בס"ד

תאריך הגשה: 24.1.23

תרגיל 3 – מערכות המלצה

הקדמה

בתרגיל זה תבנו מערכת המלצה למוצרים קוסמטיים.

התרגיל מחולק ל 3 חלקים

- ניתוח הנתונים
- בניית מערכת המלצה CF
- הערכות ביצועים

בקובץ **RAR** המצורף נתונים לכם הקבצים הבאים:

main.py – הקובץ הראשי, דרכו נקרא לכל המימושים השונים.

אין לשנות קובץ זה!

data.py – כאן תממשו פוונקציה להבנת הנתונים.

collaborative_filtering.py – כאן תממשו מערכת המלצה CF מסוג user-based ו – item-based

evaluation.py – כאן תממשו את פונקציות ההערכה השונות למערכת ההמלצה.

train.csv ו-test.csv – קבצי הנתונים בהם תשתמשו במהלך תרגיל זה.

הגשה

במהלך התרגיל תערכו את הקבצים הבאים: data.py, collaborative_filtering.py ו - evaluation.py

עליכם לשלוח קבצים אלה עם הקוד והערות שלכם. נא לא לשלוח אף אחד מהקבצים המקוריים מלבד קבצים אלה.

בראש כל אחד מקבצים אלו נא לכתוב את שם הסטודנט באנגלית ות.ז.

בנוסף לקבצים אלו עליכם להגיש דו"ח העונה על שאלות בתרגיל בקובץ בשם **report_<id>.pdf**, כאשר id יהיה ת.ז. של הסטודנט וכן קובץ פרטים אישיים בשם detail.txt והוא יכיל את שם הסטודנט ות.ז.

ההגשה דרך מערכת ה**Submit** בלבד!

חלק ראשון – נתונים

בקובץ data.py נתונה לכם הפו' watch_data_info(data). העזרו בפו' הנ"ל כדי להבין את קובץ הנתונים (train.csv) שצורף לתרגיל. (ניתן גם לפתוח את הקובץ עצמו ופשוט להסתכל על הרשומות)

נתייחס לקובץ הדירוגים:

1. כמה משתמשים ייחודיים דרגו את המוצרים? כמה מוצרים ייחודיים דורגו?
כמה דירוגים קיימים בקובץ שניתן?

2. מהו מספר הדירוגים המינימלי והמקסימלי שניתן למוצר?

3. מהו מספר הדירוגים המינימלי והמקסימלי שמשתמש דירג?

את המימוש לקבלת התשובות הנ"ל יש לכתוב בפו' print_data(data) על ידי השלמת ערכי ההדפסות המתאימות ולאחר מכן לכתוב את התשובות בדו"ח.

4. א. ישנם מקרים בהם יש מוצרים או משתמשים בעלי דירוגים מועטים או ללא דירוגים כלל.
במקרה זה מערכת ההמלצה מסוג collaborative filtering תתקשה, מדוע?
ב. באילו מערכות המלצה נשתמש כדי לטפל במוצרים חדשים, או מוצרים עליהם אנו מעוניינים להמליץ למרות מיעוט נתוני הדירוג? באיזו מערכות המלצה נשתמש עבור משתמשים חדשים, להם אין דירוגים כלל?

חלק שני - **collaborating-filltering**

בקובץ collaborative_filtering.py ישנה מחלקה למימוש מערכת המלצה מסוג CF.

הפונקציה מקבלת משתנה המייצג את סוג המימוש עבור ה Collaborative Filtering:

- User based Collaborative Filtering עבור - user
- Item based Collaborative Filtering עבור - item

פונקצית הדמיון בה תשתמשו במימוש תהיה פונקצית ה cosine.
תזכורת: המדד מחשב את הזווית בין שני וקטורים במרחב:

$$\cos(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x} \cdot \mathbf{y}^{\mathsf{T}}}{\|\mathbf{x}\| \cdot \|\mathbf{y}\|} = \frac{\sum_{i=1}^{n} \mathbf{x}_i \cdot \mathbf{y}^{\mathsf{T}}{}_i}{\sqrt{\sum_{i=1}^{n} (\mathbf{x}_i)^2}\sqrt{\sum_{i=1}^{n} (\mathbf{y}_i)^2}}$$

שימו לב – ניתן להיעזר בקוד (וחבילות) כפי שהודגם בשיעור.

5. ממשו את הפונקציה fit(matrix), המקבלת כקלט את מטריצת הנתונים, מנרמלת הדירוגים על ידי חיסור ממוצע דירוגי המשתמש ובונה את מטריצת החיזוי.
(את הערכים במטריצה זו יש לעגל ל- 2 ספרות אחרי הנקודה)

6. ממשו את הפונקציה recommend_items(user_id, k=5) המקבלת כקלט id של משתמש , ומס' המלצות (המוגדרות כברירת מחדל ל 5) ומחזירה רשימה של k המוצרים המומלצים עבור המשתמש (כל פריט מיוצג באמצעות productid). במידה וה - id לא קיים החזירו ערך None.
ניתן להוסיף פונקציות עזר ככל שתרצו.

