**אלגוריתמים ברשתות**

**בניית מערכת המלצה – מסמך מסכם**

**Olist E-COMMERCE Dataset**

**מגישים:**

**משה מנע – 313266223**

**יואב הראל – 303168231**

**סתיו הרצוג - 208852327**

**מבוא:**

כיום יש עלייה בתחום קניות האונליין ולכן חברות העוסקות בתחום מעוניינות להגדיל את המכירות שלהן על ידי אופטימזציית תהליך רכישת המוצרים של הלקוחות.

לכן אנו באים לבנותמערכת המלצה ללקוחות שקונים בחנות אונליין על סמך מוצרים פוטנציאלים אשר יכולים לעניין אותם. בהינתן מוצר מהרשת שהלקוח מתעניין בו, נרצה לדעת על איזה מוצרים נוספים להמליץ ללקוח. נבנה רשת מוצרים אשר מקושרים ביניהם במידה ונמכרו בעגלה זהה.

**הנתונים:**

סט הנתונים מכיל הזמנות אשר נאספו מפלטפורמת חנויות האונליין הברזילאית Olist.

הנתונים מכילים 100 אלף הזמנות אשר נאספו בשנים 2016-2018 ממספר חנויות שונות.

הנתונים נלקחו מאתר  [Kaggle](https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce?select=olist_products_dataset.csv), ונעשה שימוש ב- 2 טבלאות הבאות:

* A screenshot of a computer

  Description automatically generatedמוצר בהזמנה: olist\_order\_items\_dataset
* מוצרים: olist\_products\_dataset

A screenshot of a graph

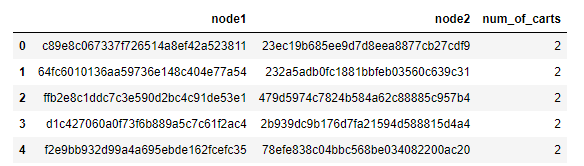
Description automatically generated

**עיבוד הנתונים:**

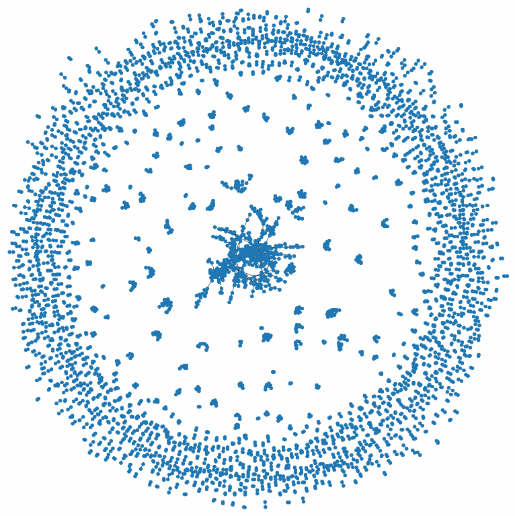
**עיבוד טבלת מוצרים:**

הורדת עמודות לא רלוונטיות והוספת תרגום קטגוריית המוצר באנגלית אשר נמצא בטבלת עזר.

**עיבוד טבלת הזמנות ויצירת טבלת קשרים:**

1. ניקוי כפילויות של מוצרים בעגלה, כדי להשאיר רק עגלות עם יותר ממוצר אחד (אך לא את אותו המוצר).
2. יצירת קשרים בין מוצרים באותה העגלה
3. הוספת משקלים לקשתות על פי כמות הפעמים שאותם מוצרים נקנו ביחד (קשתות כפולות)
4. יצירת טבלה חדשה product\_to\_product.csv אשר תשמש לבניית הרשת.

* weight = num\_of\_carts

**הרשת:**

צמתים – מוצרים

קשתות – מוצרים אשר הופיעו ביחד באותה העגלה.

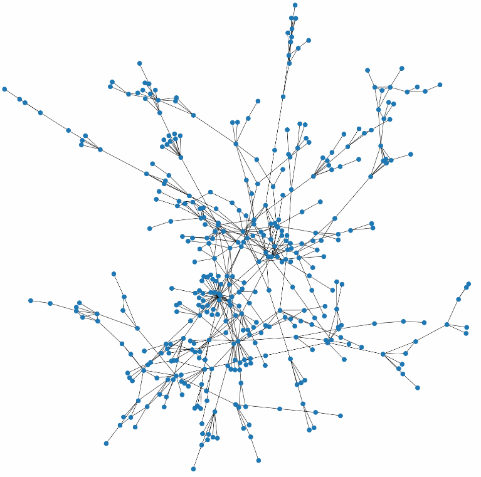
(מבנה הרשת)

