

# מודלים למסחר אלקטרוני 0960211

## תרגיל בית 1 – השפעה ברשתות

### רקע:

חברת בגדים יוקרתיים חדשה (Praducci) מעוניינת להגדיל את ערך המותג שלה ברשת החברתית "NoseBook" ולשם כך רוצה להתפרסם. לכן, בתור סטודנטים מומחים למסחר אלקטרוני, חברה זו פנתה אליכם וביקשה שתייעצו ותסייעו לה בפרסום.

הרשת החברתית "NoseBook" מאפשרת למשתמשים להעלות סרטונים (ובפרט, סרטונים בהם מראים את הבגדים החדשים שלהם). בכל פעם שמשתמש רוכש בגד, חבריו ברשת יכולים לראות זאת לאחר מכן. לצורך ביצוע המשימה קיבלתם מסד נתונים של חברת NoseBook המתאר את הקשרים החברתיים בין המשתמשים. כדי להתחיל את תהליך הפרסום, עליכם לבחור קבוצת משפיענים ראשונית (כמפורט בסעיף 'משימה'). לכל משתמש ברשת שיכול לשמש כמשפיען כזה יש "מחיר פרסום" (עלות שכירה). אם תבחרו לשכור את שירותיו במסגרת התקציב, הוא יקבל את המוצר ויחשוף אותו לחבריו ברשת, שעשויים לרכוש את המוצר בעקבות כך, בהסתברות שתתואר בהמשך. עלות השכירה של כל משפיען פוטנציאלי שונה, והיא נתונה לכם באחד הקבצים המצורפים (ראו פירוט בהמשך).

כמו שבודאי יצא לכם לחוות, לא כל חוויה עם מותגים היא חיובית. ברשת NoseBook קיימים משתמשים אשר חוו בעבר חוויה שלילית עם Praducci או מותגי יוקרה דומים (למשל, מוצר פגום, שירות לקוחות לקוי), והם משתפים את אכזבתם עם חבריהם. שיתוף חוויות אלו יוצר סנטימנט שלילי כלפי המותג. משתמשים אלו, המהווים מקור להשפעה שלילית, יכוננו "הייטרים" בתרגיל זה, והם יכולים להקטין את ההסתברות שמשתמש רגיל יקנה את מוצרי Praducci כפי שיתואר בהמשך.

## משימה:

חברת Praducci הקצתה תקציב של 1500 ש"ח לגיוס משפיענים ראשוניים בזמן  $t = 0$ . עליכם לבחור קבוצה התחלתית של משפיענים  $S_0$  מתוך כלל המשתמשים ברשת, כך שסך העלות שלהם לא תחרוג מהתקציב הנתון.

**הערה:** אסור לבחור משתמשים שהוגדרו מראש כ"הייטרים" (מקובץ haters.csv) כחלק מקבוצת המשפיענים הראשונית  $S_0$ .

כל משפיען שיבחר ישפיע על ההסתברות שחבריו ברשת החברתית יקנו את המוצר. לאחר מכן, ישנה הסתברות שמשתמש כלשהו יקנה את המוצר הקשורה למספר השכנים שלו שקנו את המוצר. הסתברות זאת תהיה תלויה גם בהייטרים ברשת. ישנם שישה סיבובי הדבקה כאלו, ובסופם נרצה שמספר המשתמשים שקנו את המוצר של Praducci יהיה כמה שיותר גדול.

**הערה:** אין צורך להכנס לשיקולים של "האם לחסוך בתקציב יוביל לעלייה בתוצאה שלנו". המטרה שלכם, כפי שתוגדר פורמלית בהמשך המסמך, היא שמספר המשתמשים שקנו את המוצר של Praducci יהיה כמה שיותר גדול. אין חשיבות לכמה מהתקציב המקורי נשאר לאחר בחירת המשפיענים.

## הייטרים:

לכל אחד מההייטרים  $a$  יש משקל ידוע מראש שמסומן ב  $w_a \in (0,1)$ . ככל שהמשקל הזה יותר גבוה, ההשפעה השלילית על הסתברות הקנייה תהיה יותר גבוהה. זהות ההייטרים ומשקליהם נתונים לכם בקובץ csv (ראו עוד בהמשך).

## הסתברות רכישה:

ההסתברות שמשתמש ירכוש בגד תלויה בחבריו ברשת NoseBook, ואילו מהחברים האלו הם הייטרים.

את הבגדים נשלח למשפיענים שבחרנו בזמן  $t = 0$ , ושכניהם (חבריהם) יוכלו לראות זאת רק בנקודת הזמן הבאה, כלומר ב  $t = 1$ . בזמנים  $t = 1 \dots 6$ , כל משתמש  $v$  בוחר אם לרכוש בגד של החברה לפי הסתברות ידועה מראש, באופן הבא.

- נסמן ב  $S_t$  את קבוצת המשתמשים שרכשו או קיבלו בגד של Praducci עד זמן  $t$ , **לא כולל**. נשים לב שאנטי-משפיענים לא יכולים להיכלל בקבוצה זאת.
- נסמן ב-  $a_1^v, \dots, a_k^v$  את השכנים ההיטרים של  $v$ , כאשר  $k$  זהו מספרם.