7. מה יהיו 5 המוצרים המומלצים עבור משתמש "**AQWF3BBBDL4QJ**" על פי user-user cf

8. מה יהיו 5 המוצרים המומלצים עבור משתמש "**A3EO0WA7R3LVBQ**" על פי item-item cf

חלק שלישי – הערכות

השתמשו בקובץ test.csv של דירוגי המשתמשים שקיבלתם, בעזרתו תעריכו את מערכות ההמלצה שבניתם. המדדים הם השוואתיים, ולכן נשווה את ביצועי המערכת מול benchmark - מערכת בסיסית פשוטה לחישוב. את ההשוואות נעשה כמובן על הtest-set ונצפה כי ביצועי המערכות שכתבנו יהיו טובים יותר.

שימו לב – גם בחלק זה ניתן להוסיף פונקציות עזר ככל שתרצו.
ניתן להניח שהמשתנה cf שפונקציות ההערכה מקבלות כבר מאותחל.

מדד עבור חיזוי רייטינג: RMSE - מדד להבדלים בין ערכים חזויים על ידי מודל או אומדן לבין הערכים האמיתיים ומוגדר כך:

$$RMSE=\sqrt{\frac{\sum_{i=1}^{N}(Predicted_i-Actual_i)^2}{N}}$$

את הערכת מערכת ההמלצה באמצעות מדד זה, נבצע מול benchmark של ממוצע דירוגי המשתמש מה train set. (בדומה למטריצת הממוצעים באמצעותה נרמלנו את מטריצת הדירוגים המקורית). כלומר מערכת החוזה דירוג קבוע עבור כל הפריטים, עבור משתמש בודד.

ממשו את פונקצית RMSE(test_set, recommender) המקבלת את ה test set ואת משתנה ה recommender שנבנה בחלק הקודם ומדפיסה את מדד ה RMSE עבור ה test set - ועבור ה banchmark עם הודעה מתאימה. (את המדד יש לעגל ל- 5 ספרות אחרי הנקודה)

9. ערכו את ההשוואה עבור user-based ועבור item-based.
את תוצאות הבדיקה ציינו בדוח בטבלה הנ"ל:

	RMSE
user-based CF	
item-based CF	
mean based (benchmark)	

מדדים עבור המלצת k פריטים:

מדד precision@k

(Relevant_Items_Recommended) / (*k*)

מדד - recall@k

(Relevant_Items_Recommended) / (Relevant_Items)

K – מיוחס למספר הפריטים שיחזרו ממערכת ההמלצה.
Relevant items – מספר הפריטים הרלוונטיים הם הפריטים שדירוגם בפועל גדול או שווה ל-3.
Recommended items@K – top-K הפריטים המומלצים ממערכת החיזוי
Recommended and Relevant items@K – (Relevant_Items_Recommended)
החיתוך בין Recommended items@K ו - Relevant items

כדי להבין את ההגדרות של recall@k ו-precision@k, נסביר באמצעות דוגמה:

```
Relevant items: item5, item10 and item1
total # of relevant items = 3

Recommended items @ 3: item7, item5 and item10
# of recommended items at 3 = 3

Recommended@3 INTERSECTION Relevant: item5 and item10
# of recommended items that are relevant @3= 2

Precision@3
=(# of recommended items that are relevant @3)/(# of recommended
items at 3)
= 2/3
= 66.67%

Recall@3
= (# of recommended items that are relevant @3)/(total # of
relevant items)
= 2/3
= 66.67%
```

את הערכת מערכת ההמלצה באמצעות מדדים אלו, נבצע מול benchmark של מערכת הממליצה לכל המשתמשים את אותו סט המלצות: **k** הפריטים עם ממוצע הדירוג הגבוה ביותר מה training set.

<u>שימו לב</u> - ציון ה recall@k ו-precision@k הסופי יהיה ממוצע על כל המשתמשים.

ממשו את הפונקיות precision_at_k(test_set, recommender, k)
ו - recall_at_k(test_set, recommender, k) כשכל אחת מהן מקבלת את ה test set, את משתנה ה recommender שנבנה בחלק הקודם ו K המציין את top-k ההמלצות למשתמש, ומדפיסה את המדד ה precision@k ו - recall@k בהתאמה, עבור ה - test set ועבור ה banchmark עם הודעה מתאימה. (את המדדים יש לעגל ל- 5 ספרות אחרי הנקודה)

10. ערכו את ההשוואה עבור user-based. את תוצאות הבדיקה ציינו בדוח בטבלה הנ"ל:

	Precision@20	Recall@20
user-based CF		
highest-ranked (banchmark)		

בהצלחה רבה!