236363 – מסדי נתונים, אביב 2023

תרגיל בית מס' 3

נושאי התרגיל: תלויות פונקציונליות, צורות נורמליות, MongoDB

- מתרגלת אחראית: שיר רוטמן.
- ההגשה **בזוגות ומוקלדת** בפורמט PDF **בלבד**, עד לתאריך 22/6/2023.
- שאלות על התרגיל יש לשאול בפורום הייעודי בפיאצה:
  https://piazza.com/technion.ac.il/spring2023/236363.
- שאלות אדמיניסטרטיביות יש להפנות לאופיר פדר.

פרטי המגישים:

ת.ז.	
200819456	1.
313511602	2.

שאלה 1 (25 נק')

נתונה סכמה $R = \{A, B, C, D, E, G, H\}$ וקבוצת תלויות פונקציונליות $F = \{AH \to B, AGH \to C, BD \to C, DH \to EC, C \to E, CG \to ABH\}$.

בנוסף, נתון פירוק $X = \{R_1(B, C, D), R_2(A, C, E, H), R_3(C, D, G, H)\}$.

א. האם X ב-BCNF? הוכיחו/נמקו

$$X\ In\ BCNF\ iff\ \ R_i \in\ X,\ \ R_i\ is\ in\ BCNF\ for\ i\ \in\ \{1{,}2{,}3\}$$

For $R_2(\mathrm{A, C, E, H})$, we have the 'C->E'. C is not a superkey of R2 because it doesn't determine 'A' and 'H'. Therefore, R2 is not in BCNF.

So, the given decomposition 'X' is not in BCNF, because 'R2' violate the BCNF condition.

ב. האם X ב-3NF? הוכיחו/נמקו

A decomposition is 3NF if for every non-trivial functional dependency 'X->Y' in a relation, one of the following holds:

1. 'X' is a superkey

2. Each attribute 'A' in 'Y \X' (the set difference) is prime attribute for the relation.

For R2(A, C, E, H), we have one functional dependency 'C->E', 'C' is not a superkey of 'R2', and 'E' is not prime, so 'R2' is not in 3NF.

So, the given decomposition X isn't in 3NF because R2 violate the 3NF condition.

ג. הראו כיצד האלגוריתם לבדיקת שימור מידע שנלמד בכיתה פועל על פירוק זה, ובהתאם לתוצאה קבעו האם X משמר מידע או לא

$$F=\{AH \rightarrow B, AGH \rightarrow C, BD \rightarrow C, DH \rightarrow EC, C \rightarrow E, CG \rightarrow ABH\}$$

$$X=\{R_1(B,C,D), R_2(A,C,E,H), R_3(C,D,G,H)\}$$

A	B	C	D	E	G	H
a1	b	c	d	e1	g1	h1
a	b2	c	d2	e	g2	h
a3	b3	c	d	e3	g	h

נסתכל על השורה השלישית.

נפעיל את הכללים C->E, CG->ABH ונקבל שורה שלמה, לכן הפירוק משמר מידע.

ד. מצאו דוגמה למסד נתונים $r \vDash F$ שמקיים את המסקנה שהגעתם אליה בסעיף הקודם, והראו זאת עפ"י הגדרת שימור המידע (תזכורת: פירוק הוא משמר מידע אם לכל יחס r מעל R כך ש-$r \vDash F$, מתקיים $\bowtie_{i=1\ldots n} \; \pi_{R_i}(r) = r$)

$$r = \{(0,1,1,1,0,0,0)\}$$
$$r_1 = \{(1,1,1)\}, r_2 = \{(0,1,0,0)\}, r_3 = \{(1,1,0,0)\}$$

ניתן לראות כי לפי הגדרת שימור מידע הjoin של r_i הוא r.

ה. האם X משמר תלויות? הראו ע"פ האלגוריתם שנלמד בביתה

We will start with $'AH \rightarrow B'$:

$$Z_f = \{A, H\}$$

$$Z_f = Z_f \cup \left(\left(Z_f \cap R_1\right)_f^+\right) \cap R_1) = \{A, H\}$$

$$Z_f = Z_f \cup \left(\left(Z_f \cap R_2\right)_f^+\right) \cap R_2) = \{A, H\}$$

$$Z_f = Z_f \cup \left(\left(Z_f \cap R_3\right)_f^+\right) \cap R_3) = \{A, H\}$$

No changes and $B \notin Z_f, 'AH \rightarrow B'$ does not preserve and hence the decomposition does not preserver dependencies.

