תרגיל בית 2 – אחזור מערכות מידע ומערכות המלצה

מגישים: קבוצה מספר 1
אופק דור-318722360,מאיה שחורי- 314082595, משה דוד דדון -207128638, גיא שייך -208720458

הסבר על המודלים:

Simple Mean - לצורך החיזוי במודל זה, בוצע שימוש במבנה נתונים של מילון (Hash Table) כאשר המפתח הוא מספר המשתמש הנמצא בסט האימון, והערך הוא ממוצע הדירוגים של המשתמש, על פני כל הפריטים שדירג. ממוצעים אלו חושבו באמצעות המתודה Group By של הספרייה Pandas.

Slope One - אלגוריתם זה מומש על ידי שימוש במערכי Numpy, תוך התבססות על הבדלי הפופולריות הקיימים בקרב המשתמשים בין הסרטים בסט הנתונים. בעזרת האלגוריתם, נוכל להשתמש בדירוג של סרט נתון על מנת לחזות דירוג של סרט אחר . נחשב את ממוצע ההפרשים בין דירוגי משתמשים אשר ביצעו דירוג עבור שני הסרטים, ונוסיף אותו לדירוג הסרט הקיים ולסך הדירוגים נבצע ממוצע משוקלל בהתאם למספר האנשים שדירגו את אותו סרט נתון ,על ידי כך נחזה את דירוג הסרט החסר.

Simple Users KNN + Simple Items KNN - לצורך ביצוע החיזוי במודלים אלו בוצע שימוש בעקרונות הבאים :

מטריצה דלילה: מבנה נתונים מסוג מטריצה, השומרת את דירוג המשתמש ה U לסרט ה I- עבורitem KNN במקום ה [U,I] ועבור user KNN במקום ה [I,U] המרת הנתונים למבנה זה מייעלת את המשך החישובים באימון המודל, מאפשרת שליפת נתונים נוחה ומביאה לזמני ריצה מהירים יותר.

חישוב ממוצעים: בעזרת numpy ניתן למצע על פני כל השורות תוך התעלמות מערכי ה 0. ערכי הדירוג מאותחלים לממוצע, כלומר כאשר נפגוש לראשונה בסרט הוא יקבל את ממוצע הדירוגים של המשתמש. זאת, על פי ההנחה כי דירוגים של משתמש מסוים ינועו סביב אותה סקאלת דירוגים בהתאם לאופי המשתמש באופן דומה, עבור משתמש חדש, נאתחל את הדירוג שלו לסרט מסוים בתור ממוצע דירוגי הסרט על ידי שאר המשתמשים הקיימים. זאת, שכן סביר להניח כי סרט מסוים יקבל דירוג נוסף שיהיה דומה לאלה שקדמו לו בממוצע, בהתאם לדעה הרווחת בקרב מרבית קהל הצופים.

שמירה יעילה : בכדי לחשב את הדימיון בשני המודלים השתמשנו במדד pearson correlation. חישוב הדמיון בין כלל הסרטים/המשתמשים (בהתאמה לכל פונקציה שמומשה) בוצע פעם אחת בלבד עבור כל זוג משתמשים ועבור כל זוג אייטמים משום שאין חשיבות לסדר ולכן הפך את השימוש במודל ליעיל יותר וחסך בזיכרון .

בכדי לייעל את זמן הריצה ביצענו בשני המודלים את שלב ה preprocessing בו אנו שומרים את מילון הדימיון בין כל שני אייטמים בזיכרון.

עבור מודל User KNN - לא קיבלנו את פונקציית save params אך כן נתבקשנו לבנות את פונקציית upload params. משום שהבנו כי לא ניתן ליצור פונקציות נוספות בקובץ ומפני שפונקציית upload params

הינה חסרת משמעות אם לא קיימת הפונקציה save params הכנסנו את המתודה ששומרת את הפרמטרים כחלק מפונקציית הפיט. עם זאת, יש לציין כי אנו סבורים שעקב הקטנת סט האימון של מודל זה ל 200,000 שורות בלבד לא היה צורך כלל בשתי הפונקציות האלו וכמו כן גם לא ראינו שיפור בזמני הריצה כאשר ביצענו preprocessing לעומת בלעדיו.

