

## מבחן לדוגמא LEVERAGING BIG DATA

סמסטר א 2014

המרצים: עידית כהן, טובה מילוא, עמוס פיאט, חיים קפלן

אין מגבלה לגבי חומר עזר

יש לפתור ארבע מתוך חמש השאלות, כל שאלה בעמוד אחד במחברת, כתוב תשובות קצרות מנוסחות היטב. אין צורך לחזור על הוכחות שניתנו בכיתה. כל שאלה 25 נקודות.

1. (25 נקודות) נתון האלגוריתם הבא, שפועל על stream של נקודות במרחב מטרי, המרחק בין כל זוג נקודות  $< 1$ :

$T \leftarrow \text{emptyset}$

$R \leftarrow 1$

Repeat forever:

While  $|T| \leq k$

do

Get next point  $p$

If  $\text{dist}(T,p) > R$  then add  $p$  to  $T$

enddo

*( $\alpha$ ) All pairwise distances in  $T > R$ ,  $|T|=k+1$  points,  
Max  $\text{dist}(T,q) < 3R$  for all  $q$  up to now*

$R \leftarrow (3/2)R$

$T' \leftarrow \text{emptyset}$

While(Exists  $q$  in  $T$  such that ( $\text{dist}(q,T') > R$  or  $T'=\text{emptyset}$ )) then add  $q$  to  $T'$

$T \leftarrow T'$

End of Repeat

Let  $OPT$  be the minimum, over all  $k$  centers, of the maximal distance between a point and its center

א. הוכח כי הטענה המסומנת ב-  $\alpha$  נכונה בכל פעם כשהלולאה לפני  $\alpha$  נגמרת.

ב. הוכח כי ב-  $\alpha$  :  $R/2 < OPT$  ולכן  $\text{Max dist}(T,q) < 6OPT$ .

ג. הוכח כי תמיד נכון ש  $\text{Max dist}(T,q) < 9OPT$ .

2. (25 נקודות) נתון DATA של יחס FOLLOW מהרשת החברתית TWITTER.

$FOLLOW(x, y)$  אם אדם  $x$  מקבל את כל הציורים של אדם  $y$ . ( $x$  עוקב אחרי  $y$ )

נגדיר את הקבוצות הבאות ביחס לקבוצת אנשים  $X$ :

- $Q_1(X)$ : היא קבוצת האנשים ב- $X$  או העוקבים (FOLLOW) אחרי לפחות אדם אחד מ- $X$ :  $Q_1(X) = X \cup \{y \mid \exists z \in X \text{ such that } FOLLOW(y, z)\}$
- $Q_2(X)$ : היא קבוצת האנשים ב- $X$ , העוקבים אחרי אדם מ- $X$  או עוקבים אחרי אדם העוקב אחרי אדם מ- $X$ :

$$Q_2(X) = Q_1(X) \cup \{y \mid \exists z \in X, \exists w \text{ such that } FOLLOW(y, w) \text{ and } FOLLOW(w, z)\}$$

שאלתות אלו מודדות את היקף ההשפעה של קבוצת האנשים  $X$ .

- תארי/י כיצד לחשב את  $|Q_2(x)|$  לכל אדם בודד  $x$  ע"י פלטפורמת MAPREDUCE הגדרי/י את פונקציות ה-MAP וה-REDUCE בכל איטרציה.
- ב. בטא/י את מספר הזוגות (KEY, VALUE) הכולל המיוצר בחישוב בסעיף הקודם ואת המספר המכסימלי של הזוגות עם אותו KEY (הממופים לאותו reducer).
- ג. תאר SKETCH  $s(x)$  בגודל  $O(k)$  לכל אדם  $x$ . על ה-SKETCH למלא את הדרישות הבאות

- ניתן לחשבו לכל האנשים בזמן פרופורציונאלי ל- $k$  כפול גודל היחס FOLLOW.
- בהנתן קבוצה  $X$  וה SKETCHES  $s(x)$  לכל אדם  $x \in X$  נוכל לאמוד את  $|Q_2(X)|$  עם

$$O(|X|k \log(k)) \text{ ומספר פעולות } O\left(\frac{1}{\sqrt{k}}\right) \text{ Coefficient of Variation (CV) לכל היותר}$$

לכל היותר.

תאר את ה-SKETCHES וכיצד יתבצעו החישובים ב-1 וב-2.

3. (25 נקודות) מפעיל של שרות ענן מחזיק שני עותקים של ווקטור מסויים  $b$  ממיימד  $n$ . העותקים מוחזקים ב DATA CENTERS מרוחקים גאוגרפית. הווקטור מעודכן ע"י עדכונים מהצורה  $(i, x)$  שמשמעותם שהכניסה ה- $i$  בווקטור משתנה ב- $x$ . כלומר  $b_i \leftarrow b_i + x$ . אותם עדכונים מופעלים על שני העותקים אבל לא בדיוק בו-זמנית.

א. מדי פעם ברצוננו לבדוק אם העותקים שונים ביותר משתי כניסות. עלינו לבצע את המשימה ע"י שליחת מיידע לשרת מרכזי וברצוננו לשלוח מעט מאד מיידע ( $n \ll$  ערכים) ברשותך Linear Sketch לפרדיקט AtMost2? : הפרדיקט מופעל על ווקטור  $b$  ממיימד  $n$  ומחזיר True אם"ם לכל היותר שתי כניסות בווקטור אינן 0. תארי/י כיצד תבצע/י את הבדיקה על ידי כלי זה.

