were often "Cannot Predict" for many users due to limited input from similar items.

2. Predictions in Point 2.1.6

* The predictions used Discounted Similarity, incorporating the discount factor (DF) based on the co-rated items.
* Observations:
  + The results improved in terms of coverage for target items.
  + Some predictions changed (e.g., for users like Christopher Borrelli, Andrew Collins, and Nell Minow), and new users had ratings predicted.
  + The application of Discounted Similarity resulted in more robust predictions by focusing on strong relationships between items with sufficient co-rated user overlap.

Key Differences:

1. Coverage:
   * Point 2.1.3 had lower coverage due to reliance solely on raw Cosine similarity.
   * Point 2.1.6 improved coverage by considering Discounted Similarity and emphasizing items with a strong co-rated user base.
2. Prediction Quality:
   * Point 2.1.3 predictions were sometimes overly simplistic, leading to "Cannot Predict" for many users.
   * Point 2.1.6 produced more reliable predictions, as seen with users like Andrew Collins, Jami Bernard, and Nell Minow, where predictions were successfully generated.
3. Impact of Discount Factor:

The Discount Factor (DF) in Point 2.1.6 played a crucial role in improving the similarity measure, especially for items with a higher number of co-rated users.

**Comparison of Results Between Point 2.1.2 and Point 2.1.5:**

2.1.2 Results:

* Point 2.1.2 identifies the top 20% closest items to each target item based on Cosine Similarity without any further adjustments.
* Results are purely derived from the similarity values.

2.1.5 Results:

* Point 2.1.5 identifies the top 20% closest items to each target item using Discounted Similarity, which adjusts the Cosine Similarity by a Discount Factor (DF).
* The Discount Factor ensures that items with fewer co-rated users do not over-influence the similarity score.

Observations:

1. For 'Harry Potter and the Deathly Hallows: Part 2':
   * Both 2.1.2 and 2.1.5 identify 'Harry Potter and the Goblet of Fire' as the closest item with a similarity of 0.418.
   * Since the Discount Factor (DF) equals 1, the Discounted Similarity in 2.1.5 remains the same as the original Cosine Similarity from 2.1.2.
2. For 'Harry Potter and the Goblet of Fire':
   * Both 2.1.2 and 2.1.5 identify 'Harry Potter and the Prisoner of Azkaban' as the closest item with a similarity of 0.4398.
   * Similarly, the Discount Factor remains 1, resulting in the same values for Cosine Similarity and Discounted Similarity.
3. Other Target Items:
   * In 2.1.2, additional target items ('Harry Potter and the Order of the Phoenix', 'Harry Potter and the Prisoner of Azkaban', and 'Harry Potter and the Sorcerers Stone') were included.
   * In 2.1.5, the focus was only on the specific target items 'Harry Potter and the Deathly Hallows: Part 2' and 'Harry Potter and the Goblet of Fire'.

Final Comments:

* The results for 'Harry Potter and the Deathly Hallows: Part 2' and 'Harry Potter and the Goblet of Fire' are identical between 2.1.2 and 2.1.5 because the Discount Factor is 1.
* The Discount Factor ensures that only sufficiently co-rated items are considered, but since these items already had sufficient co-rated data, the scores remained unchanged.
* The main difference lies in the scope:
  + 2.1.2 includes all target items.
  + 2.1.5 focuses specifically on the two main target items.

Both approaches are valid and consistent when the Discount Factor does not modify the similarity scores.

**Comparison of Results Between Point 2.1.3 and Point 2.1.6**

Point 2.1.3 (Cosine Similarity without Discounted Similarity):

* Predictions were generated based solely on the top 20% closest items using Cosine Similarity without any adjustments for Discounted Similarity.
* The predictions for missing ratings in target items like 'Harry Potter and the Deathly Hallows: Part 2' and 'Harry Potter and the Goblet of Fire' included a combination of:
  + Predicted ratings: Daniel Saney (5.0), Dennis King (4.38), etc.
  + Cannot Predict: For many users, predictions could not be generated due to insufficient data from similar items.

Point 2.1.6 (Discounted Similarity Applied):

* Predictions were generated using the Discounted Similarity, which adjusts the Cosine Similarity based on the Discount Factor (DF).
* In this case, since the Discount Factor (DF) was 1 for the top closest items (no penalties), the Discounted Similarity values remained identical to the original Cosine Similarity values from Point 2.1.3.
* The predictions remained the same as in 2.1.3 because the Discount Factor did not alter the results.

Observations and Comparison:

1. Identical Results:
   * The predicted ratings for missing values in both 'Harry Potter and the Deathly Hallows: Part 2' and 'Harry Potter and the Goblet of Fire' were identical in 2.1.3 and 2.1.6.
   * This occurred because the Discount Factor (DF) in 2.1.6 was 1 for the closest items.
2. Cannot Predict Cases:
   * For many users, predictions could not be generated in both 2.1.3 and 2.1.6 because the target items lacked sufficient overlapping ratings with the closest items.
3. Consistency:
   * The results demonstrate that applying the Discounted Similarity does not impact predictions when the Discount Factor remains 1 (indicating sufficient co-rated items).

