מסחר אלקטרוני - תרגיל בית 2

מתרגל אחראי : ג'וני

בתרגיל זה תבנו מערכת המלצה לאפליקציית ספוטיפלי (ראו תרגיל בית 1). הנתונים המצורפים מכילים תיעוד של השמעות בקרב כ- 2000 משתמשים וקרוב ל 20,000 אמנים. לצורך בניית מערכת ההמלצה, תוכלו להיעזר במידע מן הרשת החברתית המתארת קשרים בין המשתמשים. על בסיס נתונים אלו, תצטרכו לנבא את מספר הפעמים שמשתמש מסוים ישמע אמן מסוים.

נתונים

- הקובץ user_artist מכיל חלק מנתוני ההשמעות. בכל שורה, העמודה הראשונה מכילה מזהה משתמש, השנייה מזהה אמן והשלישית ("weight") את מספר הפעמים שאותו משתמש/ת שמעו את אותו אמן.
- הקובץ test מכיל זוגות של משתמשים ואמנים עבורם אתם צריכים לנבא את מספר ההשמעות.
- עבור כל זוג של משתמש ואמן שלא מופיעים באותה שורה באף אחד מהקבצים לעיל, מספר ההשמעות הוא אפס (המשתמש לא שמע את אותו אמן כלל).
- הקובץ user_friends מתאר את הרשת החברתית. כל שורה מכילה זוג שכנים ברשת. הגרף אינו מכוון ואינו ממושקל. כל זוג שכנים מופיע בקובץ פעמיים מטעמי נוחיות, אך אין לכך משמעות. כפי שיובהר בהמשך, השימוש בקובץ זה הוא אופציונאלי בלבד.

הערכת התחזיות

לכל זוג של משתמש u ואמן i בקובץ test, נסמן ב $\hat{r}_{ui}$ את התחזית וב r_{ui} את הערך האמיתי (שאינו ידוע לכם). הקנס עבור כל זוג שכזה הינו:

$$l_{ui} = (\log \hat{r}_{ui} - \log r_{ui})^2$$

(log בבסיס 10), ואילו הקנס הכולל הוא

$$L = \sum_{(ui)\in test} l_{ui}$$

מטרתכם היא למזער את הקנס הכולל.

משימה:

בתרגיל בית זה ישנן שתי משימות:

1. למזער את הקנס הכולל L ללא שימוש ברשת החברתית כלל. במשימה זו, ברשותכם להשתמש אך ורק בנתוני השמעות העבר.
2. למזער את הקנס הכולל L תוך שימוש בהשמעות העבר <u>וגם</u> ברשת החברתית.

ניתן להבין כי משימה 2 הינה מאתגרת יותר. <u>על כן, משימה 2 הינה אופציונלית בלבד</u>. מנגד, יינתן בונוס מוגדל למצליחים במשימה זו.

כפי שנרחיב מטה, כל זוג יתמודד בשתי המשימות, אך הציון יהיה המקסימום מבין שתי המשימות. היות שכך, ציון הזוג לא יוכל להיפגע במידה ויחליטו להתמודד עם משימה 2.

הגשה

כל הגשה צריכה* לכלול ארבעה קבצים:

1. קובץ ID1_ID2_task1.csv הכולל את שתי העמודות בקובץ test בתוספת עמודה "weight" עם הניבויים שיצרתם למשימה 1, כלומר ללא שימוש ברשת החברתית. הקפידו לא לשנות את סדר השורות.
2. קובץ ID1_ID2_task2.csv הכולל את שתי העמודות בקובץ test בתוספת עמודה "weight" עם הניבויים שיצרתם למשימה 2, כלומר בשימוש ברשת החברתית. הקפידו לא לשנות את סדר השורות.
3. קובץ ID1_ID2.PDF <u>באורך 2 עמודים לכל היותר</u> המסביר איך יצרתם את הנ"ל. הסבר לא מפורט יגרום לפסילה של התרגיל. בנוסף לשני העמודים המרכזיים, תוכלו לכלול נספח ארוך כרצונכם.
4. קובץ ID1_ID2.py המכיל את כלל הפונקציות בהם השתמשתם, באופן שניתן יהיה לשחזר את הניבויים שלכם.

*הערות חשובות:

- את כל הנ"ל יש להגיש בתוך תיקייה דחוסה בשם HW2_ID1_ID2.zip.
- סטודנטים שלא מעוניינים להשתתף במשימה 2, יעתיקו את הניבויים ממשימה 1 ויגישו אותם בתור ID1_ID2_task2.csv. בכל מקרה, יש להגיש ארבעה קבצים
- שמות הקבצים והתיקייה של סטודנטים המגישים לבד צריכים לכלול מספר ת"ז יחיד, למשל- ID1_test.csv

ציון

הציון הוא כרגיל תחרותי בין הסטודנטים. לכל משימה תתבצע תחרות משלה, כלומר שתי תחרויות שונות. לכל זוג ולכל תחרות יקבע ציון, וציון תרגיל הבית של הזוג יהיה המקסימום מביניהם.

בהצלחה!