ב. עתה הוטלה עליך משימה נוספת. כאשר שני העותקים שונים בדיוק בכניסה אחת, על השרת המרכזי לדעת (בסכוי לפחות 60% לבדיקה) את הכניסה בה הם שונים ואת ההפרש. באילו מהכלים שלמדנו תוכל/י להשתמש? הסבר כיצד תעשה זאת בצורה יעילה (מבחינת כמות המיידע הנשלחת לשרת המרכזי)

ג. עתה עלייך לבנות Linear Sketch למימוש AtMost2?. ה-SKETCH הוא ווקטור במימד  $3k$  ל- $k$  טבעי ולבחירתך שתי אפשרויות של פונקציות HASH (הנח/י שהן perfectly random)

- A.  $k$  פונקציות  $h_i: \{1, \dots, n\} \rightarrow \{0, 1, 2\}$  לכל  $i = 1, \dots, k$  ולכל  $t = 0, 1, 2$  ה SKETCH יכול את  $\sum_{j|h_i(j)=t} b_j$ .
- B. פונקצייה בודדה  $g: \{1, \dots, n\} \rightarrow \{0, \dots, 3k - 1\}$  וה-SKETCH יכול לכל  $t = 0, \dots, 3k - 1$  את  $\sum_{j|g(j)=t} b_j$

לצורך שאלה זו הניח/י שאנו מחשבים את הפרדיקט רק כאשר כל הכניסות בווקטור הקלט אחרי העדכון האחרון הן אי-שליליות.

1. לכל אפשרות, תארי/י את המבנה של המטריצה הכופלת את ווקטור הקלט בכדי לקבל את ה-SKETCH המתאים.
2. כיצד תעדכנ/י את ה-SKETCH בהינתן עדכון מהצורה  $b_i \leftarrow b_i + x$  לכניסה  $i$  של וקטור הקלט?
3. מהו מספר הפעולות ה HASH ומספר הכניסות ב-SKETCH שעלינו לשנות בכל עדכון בכל אפשרות?
4. כיצד תעריכ/י AtMost2? על סמך ה SKETCH בכל אפשרות? מה ההסתברות לכל סוג של טעות (False positive/negative)?
5. מה היתרון של אפשרות A על B ומה היתרון של אפשרות B על A?

אם ברצונך למזער את הסכוי המקסימלי לטעות (על פני כל הקלטים), באיזו אפשרות תבחר? נמק/י. שימ/י לב שיתכן והתשובה תלויה בבחירה של- $k$ .

4. (25 נקודות) נתון מסד נתונים המכיל  $n$  (הנח כי  $n$  מאד גדול) תמונות. כל תמונה מיוצגת על ידי תת קבוצה מתוך 256 מאפיינים הקיימים בה. מודדים מרחק בין תמונות  $p_1$  ו- $p_2$  על ידי

$$d(p_1, p_2) = 1 - \text{jaccard}(p_1, p_2)$$

- תכנן מבנה נתונים שבהינתן תמונת שאילתה  $q$  יאפשר למצוא בהסתברות  $1/2$  תמונה  $p$  באוסף כך ש- $\text{jaccard}(p, q) \geq 8/10$  אם ישנה תמונה  $x$  באוסף כך ש- $\text{jaccard}(x, q) \geq 9/10$ . תאר (לפי הסעיפים הבאים) מבנה נתונים שעונה על שאילתה כזו בזמן  $O(\sqrt{n})$  (הזנח גורמים לוגריתמיים והשתמש בקרוב  $x \approx \log(1-x)$  כאשר  $x$  קטן).
- א. תאר את המבנה עצמו כיצד נאחסן את המבנה בזכרון כך שלא יצרוך יותר מידי מקום? כמה מקום יתפוס המבנה?
- ב. תאר את אלגוריתם השאילתה במדויק.
- ג. נתח את זמן הריצה של אלגוריתם השאילתה.
- ד. הוכח את נכונות האלגוריתם.

5. (25 נקודות) מוקצה על הדיסק מטריצה  $A$  עם  $n$  שורות ו- $n$  עמודות כאשר  $n$  גדול לפחות פי 100 מגודל הזיכרון  $M$ . המטריצה מאוכסנת על הדיסק "לפי שורות" כלומר ב- $n/B$  הבלוקים הראשונים מתוך הבלוקים המכילים את המטריצה מופיעים האיברים שבשורה הראשונה ב- $n/B$  הבלוקים הבאים נמצאים האיברים שבשורה השנייה וכו' ( $B$  הוא גודל הבלוק או באופן מדויק יותר מספר האיברים מהמטריצה שנכנסים בבלוק). בהינתן וקטור  $x$  (עמודה) מבצעים את החישוב  $y=Ax$  באופן הבא

```
For r = 1 to n do
  For c = 1 to n do
     $y[r] = y[r] + A[r,c]*x[c]$ 
```

הנח כי אברי וקטורי שורה או עמודה מאוכסנים בצורה רציפה על הדיסק.

א. האם האלגוריתם הוא cache oblivious? נמק.

ב. כמה קריאות של בלוקים לזיכרון וכתיבות של בלוקים לדיסק יתבצעו במהלך הריצה? נמק.

ג. עתה נבצע את החישוב  $y=x^t A$  באופן הבא

```
For c = 1 to n do
  For r = 1 to n do
     $y[c] = y[c] + x[r]*A[r,c]$ 
```

כמה קריאות של בלוקים לזיכרון וכתיבות של בלוקים לדיסק יתבצעו במהלך חישוב זה? נמק

ד. האם תוכל להציע אלגוריתם יעיל יותר מבחינת מספר פעולות ה- $I/O$  שמבצע את החישוב שבסעיף ג? תארו במילים ונתח את מספר פעולות ה- $I/O$  שהוא מבצע. הנח כי  $M=\Omega(B^2)$ .