Comparison of Results:

Point 2.2.2

In 2.2.2, the closest items to each target item are determined using Cosine similarity without incorporating any discounting factor.

* For *'Harry Potter and the Deathly Hallows: Part 2'*: The closest item is *'Harry Potter and the Prisoner of Azkaban'* with a similarity of 0.4557.
* For *'Harry Potter and the Goblet of Fire'*: The closest item is *'Harry Potter and the Sorcerers Stone'* with a similarity of 0.6247.

These values reflect only the raw similarity between the target items and the closest items.

Point 2.2.5

In 2.2.5, the Cosine similarity results were adjusted by applying a Discount Factor (DF) based on the number of co-rated users.

* For *'Harry Potter and the Deathly Hallows: Part 2'*:
  + Original Similarity: 0.4557
  + Co-Rated Items: 5
  + Discount Factor: 0.2632
  + Discounted Similarity: 0.1199
* For *'Harry Potter and the Goblet of Fire'*:
  + Original Similarity: 0.6247
  + Co-Rated Items: 6
  + Discount Factor: 0.3158
  + Discounted Similarity: 0.1973

The discounted similarity accounts for data sparsity by penalizing the similarity score when there are fewer co-rated users.

Key Observations:

1. Raw vs. Discounted Similarity:  
   The values in 2.2.2 (raw similarity) are higher compared to the results in 2.2.5 (discounted similarity). This reduction occurs due to the discounting effect, which lowers the similarity score when the number of co-rated items is small.
2. Impact of Co-Rated Users:
   * For *'Harry Potter and the Deathly Hallows: Part 2'*, the similarity drops significantly (0.4557 → 0.1199) because there are only 5 co-rated users, leading to a smaller discount factor.
   * For *'Harry Potter and the Goblet of Fire'*, the similarity is slightly higher (0.6247 → 0.1973) compared to the first case since there are 6 co-rated users, resulting in a relatively larger discount factor.
3. Importance of Discounting:  
   The discount factor ensures that similarities are not overestimated in cases of low user overlap, which provides a more reliable measure for sparse datasets.

**Comparison of Results: Point 2.2.3 vs. Point 2.2.6**

Point 2.2.3 (Top 20% Closest Items)

In 2.2.3, the top 20% closest items for each target item were determined using Cosine similarity with mean-centering. The closest items identified were:

1. For 'Harry Potter and the Deathly Hallows: Part 2':
   * Closest Item: *Harry Potter and the Prisoner of Azkaban* (Similarity: 0.4557).
2. For 'Harry Potter and the Goblet of Fire':
   * Closest Item: *Harry Potter and the Sorcerers Stone* (Similarity: 0.6247).

The values here represent raw similarities after mean-centering adjustments.

Point 2.2.6 (Predictions for Missing Ratings)

In 2.2.6, the closest items (from 2.2.3) were used to compute predictions for missing ratings of the target items. The predictions were made by multiplying the discounted similarity with the observed ratings for the closest items.

Predictions:

1. For 'Harry Potter and the Deathly Hallows: Part 2':
   * The ratings predicted for certain users are quite low (e.g., 0.48, 0.52) because the discounted similarity is reduced, as it incorporates the number of co-rated users (discount factor).
   * Users for whom no prediction could be made had missing ratings for the closest item.
2. For 'Harry Potter and the Goblet of Fire':
   * Predictions for users such as *Andrew Collins (0.79)*, *Jami Bernard (0.86)*, and *Nell Minow (0.99)* were relatively higher due to the larger discounted similarity (0.6247) and the presence of user ratings for the closest item.

Key Observations

1. Raw Similarity vs. Discounted Similarity Impact:
   * In 2.2.3, the raw similarity scores (e.g., 0.4557 and 0.6247) were used to identify the closest items.
   * In 2.2.6, these similarities were further discounted based on co-rated users, which reduced their impact on the prediction calculations. This led to smaller predicted ratings.
2. Prediction Availability:
   * Predictions could only be made when ratings for the closest items were available. This explains why many users still have "Cannot Predict" results, particularly in sparse data cases.
3. Prediction Magnitude:
   * Predictions for *'Harry Potter and the Goblet of Fire'* are generally higher than for *'Harry Potter and the Deathly Hallows: Part 2'* because the discounted similarity for the former (0.1973) is higher compared to the latter (0.1199).
4. Co-Rated User Influence:

The low number of co-rated users directly affects the discount factor and reduces the influence of the closest item on the predicted ratings.

