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 לבנות מערכת 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 , ונעשה שימוש ב- 2 טבלאות הבאות:

- מוצר בהזמנה: olist_order_items_dataset

	order_id	order_item_id	product_id	seller_id	shipping_limit_date	price	freigl
0	00010242fe8c5a6d1ba2dd792cb16214	1	4244733e06e7ecb4970a6e2683c13e61	48436dade18ac8b2bce089ec2a041202	9/19/2017 9:45	58.90	
1	00018f77f2f0320c557190d7a144bdd3	1	e5f2d52b802189ee658865ca93d83a8f	dd7ddc04e1b6c2c614352b383efe2d36	5/3/2017 11:05	239.90	
2	000229ec398224ef6ca0657da4fc703e	1	c777355d18b72b67abbeef9df44fd0fd	5b51032eddd242adc84c38acab88f23d	1/18/2018 14:48	199.00	
3	00024acbcdf0a6daa1e931b038114c75	1	7634da152a4610f1595efa32f14722fc	9d7a1d34a5052409006425275ba1c2b4	8/15/2018 10:10	12.99	
4	00042b26cf59d7ce69dfabb4e55b4fd9	1	ac6c3623068f30de03045865e4e10089	df560393f3a51e74553ab94004ba5c87	2/13/2017 13:57	199.90	

- מוצרים: olist_products_dataset

	product_id	product_category_name	product_name_lenght	product_description_lenght	product_photos_qty	product_weight_g	produc
0	1e9e8ef04dbcff4541ed26657ea517e5	perfumaria	40.0	287.0	1.0	225.0	
1	3aa071139cb16b67ca9e5dea641aaa2f	artes	44.0	276.0	1.0	1000.0	
2	96bd76ec8810374ed1b65e291975717f	esporte_lazer	46.0	250.0	1.0	154.0	
3	cef67bcfe19066a932b7673e239eb23d	bebes	27.0	261.0	1.0	371.0	
4	9dc1a7de274444849c219cff195d0b71	utilidades_domesticas	37.0	402.0	4.0	625.0	

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.

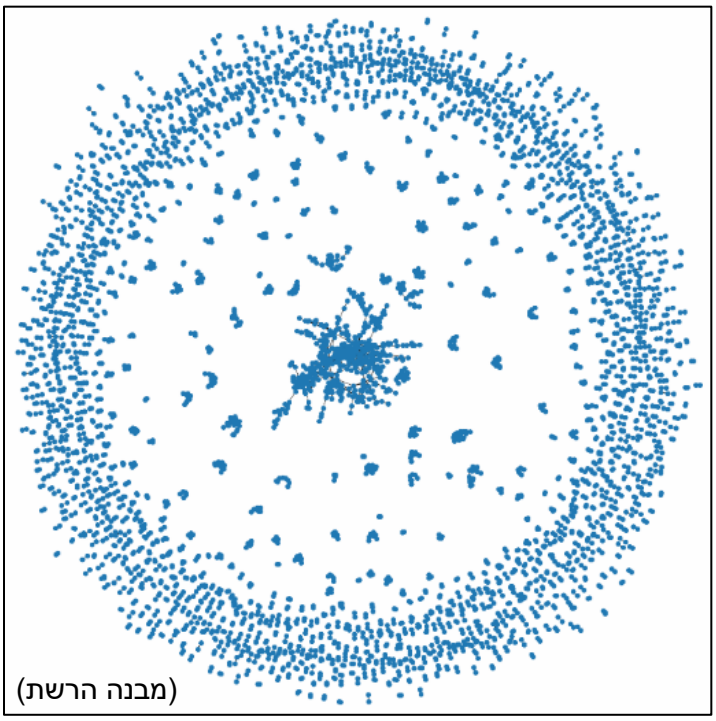
	node1	node2	num_of_carts
0	c89e8c067337f726514a8ef42a523811	23ec19b685ee9d7d8eea8877cb27cdf9	2
1	64fc6010136aa59736e148c404e77a54	232a5adb0fc1881bbfeb03560c639c31	2
2	ffb2e8c1ddc7c3e590d2bc4c91de53e1	479d5974c7824b584a62c88885c957b4	2
3	d1c427060a0f73f6b889a5c7c61f2ac4	2b939dc9b176d7fa21594d588815d4a4	2
4	f2e9bb932d99a4a695ebde162fcefc35	78efe838c04bbc568be034082200ac20	2

- weight = num_of_carts

הרשת:

צמתים – מוצרים

קשתות – מוצרים אשר הופיעו ביחד באותה העגלה.



