

## מודלים במסחר אלקטרוני - פתרון תרגילי כיתה

## אלגוריתם Market Basket

## שאלה 1

לשם הרצת A-priori צריך בשלב הראשון לספור את המופעים של פריטים בודדים לשם כך נחוצים 4,000,000 תאי זיכרון (4 לכל פריט). מידע זה מצטמצם ל-bitmap בגודל 125,000 בתים ( $\frac{1000000}{8}$ ). בשלב השני סופרים זוגות של פריטים תדירים (25,000 פריטים -  $25000^2$  זוגות), סה"כ 2,500,000,000 בתים.

כלומר, בשלב השני (ובסך הכל) נזדקק ל- 2,500,125,000 בתים.

## שאלה 2

נסמן את מספר התאים בטבלת ה-HASH ב- $x$ .

כדי שהתאים שאינם מכילים זוגות תדירים לא יהיו נפוצים נדרוש כי  $\frac{1,000,000,000}{x} < 10,000$  ומכאן  $x > 100,000$ .

כמו בשאלה 1, אנחנו זקוקים בשלב הראשון ל- 4,000,000 תאי זיכרון. כמות הזיכרון הנדרשת עבור הטבלה היא, אם כן,  $4x$ . לכן, נדרשים בסך הכל  $4000000 + 4x$  תאי זיכרון בשלב הראשון.

בשלב השני נדרשים 125,000 בתים עבור ה-bitmap של A-priori ועוד  $x/8$  בתים bitmap של ה-HASH. בשלב הראשון בדיוק 10,000 תאים ב-HASH יהיו נפוצים. מתוך ביליון הזוגות, רק מחצית מהם ייספרו בשל ה-A-priori ומתוך אלה, רק  $\frac{10,000}{x}$  מהם יפלו לתאים נפוצים. כלומר, נספור רק  $\frac{5,000,000,000}{x}$  זוגות. לכל זוג יוקצו 4 בתים ולכן בסך הכל יידרשו  $\frac{20,000,000,000}{x}$  בתים עבור הספירה. נסכם ונקבל:

$$125,000 + \frac{x}{8} + \frac{20,000,000,000}{x}$$

משום שאותו זיכרון מנוצל בשני השלבים, ניצול אופטימלי ישווה את צריכת הזיכרון בשניהם. נקבל את המשוואה

$$125,000 + \frac{x}{8} + \frac{20,000,000,000}{x} = 4000000 + 4x$$

נפתור ונקבל  $x = 5135 < 100,000$ . לכן נקבע  $x = 100,000$  ונקבל שכמות הזיכרון הדרושה היא 4,400,000 בתים.

## שאלת LSH

סעיף א'

במקרה זה יש לנו 4 רצועות של 2 ערכים כל אחת. זוג עמודות יהיה מועמד אם יש לו לפחות רצועה אחת שדומה בכל הערכים. בהנתן מידת דמיון  $s$ , ההסתברות של זוג להיות מועמד היא:

$$p = 1 - (1 - s^2)^4$$

מבין 1000 הזוגות עם מידת דמיון של 50%, נמצא שהם דומים בהסתברות 68.4% ולכן ימצאו 683 זוגות מועמדים.

מבין 1,000,000 הזוגות עם מידת דמיון של 20%, נמצא שהם דומים בהסתברות 15.065% ולכן ימצאו 150,650 זוגות מועמדים. אם נצרף את הזוגות עם מידת דמיון 50% נקבל בסך הכל 151,333 זוגות דומים.

סעיף ב'

במקרה זה יש לנו 2 רצועות עם 4 ערכים כל אחת, ולכן ההסתברות של זוג להיות מועמד היא:

$$p = 1 - (1 - s^4)^2$$

מבין 1000 הזוגות עם מידת דמיון של 50%, נמצא שהם דומים בהסתברות 12.1% ולכן ימצאו 121 זוגות מועמדים.

מבין 1,000,000 הזוגות עם מידת דמיון של 20%, נמצא שהם דומים בהסתברות 0.3197% ולכן ימצאו 3,197 זוגות מועמדים. אם נצרף את הזוגות עם מידת דמיון 50% נקבל בסך הכל 3,318 זוגות דומים.

## Clustering

Jaccard	Cosine	סה"כ	משותף	וקטרים
2/7	$2/2\sqrt{5}$	7	2	B A
3/7	3/5	7	3	C A
2/7	$2/2\sqrt{5}$	7	2	D A
3/7	3/5	7	3	E A
1/8	$1/2\sqrt{5}$	8	1	C B
2/6	2/4	6	2	D B
3/6	$3/2\sqrt{5}$	6	3	E B
2/7	$2/2\sqrt{5}$	7	2	D C
1/9	1/5	9	1	E C
2/7	$2/2\sqrt{5}$	7	2	E D