ההסתברות שצומת  $u \in N(v) \cap S_t$  (שכן של  $v$  שנדבק עד זמן  $t$  לא כולל) ידביק את  $v$  באיטרציה  $t$  הינה

$$P_{u \rightarrow v}^t = 0.2 \cdot \prod_{i=1, \dots, k} (1 - w_{a_i^v})$$

במילים אחרות, הסתברות זאת היא ההסתברות הבסיסית להדבקה בעזרת קשת  $(0.2)$ , עם הכפלה בקבוע קטן מ-1 שקטן ככל שיש יותר שכנים היטרים חזקים.

בכל סיבוב, כל אחד משכניו של  $v$  שהודבק בעבר מנסה להדביק אותו. לכן סה"כ, ההסתברות ש  $v$  יודבק באיטרציה  $t$  הינה ההסתברות המשלימה לכך שאף צומת שכן של  $v$  שאינו אנטי-משפיען הצליח להדביק אותו, כלומר:

$$\Pr(v \text{ is infected in } t) = 1 - \prod_{u \in N(v) \cap S_t} (1 - P_{u \rightarrow v}^t)$$

נשים לב שככל שמספר השכנים של  $v$  שקנו/קיבלו את המוצר יותר גדול, אז הסתברות זאת יותר גדולה. לעומת זאת, אם ל- $v$  יש הרבה שכנים היטרים עם משקל גבוה, אז הסתברות זו תקטן.

**הערה:** אם משתמש בחר שלא לקנות בגד בנקודת זמן מסוימת, הוא עדיין יכול לעשות זאת בנקודות הזמן הבאות – עפ"י ההסתברות המתאימה. בנוסף, שלא כמו במודל ה IC הרגיל שראינו בהרצאה, במודל הנוכחי אם  $u \in N(v) \cap S_t$  לא הצליח להדביק את  $v$  בזמן  $t$ , הוא יוכל לנסות שוב בזמן  $t + 1$  (ובזמנים הבאים) כל עוד  $v$  לא הודבק על ידי אחד משכניו האחרים.

## מטרה

במילים, המטרה שלכם היא למצוא את קבוצת המשפיענים ההתחלתית  $S_0$  שתביא למקסום התוחלת (הערך הממוצע הצפוי) של מספר המשתמשים הכולל אשר יחזיקו בפריט של Praducci בתום 6 סיבובי ההשפעה. פורמלית, עליכם למקסם את:

$$Objective = \mathbb{E} \left( \sum_{v \in V} I\{v \text{ has Praducci item after round 6} \mid S_0\} \right) \quad (1)$$

כאשר  $V$  היא קבוצת הקודקודים (משתמשים) בגרף  $G$  ו- $S_0$  זאת קבוצת המשפיענים ההתחלתיים שבחרתם. התוחלת נלקחת על פני הסתברויות הרכישה השונות בהתאם למשפיענים שבחרתם.

**שימו לב!** לנוחותכם, מצורף לקבצי התרגיל, קובץ קוד "Praducci\_simulation.py" המבצע סימולציה של המשימה כולה ומחזיר לכם את התוצאה שלכם. שימו לב כי התוצאה שהסימולציה מחזירה היא ריאליזציה יחידה של משתנה מקרי, ומטרתכם היא מקסום התוחלת של אותו משתנה מקרי.

## קבצים שנתונים לכם:

בתיקיית התרגיל שקיבלתם ישנם את הקבצים הבאים:

1. קובץ pdf זה עם הוראות התרגיל.
2. קובץ 123456789\_987654321.csv בו נבחרו משפיענים אקראיים בשביל הדגמת הסימולציה (ראו את הפריט הבא). קובץ זה עומד בפורמט בו אתם מתבקשים להגיש את המשפיענים שלכם.
3. קובץ קוד "Praducci\_simulation.py" המבצע סימולציה של המשימה כולה ומחזיר לכם ריאליזציה התוצאה שלכם. לקובץ זה 3 מטרות:
  - a. ביצוע סימולציה על המשפיענים מהקובץ 123456789\_987654321.csv והדגמת מבנה התהליך בפונקציה `simulate_influence`.
  - b. הדגמה של כיצד תתבצע קריאת המשפיענים שבחרתם באמצעות הפונקציה `read_influencers_from_csv`. בפונקציה זו יש הרבה בדיקות תקינות על קובץ ההגשה שלכם כדי לעזור לכם להימנע מטעויות, אך אינה מתיימרת לתפוס את כל הטעויות האפשרויות בפורמט ההגשה. פונקציה זו אינה משנה את קובץ ההגשה ולכן "בטוחה לשימוש" עם קובץ ההגשה הסופי שלכם (ראו עוד בהמשך).
  - c. פונקציה בשם `submit_influencers` שמקבלת רשימת משפיענים ואת תעודות הזהות שלכם, בודקת שהמשפיענים שבחרתם עומדת בתקציב הרצוי ואינם כוללים הייטרים ויוצרת בשבילכם קובץ csv העומד בהנחיות ההגשה. אנחנו מעודדים אותכם להשתמש בפונקציה זאת כדי ליצור את קובץ ההגשה שלכם (ראו על זה עוד בהמשך).

