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.

.3

שמות השחקנים היא איזור מועד לשגיאות.

1. נניח שסרקתם את ויקיפדיה ובניתם רשימת שמות שחקנים. אלו בעיות הרשימה יכולה לעזור לזהות וכיצד? מה ההשלכות של בעיות כאלו על מודלי המלצה?
2. נניח שסרקתם את ויקיפדיה ובניתם רשימת שמות פוליטיקאים. אלו בעיות הרשימה יכולה לעזור לזהות וכיצד? מה ההשלכות של בעיות כאלו על מודלי המלצה?
3. רונלד רייגן היה שחקן ופוליטיקאי. נניח שהוא ואחרים נמצאים בשתי הרשימות. מה תהיה ההשפעה?
4. חישבו על שלושה מקרים שיכולים להוביל לשמות לא תקינים.
    a. ציינו את הבעיה
    b. הסבירו כיצד יראו הנתונים.
    c. הסבירו איך הנתונים ישפיעו על מדידת מערכת ההמלצות.
    d. הציעו דרך לזהות את הנתונים. ממשו ב sql או הסבירו מדוע sql אינה כלי מתאים למימוש הרעיון שלכם.
    e. הציעו דרך לתקן או להסיר את הנתונים. ממשו ב sql מדוע sql אינה כלי מתאים למימוש הרעיון שלכם.
5. אנחנו מקבלים כמובן מאליו שהשמות הם באנגלית. האם זו בעיה בנתונים? כיצד הייתם תומכים בריבוי שפות?

.4

המושג null records מציין את ה true negatives - רשומות שהן שליליות ומסווגות כשליליות.
בהמלצות שלנו הרוב המוחלט של הזוגות הן רשומות שליליות ויסווגו כך משום שאין כל אינדיקציה לקשר בניהן.
רשומות כאלו עלולות לפגוע בהערכת הביצועים, בוודאי כשהוזנו במקום רשומות מועילות יותר (המלצות בינוניות).
ניתן לאתר null records בעזרת מודלים בעלי recall מאד גבוה. רשומה של המלצה שלילית שלא מסווגת כחיובית על ידי אף אחד מהמודלים היא בהסתברות גבוהה null record.

1. מה תהיה ההשפעה של null records על המדדים accuracy, precision, recall
2. בנו מודל בעל recall גבוה על בסיס
    a. במאי משותף
    b. שחקן משותף
    c. ז'אנר משותף
    d. תפקיד משותף
    e. שנה קרובה
    f. על בסיס collaborative filtering "אנשים שאהבו את א', אהבו את ב'"
    g. בונוס: למה מודל של : "אנשים שדירגו את א', דירגו גם את ב'" יכול להועיל גם? ממשו
3. אחדו את כל המודלים לטבלה אחת
4. מיצאו את ההמלצות שהן null records
5. בידקו האם יש ממליצים הנוטים במיוחד להמליץ על null records

.5

טבלת personal_movies_ranking ב collaborative filtering מציגה את הדירוג שננתם לסרטים. הדירוגים עלולים להיות לא יציבים, כלומר שאותו אתם יתן ציון שונה לאותו הסרט בהזדמנויות שונות.

נניח שהוספנו שדה תאריך המלצה וביקשנו המלצות חוזרות.

1. מה השאילתה המציגה את זוגות ההמלצות של אותו אדם לאותו סרט?
2. חשבו את הסיכוי לציון זהה.
3. חשבו את הסיכוי למהפך (מעל לחמש מול חמש ומטה).
4. הוצע להשאיר רק את ההמלצות שהיו זהות פעמיים. מה היתרונות והחסרונות?
5. הוצע להשאיר רק ממליצים עם סטיית תקן נמוכה. מה היתרונות והחסרונות?
6. הוצע להשתמש בשתי הגרסאות. מה היתרונות והחסרונות?
7. ההמלצות הן על סולם טעם אישי. הוצע לנרמל את ההמלצות. מה היתרונות והחסרונות?

.6

המלצות על זוגות סרטים עלולות להיות שגויות. ניתן למצוא שגיאות כאלו על ידי השוואה לתחזית של מודל מוצלח. נניח שאיתרנו מקרה כזה, ביררנו עם הממליץ שאישר שזו שגיאה.

1. מה היתרונות והחסרונות בתיקון הרשומה?
2. מה היתרונות והחסרונות במחיקת הרשומה?
3. נניח שאנחנו מתקנים שגיאות במשך תקופה ארוכה. מה ההשלכה על מדידת ביצועי החיזוי של המודל לעמות מדידה על נתונים שלא עברו תהליך זה?

.7

בעיות נתונים עלולות להתעורר לא רק מבעיות ברשומות הקיימות אלא מאי קיום רשומות או שאלות יצוג.

1. ביחרו 10 סרטים ישראלים אהובים עליכם. בידקו מי נמצאים בבסיס הנתונים. בשל בעיות שפה חפשו גם לפי במאי ושחקנים כדי לוודא שהם לא קיימים. איך כדאי להתייחס להעדר סרטים ישראלים או ממקור אחר?
2. האם יתכן שחלק מהסרטים בבסיס הנתונים כלל לא קיימים? כיצד ניתן למצוא אותם?
3. מיצאו זוגות של סרטים בעלי אותו השם שיצאו בשנים עוקבות. מה יכולה להיות הסיבה לקיומם? איך קיומם ישפיע על מערכת ההמלצות?
4. נשאלת השאלה האם להחשיב נאומים מוסרטים של פוליטיקאים כסרטים. כיצד ההחלטה תשליך על הנתונים והניתוח שלהם?
5. משיקולי כלכלת פשע, פושעי סייבר עושים malware coloring ומייצרים הרבה גרסאות של אותה התוכנה זדונית. כך לתוכנה הזדונית יש את אותה הפונקציונליות אבל החתימות האוטומטיות של כל גרסה שונה וכך החתימה לא קיימתבמידע היסטורי. חברת סייבר הגנתי מתלבטת אם לשמור גרסאות מרובות של אותה תוכנה זדונית. מה תמליצו בתרחישים שונים? על בסיס מה כדאי לבסס את ההחלטה?