*כאשר נרצה לבדוק את ביצועי המודל על סט הווולידציה, במידה ונרצה לחשב את הדימיון עבור ערך שלא דורג בסט ה train נשלוף ערך זה מסט הווולידציה. כתוצאה מכך התוספת לחישוב ה RMSE תהיה 0. ניתן לראות זאת בפרט עבור מודל User KNN בו קיבלנו סט נתונים בגודל של 200,000 שורות ראשונות בלבד ובמידה ובסט הווולידציה יופיעו משתמשים אשר לא מופיעים בשורות אלו בסט ה train נפעל כפי שציינו לעיל.

<u>תוצאות המודלים:</u>

מודל	תוצאה - RMSE	זמן ריצת האלגוריתם לא כולל preprocessing	זמן ריצת preprocessing
Slope One	0.9481523385597798	17.55 seconds	30 minutes
Simple Mean	1.049310719086559	5.46 seconds	
Item KNN	1.0112376692222997	13.6 seconds	25 minutes
User KNN	0.31790024029226605	45.02 seconds	2.5 minutes

<u>הסבר וניתוח התוצאות</u>

ניתן לראות כי קיבלנו עבור User KNN קיבלנו את ה RMSE הנמוך ביותר כלומר התוצאה הטובה ביותר. המשמעות היא שמודל זה ביצע את החיזוי בצורה הטובה ביותר. עם זאת , אנו סבורים כי ייתכן ותוצאת המדד מוטת ויצאה כך משום שמודל זה קיבל סט אימון נתונים קטן יותר משום שהיה מורכב רק מ 200,000 השורות הראשונות עוד שאר המודלים קיבלו את סט הנתונים המלא. בנוסף, כפי שציינו לעיל הוספנו למדד את הערך 0 עבור כל משתמש שהופיע בסט הווולידציה בלבד ולא בסט ה train . אנו סבורים שגודל סט הנתונים עלול להשפיע על מדד ה RMSE ממספר סיבות: עבור מערכי נתונים קטנים יותר, קיים סיכון גבוה יותר להתאמת יתר, מה שעלול להוביל לערכי RMSE אופטימיים מדי במהלך הערכת המודל. בנוסף עבור סט נתונים גדול עשויה להיות שונות גבוהה יותר השונות המוגברת הזו יכולה להשפיע על RMSE מכיוון שהיא מייצגת את השורש הריבועי של ההבדלים הממוצעים בריבוע בין הערכים החזויים והנצפים. לכן, אם מערך הנתונים גדול ומגוון, ה-RMSE עשוי להיות גבוה יותר. כמו כן מערכי נתונים גדולים יותר נוטים לספק הערכות אמינות וחזקות יותר של ביצועי המודל, מה שמוביל לערכי RMSE אמינים יותר. לאור הסיבות שציינו לעיל אף על פי שתוצאת RMSE שלו הינה הטובה ביותר לא ניתן לומר בודאות כי מודל User KNN מבצע את החיזוי בצורה הטובה ביותר. לגביי שאר המודלים קיבלנו ערכים דיי דומים ומכיוון שאין הבדל בסט הנתונים שקיבלו נשער כי מבין שלושתם המודל שיבצע את החיזוי בצורה הטובה ביותר עבור בעיה זו הינו Slope Oneמשום שמדד ה RMSE עבורו הוא הנמוך ביותר. חשוב לציין כי אין זה אומר שבאופן גורף מודל זה יניב את התוצאות הטובות ביותר משום שישנה השפעה לסט הנתונים ולבעיה אותה אנו פותרים.