# Watch it

# הסבר על השיטות הסטטיסטיות

## מבנה נתוני הדירוג

מטריצה שבה השורות הן X צופים והעמודות הן Y סרטים.

הערכים בכל תא במטריצה הם הדירוג שנתן הצופה לסרט.

הדירוגים הם ערך מספרי בין 1 ל-5.

אם אין דירוג, מוצב ערך 0. באופן טבעי, ברוב הגדול של התאים אין דירוג (ערך 0).

## שיטת ההמלצה

האפליקציה מציגה בפני כל משתמש מבחר סרטים לדירוג.

המשתמשים מדרגים "אהבתי", "אדיש", "לא אהבתי" עבור סרטים שראו (מתורגם לציונים 1,3,5 בהתאמה)

כעת האפליקציה מוצאת לכל משתמש את k השכנים הקרובים ביותר לו במטריצת הנתונים. הם מכונים "ממליצים".

לכל ממליץ יש וקטור של דירוגים לכל סרט, מתוך מטריצת הנתונים. מתרגמים את זה להמלצות באופן הבא:

* ציונים 4,5 מתורגמים לנקודה בעד (+1)
* ציונים 1,2 מתורגמים לנקודה נגד (-1)
* ציון 3 מתורגם לאדישות (0), ובאופן זה שקול לדירוג חסר (ערך 0)

כעת סוכמים את הווקטורים של הממליצים של כל משתמש ומקבלים את וקטור הדירוג הקבוצתי.

הסרט המומלץ לקבוצה הוא זה שזכה בדירוג הגבוה ביותר.

## בחירת הסרטים לדירוג

ברצוננו להציג למשתמשים רשימת סרטים קצרה שתספק לנו מידע מירבי על מנת לזהות טעמים.

נשקלו מספר שיטות לעשות זאת:

* הצגת הסרטים עם מספר הדירוגים הרב ביותר
* הצגת הסרטים עם הציונים הטובים ביותר והציונים הגרועים ביותר

שיטות אלו נדחו משום שאין סיבה להניח שהן יעזרו לנו להבחין בצורה טובה בין סוגי צופים.

* הפעלת PCA על מטריצת נתונים מנורמלת

PCA היא שיטה מקובלת לזיהוי פקטורים בעלי השפעה מירבית על שונות. על מנת להתמודד עם בעיה של ערכים חסרים, התכוונו לנרמל את הציונים מסקאלה של (1,2,3,4,5) לסקאלה של (-2,-1,0,1,2), כך שציון 3 שקול לערך חסר.

עם זאת מספר הפקטורים (סרטים) גדול יותר ממספר השורות (צופים מדרגים), ובמצב כזה לא ניתן להפעיל PCA.

לבסוף החלטנו להשתמש במודל:

* שיטת Reddit לזיהוי פריטים שנויים במחלוקת

ההנחה היא שסרטים שנויים במחלוקת יכוונו אותנו לסוגים שונים במיוחד של צופים.

בשיטה זו לוקחים את סכום הערכים המוחלטים של הציונים בעד ונגד כל סרט (זה נקרא magnitude), ומנרמלים באמצעות הכפלה ביחס בין סכום הציונים בעד לסכום הציונים נגד.

אם יש יותר ציונים בעד, אז מחשבים נגד \ בעד

ואם יש יותר נגד, אז מחשבים בעד \ נגד

באופן הזה מקבלים פקטור בין 0 ל-1, שמתקרב ל-1 כאשר סכום הבעד והנגד מתקרבים זה לזה, ומתקרב ל-0 כאשר אחד גדול משמעותית מהשני.