שאלות Data Integrity

.1

כדי לתת המלצות סרטים משתמשים ברעיון שסרטים שיש בהם את אותם התפקידים הם המלצות טובות.
הכלל ממומש בשאילתה הבאה:

```
select
f.movie_id as fid
s.movie_id as sid ,
count(distinct f.role) as roles ,
from
imdb_ijs.roles as f
join
imdb_ijs.roles as s
on
f.role = s.role
group by
fid
sid ,
having
count(distinct f.role) >= 3
```

הוטל עליכם לשפר את ביצועי השאילתה. עליכם ליצור ואריציה עם precision עדיף וואריציה עם recall עדיף.
האילוץ הוא שהשאילתה נשארה כמו שהיא ומותר לכם לשנות רק את טבלת roles.

כיוונים אפשרים הם התייחסות ל:

1. התפקיד הנפוץ ביותר "" (שמשמש כשהתפקיד לא ידוע).
2. התפקיד Himself
3. התפקיד extra (ניצב)
4. תפקידי דמויי ניצב (נהג, פקיד קבלה)
5. תפקידים בעלי מופעים רבים בסרט (חיילים בסרט מלחמה)
6. תפקיד ראשי (והזיהוי שלו)
7. תפקידים ראשיים חוזרים (ג'יימס בונד, הארי פוטר)
8. תפקיד סופרמן\קלארק קנט
9. שמות דומים לאותו התפקיד (רוצח\מתנקש)
10. תפקידים דומים סמנטית (ביולוג/ארכיאולוג על תקן מדען מטורף)

1. הציעו 3 שיטות לכל ואריאציה.
2. ממשו ב sql או הסבירו מדוע sql אינה כלי מתאים למימוש הרעיון שלכם.

.2

בבנית ה gold standrd התבקשת לתת 200 זוגות של סרטים שמחציתם המלצות טובות ומחציתם המלצות בינוניות (בעלות קשר סביר אבל לא מספיק חזק להיות המלצה).

לקורס התגנב איש פלאי שמטרתו להזין המלצות לא מתאימות.
המניעים להמלצות הלא טובות עשויות להיות חוסר הבנה, חוסר זמן, רצון לחמוד לצון, זדון וכדומה.
חשוב מאד לאתר המלצות אלו כי המודלים שלכם ימדדו עליהן.

1. ציינו מוטיבציה וסוג המלצות לא מתאימות.
2. הסבירו כיצד יראו הנתונים.
3. הסבירו איך הנתונים ישפיעו על מדידת מערכת ההמלצות.
4. הציעו דרך לזהות את הנתונים. ממשו ב sql או הסבירו מדוע sql אינה כלי מתאים למימוש הרעיון שלכם.
5. הציעו דרך לתקן או להסיר את הנתונים. ממשו ב sql מדוע sql אינה כלי מתאים למימוש הרעיון שלכם.

ענו על הסעיפים למעלה עבור שלושה מקרים.