שאלה 2 (20 נק')

נתונה סכמה $R = \{A, B, C, D, E, G\}$ וקבוצת תלויות פונקציונליות $F = \{AC \to D, BD \to G, G \to ABCD\}$.

א. מצאו שני מפתחות **קבילים** שונים של R. עבור אחד מהם הראו כיצד הוא מתקבל מהרצת האלגוריתם למציאת מפתח.

$\{B, D, E\}, \{G, E\}$

$\{G, E\}$:

Start with R = {A, B, C, D, E, G}.
Try to remove attributes one by one:

- Removing A, the closure of {B, C, D, E, G} under F is R, so A can be removed.
- Removing B, we get {C, D, E, G}}. The closure of this set under F is R, so B can be removed.
- Removing C, we get {D, E, G}}. The closure of this set under F is R, so C can be removed.
- Removing D, we get {E, G}. The closure of this set under F is R, so D can be removed.
- Removing E, we get {G}. The closure of this set does not include E, so E is necessary.
- Removing G, the closure of {E} does not include G, so G is necessary.

No more attributes can be removed, so {E, G} is a proper keys.

ב. מצאו פירוק X משמר מידע של R, כך ש-X הוא מהצורה $X = \{R_1(\circ,\circ,\circ), R_2(\circ,\circ,\circ), R_3(\circ,\circ,\circ)\}$ ו-$R_1 \cap R_2 \cap R_3 = \emptyset$ והוכיחו ע"פ האלגוריתם שהוא אכן משמר מידע. אין צורך להראות כיצד מצאתם את הפירוק.

שימו לב – פתרונות **נכונים** (במובן של שימור מידע) המשתמשים ב-2 תתי-סכמות יקבלו 60% מהניקוד ופתרונות **נכונים** המשתמשים ביותר תתי-סכמות או בתתי-סכמות גדולות יותר יקבלו 80% מהניקוד (פתרונות נכונים עם 3 תתי-סכמות קטנות יותר יקבלו ניקוד מלא)

$$X = \{R1(B,C,D), R2(D,B,G), R3(A,E,G)\}$$

The intersections are R1 ∩ R2 = {B, D}, R1 ∩ R3 = ∅, R2 ∩ R3 = {G} so R1 ∩ R2 ∩ R3 = ∅.

A	B	C	D	E	G
	b	c	d		
	b		d		g
a				e	g

$$BD \rightarrow G$$

A	B	C	D	E	G
	b	c	d		g
	b		d		g
a				e	g

$$G \rightarrow ABCD$$

A	B	C	D	E	G
a	b	c	d	e	g
a	b		d	e	g
a	b		d	e	g

הסעיפים הבאים בלתי תלויים בסעיפים הקודמים של השאלה.

תהי סכמה $R = (A_1, A_2, \dots, A_n)$ וקבוצת תלויות פונקציונליות F הכוללת תלות אחת בלבד f.

ג. הוכיחו/הפריכו: R היא ב-BCNF אמ"מ לכל $i \in \{1, \dots, n\}$ מתקיים ש-A_i מופיע ב-f לפחות פעם אחת או ש-f טריוויאלית.

Proof:

The first direction is trivial. If 'R' in BCNF, then either f is trivial or all features appear in f since there is only one dependence function and X (where 'X -> Y') must be superkey then necessarily all features are in f.

The second direction. There are two cases to consider:

1. 'f' is trivial: A trivial functional dependency is of the form 'X -> Y' where 'Y' is a subset of 'X'. Trivial functional dependencies always hold, so 'R' will be in BCNF if 'f' is trivial, as there is no nontrivial dependency that violates the BCNF condition.

