Assignment3 - Report

Dorin Keshales - 313298424

חלק א' – Non-personalized

1. בחרתי במדד Weighted Average Ratings כלומר, עבור כל ספר נעשה ממוצע ממושקל של הratings יחד עם מס' הקוראים שדירגו אותו. בחרתי במדד זה מכיוון שהוא נותן ציון אמין אשר יכול לשקף את הספרים המומלצים במערכת בהתאם למס' הדירוגים שניתנו להם ואיכותם. לצורך חישוב מדד זה, השתמשתי בקבצים books.csv ו- ratings.csv, בכדי שתהיה לי גישה לנתוני ה- book_id, ה-title וה-rating. עבור כל ספר אני מחשבת את כמות הקוראים שדירגו את הספר ואת ממוצע הדירוגים עבור ספר זה ועל בסיס נתונים אלו ובעזרת מדד הדמיון Weighted Average Ratings אני מחשבת את ה-Score (ה-rating המשוקלל) עבור כל ספר במערכת. באמצעות אותם Scores שהתקבלו ניתן להמליץ על K הספרים הכי מומלצים שקיימים במערכת.

2. להלן, עשרת הספרים המומלצים במערכת:

rank	book_id	title	score
1	25	Harry Potter and the Deathly Hallows (Harry Potter, #7)	4.33803
2	4	To Kill a Mockingbird	4.29984
3	102	Where the Wild Things Are	4.27321
4	85	The Giving Tree	4.24031
5	50	Where the Sidewalk Ends	4.23972
6	31	The Help	4.23885
7	144	Unbroken: A World War II Story of Survival, Resilience, and Redemption	4.22186
8	27	Harry Potter and the Half-Blood Prince (Harry Potter, #6)	4.21391
9	1	The Hunger Games (The Hunger Games, #1)	4.18738
10	133	Anne of Green Gables (Anne of Green Gables, #1)	4.18149

3. להלן, עשרת הספרים המומלצים עבור משתמש הגר ב- Ohio:

rank	book_id	title	score
1	126	Dune (Dune Chronicles #1)	4.36796
2	143	All the Light We Cannot See	4.31779
3	144	Unbroken: A World War II Story of Survival, Resilience, and Redemption	4.26609
4	24	Harry Potter and the Goblet of Fire (Harry Potter, #4)	4.24973
5	102	Where the Wild Things Are	4.22688
6	490	Maus I: A Survivor's Tale: My Father Bleeds History (Maus, #1)	4.21366
7	1462	The Orphan Master's Son	4.21366
8	983	Between the World and Me	4.21366
9	119	The Handmaid's Tale	4.19957
10	89	The Princess Bride	4.19006

4. להלן, עשרת הספרים המומלצים עבור משתמש שגילו 28:

rank	book_id	title	score
1	25	Harry Potter and the Deathly Hallows (Harry Potter, #7)	4.32625
2	4	To Kill a Mockingbird	4.2942
3	85	The Giving Tree	4.28961
4	89	The Princess Bride	4.2447
5	133	Anne of Green Gables (Anne of Green Gables, #1)	4.22491
6	50	Where the Sidewalk Ends	4.21641
7	102	Where the Wild Things Are	4.20468
8	70	Ender's Game (Ender's Saga, #1)	4.2041
9	31	The Help	4.20289
10	21	Harry Potter and the Order of the Phoenix (Harry Potter, #5)	4.19638

חלק ב' – Collaborative filtering

5. מומש בקוד.

6. מומש בקוד.

7. מומש בקוד.

חלק ג' – Content-based Filtering

8. בחרתי להשתמש בפיצ'רים הבאים מתוך הקובץ books.csv :

"title", "authors", "original_publication_year", "language_code"

בעבור הפיצ'ר **"authors"** אני ראשית מסננת כותבים נוספים ומשאירה רק את הכותב הראשון ברשימה ממחשבה או ציפייה שהוא הכותב המרכזי של הספר. לאחר מכן אני הופכת את הפיצ'רים **"title"**, **"authors"** ו- **"language_code"**, שהם פיצ'רים קטגוריאליים , לפיצ'רים בינאריים באמצעות הפונקציה `get_dummies` של pandas.
עבור הפיצ'ר **"original_publication_year"** אני יוצרת bins באמצעות הפונקציה `cut` של pandas. הערכים של ה- bins הם הערכים שמופיעים כאשר עושים describe() לפיצ'ר **"original_publication_year"** - אלו בדיוק הערכים של min, 25%, 50%, 75%, max.
לאחר החלוקה לbins, אני משתמשת בפונקציה `get_dummies` בכדי לקבל ערכים בינאריים לפיצ'ר.