4. קובץ csv בשם NoseBook\_friendships.csv המכיל את רשימת השכנויות בגרף. לנוחותיכם, מודגם בקובץ Praducci\_simulation.py כיצד ליצור אובייקט networkx מרשימה זו וכיצד לעבוד עם האובייקט הזה כדי לעשות סימולציה של התהליך שהוגדר בשיעורי בית אלה.
5. קובץ csv בשם haters.csv המכיל את זהות ההיטרים בגרף ואת משקלם. לנוחותיכם, מודגם בקובץ Praducci\_simulation.py כיצד להשתמש במשקלים אלו כדי לעשות סימולציה של התהליך שהוגדר בשיעורי בית אלה.
6. קובץ csv בשם costs.csv המכיל את מחירי המשפיענים בגרף. לנוחותיכם, מודגם בקובץ Praducci\_simulation.py כיצד לקרוא את המשקלים האלו ולבדוק שהתקציב שבחרתם לא עולה על זה שניתן לכם.

## הגשה

עליכם להגיש שלושה קבצים:

1. קובץ csv יחיד המכיל את המשפיעים שבחרתם. הפורמט המבוקש הינו הפורמט המוחזר על ידי הפונקציה submit\_influencers. שימו לב – מנסיון של שנים קודמות, התעסקות ידנית עם הקובץ אותו אתם מגישים (כמו פתיחתו באקסל או תוכנות דומות) עלולה לפגוע בפורמט הדרוש להגשה. קובץ בפורמט לא נכון לא יעבור בקובץ הבדיקה שלנו וההגשה תקבל אוטומטית ציון 0. כדי להמנע ממקרים כאלו, השתמשו בפונקציה submit\_influencers כדי ליצור את הקובץ ואל תערכו אותו לאחר מכן.
2. קובץ הסבר בפורמט pdf, ששמו ID1\_ID2.pdf, באורך של **לכל היותר עמוד**, הכולל את ההסבר שלכם לבחירת המשפיענים. הקובץ צריך לפרט את הגישה שלכם לפתרון הבעיה. המטרה של הקובץ הזה היא לנסח בבהירות את תהליך ההתמודדות שלכם עם המשימה.
3. בנוסף עליכם להגיש את הקוד שבעזרתו פתרתם את התרגיל - יש להגיש קובץ py יחיד שכותרתו ID1\_ID2.py.

יש לכלול את שלושת הקבצים הללו בקובץ zip ששמו ID1\_ID2.zip. אין להוסיף תיקיות נוספות בתוך ה-zip!

על כל הסטודנטים לעמוד בהוראות הגשה אלו. בפרט, סטודנטים המשתמשים ב macOS שמייצר בדרך כלל תיקייה נוספת כאשר דוחסים קבצים לתיקייה, מוזמנים לעיין [בקישור הבא](#) כדי להימנע מיצירת התיקייה הנוספת.

**הערה:** במהלך תרגיל זה, ניתן לייבא את הספריות הבאות בלבד:

```
import numpy as np  
  
import networkx as nx  
  
import random  
  
import pandas as pd  
  
import csv  
  
import os
```

יבוא של ספריות אחרות יגרור פסילה של תרגיל הבית!

## הערכה וציון

הערכת הסטודנטים תתבצע לפי פונקציית המטרה במשוואה (1). כיוון שאין ביכולתנו לחשב את התוחלת במפורש נבצע אלפי סימולציות עד זמן  $t = 6$  עבור כל קבוצה, כדי למצוא רווח סמך קטן ככל האפשר. ציון הקבוצה יהיה מבוסס על אותו רווח סמך, כאשר הציונים יהיו על בסיס תחרות בין הגשות הסטודנטים.

**שאלה:** אם הציון שלנו מבוסס על משתנה מקרי, איך נוכל להבטיח הצלחה?

**תשובה:** ראשית, הגשה שתשיג תוצאה של יותר מ-1000 (בריצת הבדיקה שנבצע אצלנו) תזכה את המגישים בציון 75 לפחות. במקרה והגשתכם לא עברה את הסף, לא יתקבלו ערעורים בטענה שבריצת הקוד שלכם כן הצלחתם לעבור אותו.

שנית, פונקציית הרווח במשוואה (1) אינה משתנה מקרי. מדובר בערך **קבוע** התלוי בבחירתכם במשפיענים, שאמנם לא קל לחשבו במדויק (כיוון שהוא מייצג את תוחלת ההשפעה המתקבלת), אך הוא אינו אקראי. מהות התרגיל היא לבחור את המשפיענים בצורה נכונה וטובה יותר מאחרים. לכן, יצירתיות, השקעה ויעילות יבטיחו לכם הצלחה.