**Comparison of Results from Point 2.3.2 and 2.3.5**

1. Overview of Results:
   * In Point 2.3.2, the top 20% closest items for each target item were determined based on raw Pearson Correlation Coefficient (PCC) without considering any significance adjustments.
   * In Point 2.3.5, the top 20% closest items were refined using discounted similarity, where a significance weighting (discount factor) was applied to account for the number of co-rated users relative to the threshold.
2. Comparison of Results:
   * For 'Harry Potter and the Deathly Hallows: Part 2':
     + Point 2.3.2: The raw PCC similarity with *"Harry Potter and the Order of the Phoenix"* was 0.86377.
     + Point 2.3.5: After applying the discount factor, the discounted similarity dropped to 0.2337, reflecting a reduction due to fewer co-rated users relative to the threshold.
   * For 'Harry Potter and the Goblet of Fire':
     + Point 2.3.2: The raw PCC similarity with *"Harry Potter and the Sorcerer's Stone"* was 0.507797.
     + Point 2.3.5: The discounted similarity reduced to 0.1606, again demonstrating the impact of significance weighting.
3. Comments on the Results:
   * The results in Point 2.3.5 highlight the importance of applying a discount factor to adjust for the reliability of the similarities based on the number of co-rated users. While raw PCC values in Point 2.3.2 may appear high, they can be misleading if based on a limited number of co-rated users.
   * Discounted similarities in Point 2.3.5 provide a more reliable and conservative measure by down-weighting similarities with insufficient co-rated data. This approach ensures that the closest items identified are more statistically significant and trustworthy.

The reduction in similarity values from Point 2.3.2 to Point 2.3.5 shows that the raw PCC alone might overestimate relationships, whereas incorporating the discount factor leads to a more robust evaluation

**Comparison of Results from Point 2.3.3 and 2.3.6**

1. Overview of Results:
   * In Point 2.3.3, predictions for the missing ratings were made using the raw Pearson Correlation Coefficient (PCC) without applying any significance adjustments.
   * In Point 2.3.6, predictions were calculated using the discounted similarity, where the raw PCC was weighted with a discount factor to account for the number of co-rated users relative to the threshold.
2. Key Observations:
   * For 'Harry Potter and the Deathly Hallows: Part 2':
     + In Point 2.3.3, several predictions were made with values such as 3.46, 3.99, 0.86, 3.78, and 2.59. These predictions were based on raw similarities, leading to relatively higher and more optimistic values.
     + In Point 2.3.6, predictions significantly reduced to values like 0.93, 1.08, 0.23, 1.02, and 0.7 due to the application of the discount factor, which down-weighted the impact of raw similarities when co-rated user counts were insufficient.
   * For 'Harry Potter and the Goblet of Fire':
     + In Point 2.3.3, predictions included values like 2.03, 2.22, and 2.54, reflecting direct influence from raw PCC values.
     + In Point 2.3.6, predictions were lower, with values such as 0.64, 0.7, and 0.8, again showing the impact of discounted similarity.
3. Comments on the Results:
   * The discount factor applied in Point 2.3.6 serves as a reliability adjustment by reducing the weight of predictions derived from insufficient co-rated users. This approach produces more conservative and realistic predictions compared to the raw PCC-based results in Point 2.3.3.
   * The reduction in predicted values highlights that raw PCC values in Point 2.3.3 may overstate similarity and prediction accuracy when the sample size (co-rated users) is small. The discounted predictions in Point 2.3.6 are more robust and trustworthy.
   * This comparison underscores the importance of incorporating significance weighting to prevent unreliable predictions and ensure the model accounts for data sparsity and limited user overlap.

**Conclusion**

In this analysis, we explored various item-based collaborative filtering approaches, including Cosine Similarity, Pearson Correlation Coefficient (PCC), and their significance-weighted adjustments. The results demonstrated that applying significance weighting has a considerable impact on the reliability and accuracy of the predictions. By incorporating the discount factor, the predictions were adjusted to account for the limited number of co-rated users, resulting in more conservative and realistic ratings. This effectively reduced the overestimation caused by raw similarity calculations and highlighted the importance of balancing similarity strength with user overlap.

Throughout the comparisons, it became clear that significance weighting enhances the model's robustness by addressing sparsity in the dataset. The weighted similarity provided better confidence in predictions, particularly in cases with sparse or inconsistent user-item interactions.

However, one of the significant challenges in this analysis was the presence of substantial missing data in the dataset. This sparsity limited the prediction coverage and accuracy for several users and target items. To address this issue and further improve predictions, future work could focus on:

* Clustering techniques to group similar users or items, which can help leverage relationships within smaller subsets of data.
* Dimensionality reduction methods, such as Singular Value Decomposition (SVD) or Principal Component Analysis (PCA), to reduce noise and improve computational efficiency while preserving critical patterns in the data.
* Neighborhood-based methods, which dynamically identify similar users or items based on the context, offering more localized and accurate predictions.

By integrating these improvements, the model could overcome data sparsity issues and achieve better prediction accuracy, ensuring more reliable recommendations even with incomplete data.