**חקירת הרשת:**

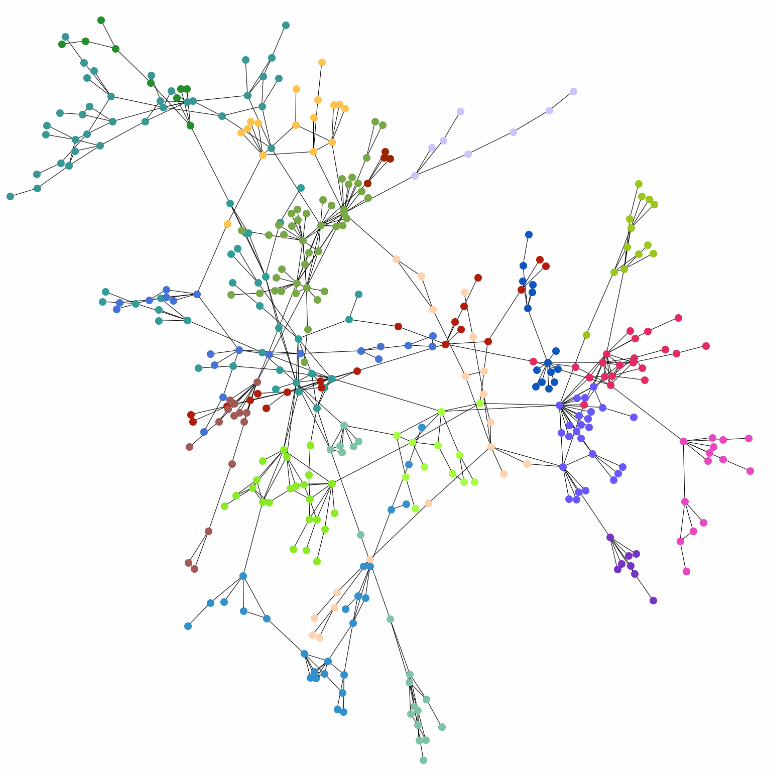
הרשת מכילה 4885 צמתים (מוצרים) ו4058 קשתות.

קיימים 1652 רכיבי קשירות בגרף ורובם קטנים מאוד ולא תורמים מידע משמעותי למערכת ההמלצה.

ולכן בחרנו להתמקד ברכיב הקשירות הגדול ביותר בגרף בעל 398 צמתים (מוצרים).

****

(מבנה הGCC)

**חקר קהילות:**

A grid of numbers and letters

Description automatically generatedהשוואת קהילות ע"י שימוש באלגוריתמים של Louvain ו Girvan newman.

)קהילות לפי (Louvain

מדד המודולריות בחלוקה לקהילות ע"י Louvain גבוהה יותר וכן חלוקת הקהילות  
הייתה מדוייקת יותר מבחינת חלוקת הקטגוריות של המוצרים. ולכן האלגוריתם הנבחר לחלוקת קהילות הוא Louvain.

**אלגוריתם מערכת ההמלצה:**

מערכת ההמלצה מקבלת id של מוצר מהרשת:

1. מחשבת מיהו המוצר מתוך הקהילה שלו עם ה betweenness centrality הגבוהה ביותר.
2. מחשבת מיהו המוצר בעל מדד ה Jaccard coefficient הגבוהה ביותר בקהילה שלו.
3. מחשבת מיהו המוצר בעל מדד ה Jaccard coefficient הגבוהה ביותר בגרף כולו.
4. מחזירה קבוצת (1-3) מוצרים מומלצים.

הסבר מדדים:

betweenness centrality – מדד זה נבחר כיוון שהוא מייצג את המוצר בתוך הקהילה של המוצר שהגיע כקלט למערכת אשר מחבר בין הכי מוצרים.

Jaccard coefficient – מדד זה נבחר על מנת למצוא מוצר בעל הדמיון הגבוהה ביותר עם המוצר שהגיע כקלט למערכת גם ברמת הקהילה וגם ברמת הרשת.

**הערכת המערכת:**

שיטה ראשונה - מדדים:

- אם כל שלושת המדדים המליצו על אותו המוצר איכות ההמלצה היא חזקה.

- אחרת ההמלצה עדיין בוחרת במוצרים רלוונטים אך פחות מובהקת.

שיטה שנייה – קטגורייה:

- להתסכל על הקטגוריות של המוצרים שיצאו בהמלצה ולבדוק האם יש קשר בין הקטגוריות, לקטגוריה של המוצר המדובר.

**A screenshot of a computer code

Description automatically generated**

* פלט מערכת ההמלצה להערכת שיטות 1 ו2.

שיטה שלישית – עגלת המבחן:

הוצאה של עגלה ובדיקה האם מערכת ההמלצה תמליץ על מוצרים שנקנו יחד בעגלה זו.

לפני בניית הרשת הוצאנו עגלה עם id = 1e0cc5a03c811818d8f95ba53e014589 אשר מכילה 2 מוצרים:

1. Id מוצר ראשון - eebbed5ed3b134eceb717496c47652ba

2. Id מוצר שני - a4788cb24856965c36a24e339b605899

הסבר בחירת עגלה:

עגלה המכילה לפחות 2 מוצרים אשר קיימים ברכיב הקשירות הגדול ביותר. בנוסף כל מוצר נקנה גם כן בעגלות נוספות על מנת שיופיע ברשת לאחר הוצאת העגלה.

תוצאות שיטת הערכה שלישית:

כשאר הוזן למערכת המוצר הראשון מהעגלה ניתן לראות שאכן המוצר שהנוסף שהיה איתו בעגלה הומלץ על ידי המערכת ולהפך.

מוצר ראשון:

A close-up of a computer code

Description automatically generated

מוצר שני:

A close-up of a computer screen

Description automatically generated

**מסקנות**

בניית הרשת על ידי חיבור מוצרים אשר נקנו ביחד באותה העגלה (2 מוצרים ומעלה), הניבה תוצאות טובות וניתן לראות זאת בתוצאות לעיל שאכן יש קשר בין המוצרים המומלצים אשר מגיעים מאותה הקטגוריה של מוצר הקלט. בנוסף, השימוש במדד Jaccard ברמת הרשת כולה מניב לעיתים תוצאות טובות יותר מאשר מדד Jaccard ברמת הקהילה. לסיכום, אנו מניחים שבעזרת נתונים נוספים היה ניתן לקבל רשת גדולה ומקושרת יותר וכך ניתן היה להגיע לתוצאות מדוייקות ואמינות יותר.