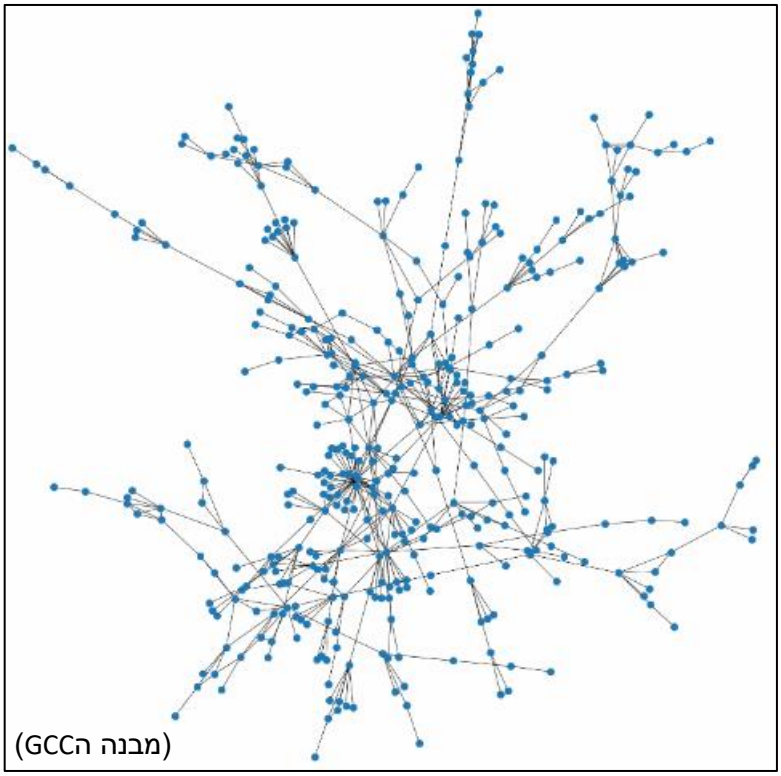
(מבנה הרשת)

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).

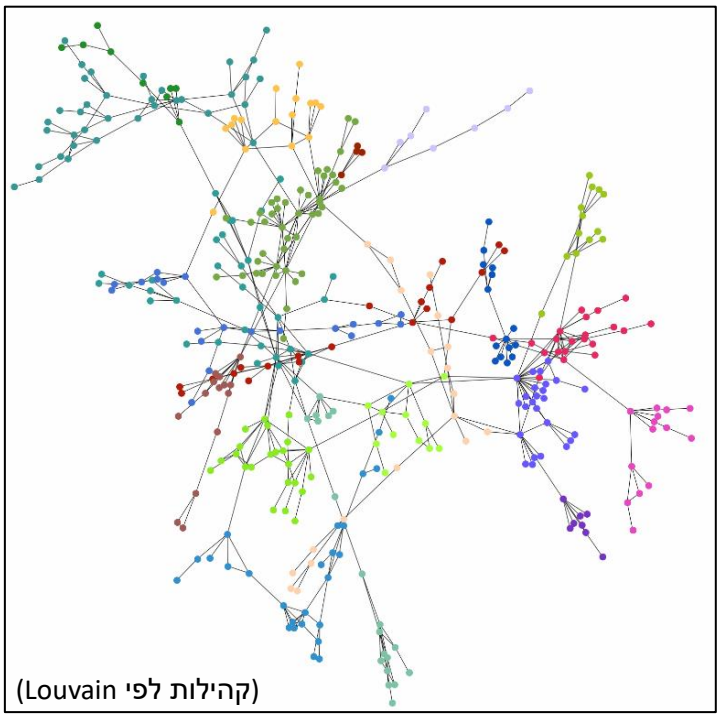


(מבנה הGCC)

חקר קהילות:

השוואת קהילות ע"י שימוש באלגוריתמים של Louvain ו Girvan newman .

	Girvan Newman	Louvain
Modularity	0.64	0.86
Number of communities	3	21



(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.

<u>הסבר מדדים:</u>

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:

<u>שיטה ראשונה - מדדים:</u>

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

- אחרת ההמלצה עדיין בוחרת במוצרים רלוונטים אך פחות מובהקת.

<u>שיטה שנייה – קטגורייה:</u>

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

```
Product Id: 52b668edc0d0c20cd9319bfda4019597, category: furniture_decor

Max betweenes id: 3e5201fe0d1ba474d9b90152c83c706c, score: 0.5473684210526316, category: bed_bath_table
Max jaccard by community id: cd3de1984e1a77b441e1b39b8e334330, score: 0.2, category: furniture_decor
Max jaccard by gcc id: cd3de1984e1a77b441e1b39b8e334330, score: 0.2, category: furniture_decor
```

- פלט מערכת ההמלצה להערכת שיטות 1 ו2.

<u>שיטה שלישית – עגלת המבחן:</u>

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 - 99a4788cb24856965c36a24e339b6058

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:

```
Product Id: eebbed5ed3b134eceb717496c47652ba, category: bed_bath_table

Max betweenes id: 99a4788cb24856965c36a24e339b6058, score: 0.83, category: bed_bath_table
Max jaccard by community id: 3458b4c1fcbe46e2eedb48e00960a60e, score: 0.33, category: bed_bath_table
Max jaccard by gcc id: 4f88323d03ffaf090b8fb0116b33c95e, score: 0.5, category: bed_bath_table

Recommended items: {'4f88323d03ffaf090b8fb0116b33c95e', '3458b4c1fcbe46e2eedb48e00960a60e', '99a4788cb24856965c36a24e339b6058'}
```

מוצר שני:

```
Product Id: 99a4788cb24856965c36a24e339b6058, category: bed_bath_table

Max betweenes id: f2e53dd1670f3c376518263b3f71424d, score: 0.53, category: bed_bath_table
Max jaccard by community id: eebbed5ed3b134eceb717496c47652ba, score: 0.08, category: bed_bath_table
Max jaccard by gcc id: 592962829d5a715304344e656e39108a, score: 0.12, category: bed_bath_table

Recommended items: {'eebbed5ed3b134eceb717496c47652ba', '592962829d5a715304344e656e39108a', 'f2e53dd1670f3c376518263b3f71424d'}
```

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.