2. Every attribute in 'R' appears in 'f': If 'f' is a nontrivial dependency of the form 'X -> Y' such that 'X U Y = R' (because every attribute in 'R' appears in 'f'), then 'X' is a superkey of 'R' since it functionally determines every other attribute in the relation. Therefore, 'R' will be in BCNF.

So in either case, 'R' will be in BCNF. Therefore, the statement is correct.

ד. הוכיחו/הפריכו: F היא כיסוי מינימלי של R אמ"מ כל A_i מופיע ב-f בדיוק פעם אחת.

The claim is incorrect.

It's possible for a functional dependency to have a right side (Y) with more than one attribute and still have each attribute appear exactly once in f. This would violate condition 1 of a minimal cover.

Therefore, F is not necessarily a minimal cover of R if every Ai appears in f exactly once.

Here's a counterexample:

R = {A, B, C, D}

F = {AB -> CD}

Each attribute appears exactly once on the left side and on the right side of f, but F is not a minimal cover for R since $|CD| = 2$, so the claim is incorrect.

שאלה 3 (25 נק')

נתונה סכמה $R = \{A, B, C, D, E, G, H\}$ וקבוצת תלויות פונקציונליות $F = \{B \to G, BG \to E, C \to D, E \to CH, EH \to D, G \to B, G \to H\}$.

א. מצאו מפתח קביל (אין צורך להראות כיצד הגעתם אליו), והראו כיצד ניתן להסיק ממנו את כל הסכמה R בעזרת אקסיומות ארמסטרונג:

$$\{\boldsymbol{A}, \boldsymbol{B}\}$$

$$\{A, B\} - B \to G - \{A, B, G\}$$

$$\{A, B, G\} - BG \to E - \{A, B, G, E\}$$

$$\{A, B, G, E\} - E \to CH - \{A, B, G, E, C, H\}$$

$$\{A, B, G, E, C, H\} - C \to D - \{A, B, G, E, C, H, D\}$$

ב. הוכח/הפרך; במקרה של הפרכה, ציינו דוגמה נגדית, כלומר יחס המספק את F אך לא את התלות.

1. $F \vdash BG \to EGH$

Proof:

- $BG \to E\ (F)$
- $BG \to H$

$$-\text{transitivity of } BG \to G\ (\text{reflexivity})\ and\ G \to H\ (F)$$

- $BG \to G(\text{reflexivity})$

$$BG \to EGH$$

2. $F \vdash CH \to AE$

A counterexample would be a relation that satisfies F but not the dependence CH → AE. For example, consider the relation r(A,B,C,D,E,G,H) = {(1,1,1,1,1,1,1), (1,1,0,0,1,1,0)}. This relation satisfies all the given functional dependencies but violates CH → AE because for the same CH, we have two different values of A.

כעת נתונים גם הפירוקים הבאים:

$$X_1 = \{R_1(A,B,G), R_2(B,C,E), R_3(D,E,H)\},$$
$$X_2 = \{R_1(A,G), R_2(B,E,G), R_3(C,D), R_4(D,E), R_5(G,H)\},$$
$$X_3 = \{R_1(A,G), R_2(B,G), R_3(C,D,G), R_4(D,E), R_5(G,H)\}$$

ג. האם X_1 משמר מידע? הוכיחו את תשובתכם

$$F = \{B \to G, BG \to E, C \to D, E \to CH, EH \to D, G \to B, G \to H\}$$

A	B	C	D	E	G	H
a	b				g	
	b	c		e		
			d	e		h

$$B \to G, E \to CH$$

A	B	C	D	E	G	H
a	b				g	
	b	c		e	g	h
		c	d	e		h

$$BG \to E, G \to H$$

A	B	C	D	E	G	H
a	b			e	g	h
	b	c		e	g	h
		c	d	e		h

$$EH \to D, E \to CH$$

A	B	C	D	E	G	H
a	b	c	d	e	g	h
	b	c	d	e	g	h
		c	d	e		h

According to the algorithm, decomposition X_1 preserves information.