9. מומש בקוד.

10. להלן, עשרת הספרים הכי דומים לספר 'Twilight (Twilight, #1)' (book id = 3, שורה 4 בקובץ books.csv):

rank	book_id	title
1	52	Eclipse (Twilight, #3)
2	73	The Host (The Host, #1)
3	56	Breaking Dawn (Twilight, #4)
4	992	The Twilight Saga (Twilight, #1-4)
5	2233	Possible Side Effects
6	1630	Turn Coat (The Dresden Files, #11)
7	4723	The Lost Painting
8	1398	Extras (Uglies, #4)
9	4446	Fade Out (The Morganville Vampires, #7)
10	1397	The Lost Colony (Artemis Fowl, #5)

חלק ד' – מדדי הערכה

.11

	precision_k	ARHR	RMSE
cosine	0.08	0.323	0.901
euclidean	0.008	0.033	0.919
jaccard	0.08	0.322	0.905

.12 ניתן לסדר את מדדי ההערכה לפי טיב הדיוק שקיבלנו בכל אחד מהם, בצורה הבאה:

$$Percision@k < ARHR < RMSE$$

ראשית, נשים לב שהדיוק שקיבלנו במדד ההערכה $Percision@k$ הוא הכי נמוך בעבור כל מטריקות הדמיון, בפרט הוא גם נמוך מהדיוק של $ARHR$, הסיבה לכך היא שמדד ההערכה $ARHR$ הוא יותר מדויק, שכן ב- $ARHR$ כל חיזוי מקבל ניקוד לפי מיקומו ברשימת ההמלצות ולכן יש הבדל בין hit עבור ספר הנמצא במקום הראשון או השני ברשימת ההמלצות (ניקוד של $\frac{1}{1} = 1$ ו- $\frac{1}{2}$ בהתאמה) לבין hit עבור ספר הנמצא במקום התשיעי או העשירי ברשימה (עבור k=10). כלומר, בסוף רשימת ההמלצות (ניקוד של $\frac{1}{9}$ ו- $\frac{1}{10}$ בהתאמה). בנוסף, נשים לב שדווקא עניין זה מצביע על כך שיחסית שליש מהמיקומים של ההמלצות שלי היו נכונים כאשר מדובר במטריקות הדמיון Cosine ו-Jaccard. לעומת זאת, מדד ההערכה $Percision@k$ מנקד כל hit שיש לנו ברשימת ההמלצות של ה- top k בלי קשר למיקומו – בעבור כל hit שזיהינו נקבל ניקוד של 1. ולכן, כפי שציינתי קודם, מדד ההערכה $ARHR$ הינו יותר מדויק כאן.
כעת, נשים לב לפער הגדול שקיים בין תוצאות מדדי ההערכה של $Percision@k$ ו- $ARHR$ לבין תוצאות המדד $RMSE$. קל להבין שהפער הזה נובע מהשוני המהותי בין המדדים, שכן בשני המדדים הראשונים החישוב מתבצע רק בעבור k המקומות הראשונים ברשימת ההמלצות ואילו במדד ה- $RMSE$ החישוב מתבצע לאורך כל רשימת ההמלצות. כלומר, בשני המדדים הראשונים אנחנו יכולים לקבל ניקוד רק עבור k ההמלצות הראשונות, בעוד שב-$RMSE$ אנחנו נותנים 'משקל' לכל ההמלצות ברשימת ההמלצות. כלומר, גם ספר שדורג נמוך וחזיתי אותו כנמוך יתמוך בהעלאת הניקוד, כלומר יתרום לציון הסופי של מדד זה, שכן במדד זה אנחנו 'נענשים' רק על הפער בדירוג.