מערכות המלצה, תרגיל מספר 3

חישוב המלצות

כללי: בתרגיל זה תתבקשו לממש מערכת המלצה (בניגוד למערכת החיזוי הקיימת). לפיכך תדרשו לממש גם הסבה של המערכות שבניתם קודם לכן לביצוע המלצות (ולא רק חיזוי). כמו כן תדרשו לממש טכניקות שונות להשוואה בין מערכות המלצה – Precision-Recall.

1. בסעיפים הבאים ננסה לחזות אילו סרטים המשתמש ידרג (יצפה). נשאל את השאלה – בהינתן מספר סרטים שהמשתמש דירג, אילו עוד סרטים הוא ידרג?
   1. ממשו את השיטה Recommend המקבלת שם אלגוריתם ומספר המלצות n, ומחזירה רשימת המלצות באורך n (בדיוק) שחושבו על פי האלגוריתם.
   2. עליכם לממש את האלגוריתמים המופיעים למטה.
   3. הערה, בכל המקרים למטה יש לזכור להסיר מרשימת ההמלצות סרטים שכבר דורגו על ידי המשתמש.
2. ממשו מערכת המלצה מבוססת פופולאריות, הממליצה על הסרטים אותם המשתמש טרם דירג בסדר יורד של פופולאריות (מספר דירוגים). שם האלגוריתם יהיה “Popularity”.
3. ממשו מערכת כללית המשתמשת במערכת חיזוי על מנת לחשב המלצות. המערכת תיקח אלגוריתם חיזוי, ותבקש ממנו חיזוי על כל הפריטים שהמשתמש טרם דירג. ההמלצה תהיה רשימת הפריטים מדורגים בסדר יורד של ציון חזוי.
   1. אלגוריתמים אפשריים – Cosine, BaseModel, Pearson, Stereotypes
4. ממשו מערכת כללית המשתמשת במערכת משקלות (למשל Pearson או Cosine) על מנת לחשב המלצות. המערכת תשתמש בסביבת המשתמש (k המשתמשים הדומים לו ביותר, ניתן להשתמש ב-k=20), ותמליץ על הפריטים שהסביבה דירגה באופן חיובי, ממוינים בסדר יורד על פי משקל.
   1. משקלו של סרט על ידי הסביבה הוא סכום המשקלות של המשתמשים שדירגו את הסרט. יש לנרמל את המשקלות.
   2. אלגוריתמים אפשריים – NNCosine, NNBaseModel, NNPearson. כאשר NN=Nearest Neighbors.
   3. הוסיפו חישוב דומות בין משתמשים על פי מקדם Jaccard:

שם השיטה - NNJaccard

1. ממשו מערכת המלצה (לא חיזוי) מבוססת פריטים (Item-based). לשם כך, עליכם לחשב הסתברות מותנית עבור זוגות פריטים pr(i2|i1). בהינתן פריט i1 ההמלצה תהיה רשימת פריטים מסודרים (בסדר יורד) על פי ההסתברות המותנית לצפייה בהם בהינתן הצפייה ב-i1. במידה ונתון לנו פרופיל משתמש (יותר מפריט אחד כקלט) נגדיר pr(i2|u)=maxi1pr(i2|i1), כלומר, ההסתברות המקסימלית לדירוג סרט בהינתן סרט כלשהו שהמשתמש דירג. זכרו לסנן מרשימת ההמלצות פריטים שהמשתמש כבר ראה (דירג).
   1. הערה – רצוי לסנן זוגות פריטים שנצפו יחד מספר פעמים נמוך יחסית (נאמר, פחות מ-10), מאחר וחישוב ההסתברות המותנית עבורם אינו מדויק.
   2. שם האלגוריתם – CP.
2. ממשו מערכת מבוססת פריטים (כמו בסעיף הקודם) כאשר הדומות בין הפריטים משתמשת ב-Jaccard.
3. ממשו את השיטה ComputePrecisionRecall המחשבת את טכניקת ההשוואה Precision-Recall:
   1. הערכים מחושבים עבור אורך רשימת המלצות קבוע n.
   2. נבחר משתמש מה-test set.
   3. נשתמש באלגוריתם המלצה כלשהו לחשב רשימת סרטים מומלצים באורך n על פי רשימת הסרטים ב-train.
   4. כעת נחשב את הסטטיסטיקות הבאות:
      1. #tp מספר הסרטים שהומלצו ודורגו (מופיעים בחצי החבוי של רשימת הסרטים שהמשתמש דירג)
      2. #fp מספר הסרטים שהומלצו אך לא דורגו
      3. #fn מספר הסרטים שלא הומלצו אך המשתמש דירג אותם
      4. Precision=#tp/(#tp+#fp)
      5. Recall=#tp/(#tp+#fn)
   5. יש לחזור על הניסוי עבור כל ה-test set עבור רשימת המלצות באורך n ולדווח את הממוצע.
   6. שימו לב, מבחינה חישובית, קל יותר לבצע לכל משתמש וכל שיטה את הבדיקות עבור כל ערכי n הנדרשים – מבקשים רשימה באורך המקסימלי הנדרש, ומחשבים precision/recall עבור חיתוכים שונים של הרשימה.
   7. השיטה ComputePrecisionRecall מקבלת רשימת אלגוריתמים להשוואה, ורשימת ערכי אורך רשימה, ומחשבת precision ו-recall עבור כל האורכים. היא מחזירה לכן Dictionary הממפה אורך רשימה, לאלגוריתם, ל-precision ול-recall.
4. בצעו את ניסוי ה-Precision-recall עבור ערכים שונים של n (1,5,10,20,50(, והאלגוריתמים מסעיפים 1-5. יש לצייר עקומות של Precision-recall ב-Excel. הגישו עבור חלק זה גם את קובץ ה-Excel המכיל את ערכי ה-Precision-recall עבור ערכי ה-n השונים והאלגוריתמים השונים וגם את העקומות.

הגשה: את התרגיל יש להגיש ביום ראשון, ה-5 ביוני , עד השעה 23:59.

בהצלחה!