פתרון שאלה 4 - ליה גולדינג

על פי רוב אנחנו מניחים סימטריה, כלומר ההמלצה על AB תהיה באותו חוזק כמו ההמלצה BA היצוג AB פירוש המלצה על B למי שאהב את A

1. כיצד ניתן לבדוק באיזו רמה סימטריה אכן מתקיימת?

נוכל לבדוק באיזו רמה מתקיימת סימטריה, על ידי השוואה בין הציון שיקבל הסרט A על בסיס חיבה ל- B ונשווה לציון של B על בסיס חיבה לA. הפער מהווה מדד לסימטריה, ככל שהפער קטן רמת הסימטריה גבוהה. רמת הסימטריה של ההמלצות היא 0.54, כאשר 0 היא סימטריה מושלמת ו- 1 מהווה אסימטריה מושלמת.

2. רוב המודלים שלנו סימטרים. חישבו על פירוש ל"אנשים שאהבו את A גם אהבו את B "באופן שיאפשר מודל לא סימטרי.

מכיוון שלא תמיד תהיה סימטריה בין חיבה לסרטים, כלומר שאם סרט A הוא סרט קצר ונישתי בסגנון מסויים וסרט B הוא סרט גדול, מוכר ועתיר תקציב באותו הז'אנר. סביר שכל מי שצפה בA צפה גם בB ואהב אותו אבל לא כל מי שצפה בB צפה או אהב את A.

דוגמה נוספת היא סרט אהוב ולו סרט המשך פחות מוצלח, סביר שמי שבחר לצפות בסרט ההמשך ראה ואהב את הסרט הראשון, אך לא בהכרח להפך.

מהסיבה הזו יש הרבה יותר דיוק במודל שאינו סימטרי, רוב המודלים שלנו מחשבים מרחק באופן סימטרי - חישוב מרחק בין שתי נקודות ללא משמעות ל"נקודת המוצא". כדי לחשב מודל שאינו סימטרי צריך לקחת בחשבון את גודל קהל האוהדים של הסרט, כלומר לתת משמעות לכמה הסרט מוכר ואהוב כדי להמנע מתוצאה מוטית שעלולה להתקבל המקרה המתואר כאן.

פירוש בר מימוש שיאפשר מודל אסימטרי יהיה - "כמה מתוך האנשים שאהבו את B אהבו גם את A?" כלומר הסתברות מותנית- כמה חיתוך (כאלו שאהבו גם וגם) ישנם מתוך B : $\frac{A \cap B}{B}$

3. ממשו את המודל

```
CREATE INDEX idx_pmr_suggested_by ON personal_movies_ranking(suggested_by);
#לייעול ההרצה
DROP TABLE IF EXISTS movie_user_counts;
CREATE TABLE movie_user_counts AS
SELECT movie_id, COUNT(*) AS total_source_users
FROM personal_movies_ranking
GROUP BY movie_id;


DROP TABLE IF EXISTS shared_movie_voters;
CREATE TABLE shared_movie_voters AS
SELECT
   r1.movie_id AS movie_id_source,
   r2.movie_id AS movie_id_target,
   COUNT(*) AS shared_users_count
FROM personal_movies_ranking r1
JOIN personal_movies_ranking r2
   ON r1.suggested_by = r2.suggested_by
WHERE r1.movie_id <> r2.movie_id
GROUP BY r1.movie_id, r2.movie_id
HAVING COUNT(*) >= 2 ;     #סף שרירותי להקטנת הטבלה וייעול ההרצה


DROP TABLE IF EXISTS asymmetric_preference;
CREATE TABLE asymmetric_preference AS
SELECT
   sv.movie_id_source,
   sv.movie_id_target,
   (sv.shared_users_count * 1.0) / NULLIF(smc.total_source_users, 0) AS model_score
FROM shared_movie_voters sv
JOIN movie_user_counts smc
   ON sv.movie_id_source = smc.movie_id
WHERE sv.shared_users_count >= 2;
```

#ההגדרה של 2 ומעלה היא סף שרירותי הנועד לסנן רעש, מקרה בו אדם בודד אהב את שני הסרטים יהיה מקרה קיצון שיבזבז זיכרון ועלול להטות את תוצאות המודל.

חישוב רמת הסימטריה של המלצות:

#יצרנו טבלת השוואה בין הציון של AB לBA לבדוק את מידת הסימטריה של כל זוג,
score_difference הוא בעצם מדד אסימטריה, אם הוא 0 קיימת סימטריה מושלמת, ככל שהוא גדול הסימטריה חלשה.

```
SELECT
   COUNT(*) AS total_pairs,
   AVG(score_difference) AS avg_score_difference,
   SUM(CASE WHEN score_difference < 0.0001 THEN 1 ELSE 0 END) * 1.0 / COUNT(*)
AS prob_perfect_symmetry,
   SUM(CASE WHEN score_difference >= 0.5 THEN 1 ELSE 0 END) * 1.0 / COUNT(*) AS
prob_poor_symmetry
FROM (
   SELECT
      r1.movie_id AS movie_a,
      r2.movie_id AS movie_b,
      (COUNT(*) * 1.0) / m1.total_users AS score_a_to_b,
      (COUNT(*) * 1.0) / m2.total_users AS score_b_to_a,
      ABS(((COUNT(*) * 1.0) / m1.total_users) - ((COUNT(*) * 1.0) / m2.total_users)) AS
score_difference
   FROM personal_movies_ranking AS r1
   JOIN personal_movies_ranking AS r2
      ON r1.suggested_by = r2.suggested_by
     AND r1.movie_id < r2.movie_id
   JOIN (
      SELECT movie_id, COUNT(*) AS total_users
      FROM personal_movies_ranking
      GROUP BY movie_id
      HAVING COUNT(*) >= 2
   ) AS m1 ON r1.movie_id = m1.movie_id
   JOIN (
      SELECT movie_id, COUNT(*) AS total_users
      FROM personal_movies_ranking
      GROUP BY movie_id
      HAVING COUNT(*) >= 2
   ) AS m2 ON r2.movie_id = m2.movie_id
   GROUP BY r1.movie_id, r2.movie_id, m1.total_users, m2.total_users
   HAVING COUNT(*) >= 2
) AS direct_pairs;
```

סיכום תוצאות שאילתת חישוב רמת הסימטריה

	total_pairs	avg_score_difference	prob_perfect_symmetry	prob_poor_symmetry
▶	136312	0.542568929	0.06510	0.46149

ניתן לראות מסיכום התוצאות כי קרוב למחצית מהזוגות לא היו סימטריים (כ- 46 אחוזים) ואילו רק 6.5 אחוזים אכן היו סימטריים לחלוטין.

דוגמה לתוצאה שהתקבלה להשוואה בין הזוגות (הרצת השאילתה ללא מעטפת סיכום הנתונים)

movie_a	movie_b	score_a_to_b	score_b_to_a	score_difference
2	108	1.00000	0.66667	0.33333
2	121	0.75000	0.60000	0.15000
2	152	0.50000	1.00000	0.50000
2	153	0.75000	1.00000	0.25000
2	154	0.50000	1.00000	0.50000