ד. סמנו ✓ או ✗ בטבלה הבאה בהתאם לפירוקים הנתונים:

	X_2	X_3
BCNF	v	v
3NF	v	v

במידה וסימנתם ✗ באחד התאים, ציינו הפרה של הצורה הנורמלית (תלות שלא מקיימת את התנאים):

שאלה 4 (10 נק')

בהרצאה ראינו כי אקסיומות ארמסטרונג – רפלקסיביות, הכללה וטרנזיטיביות – מהוות מערכת הוכחה נאותה ושלמה.

הסעיפים הבאים מתייחסים למערכת אקסיומות קטנה יותר המורכבת מרפלקסיביות וטרנזיטיביות (ללא הכללה).

א. האם המערכת נאותה? הסבירו בקצרה

Yes, the system consisting of reflexivity and transitivity axioms is sound. Soundness refers to the property that every derived functional dependency is logically implied by the given set of functional dependencies.

Even if we don't have the augmentation axiom (which allows you to infer AC -> BC from A -> B), the system is still sound because it doesn't allow us to make any incorrect inferences (That means even if the system doesn't tell the whole truth, but what it says is true and it won't lie).

Since the reflexivity and transitivity axioms are valid logical inferences, any functional dependency derived using these axioms is guaranteed to be logically implied by the given set of functional dependencies. Therefore, the system is sound.

ב. האם המערכת שלמה? אם כן, הסבירו איך תשתנה ההוכחה מההרצאה. אם לא, הוכיחו בעזרת דוגמא נגדית.

No. The reason is that there are functional dependencies that we can't derive using only reflexivity and transitivity. Specifically, the augmentation axiom plays a critical role in enabling the derivation of additional dependencies.

For example, if we know that {A} functionally determines {B} (A -> B), and we have an additional attribute C, the augmentation rule would allow us to infer that {A, C} functionally determines {B, C} (AC -> BC). Without the augmentation rule, this dependency can't be derived from A -> B using only reflexivity and transitivity.

The difference between perfection and soundness. soundness ensures that the system doesn't tell you any lies (if it proves something, it's true), while completeness ensures that the system tells you **all** the truth (if something is true, the system can prove it).

שאלה 5 – MongoDB (20 נק')

כידוע לכם, לאחרונה פרצה שביתה במוסדות האקדמיים בישראל על רקע פערים בהסכמי השכר של הסגל הבכיר.

בשאלה זו נבחן מסד נתונים המתאר את חברי הסגל וחלק מהנתונים שנלקחים בחשבון במהלך המו"מ. מסד הנתונים מכיל אוסף (collection) יחיד, SeniorStaff, כאשר כל מסמך (document) באוסף הוא מהצורה הבאה:

```
{
    _id: <ObjectId>,
    staff_member_name: <string>,
    staff_member_id:<string>,
    university: <string>,
    study: <string>,
    salary:<int>,
    estimated_retirement_salary:<int>,
    seniority_year:<int>,
    estimated_retirement_year:<int>
}
```

דוגמה למסמך אפשרי:

```
{
    _id: ObjectId("63d536a10b25ad09196b08e7"),
    staff_member_name: "Ted Mosby",
    staff_member_id: "240707089",
    university: "Tel-Aviv",
    study: "architecture",
    salary: 33,000,
    estimated_retirement_salary: 29,000,
    seniority_year: 2013,
    estimated_retirement_year: 2034
}
```

א. כתבו שאילתת MongoDB המחזירה את שמות האוניברסיטאות שיש בהן חברי סגל חדשים, כלומר חברי סגל שהפכו לבכירים בשנת 2023.

```
db.SeniorStaff.distinct("university", { "seniority_year": 2023 })
```

ב. כתבו שאילתת MongoDB המחזירה עבור כל אוניברסיטה את מספר חברי הסגל עם ותק של פחות מ-10 שנים (seniority_year אחרי 2013).

```
db.SeniorStaff.aggregate({
  $match: { "seniority_year": { $gt: 2013 } } },
 { $group: { _id: "$university", count: { $sum: 1 } }
})
```

ג. כתבו שאילתת MongoDB המחזירה עבור כל אוניברסיטה וכל שנה בה יש חבר סגל שצפוי לצאת לפנסיה, את ההפרש בין סכום המשכורות הנוכחי של חברי הסגל באוניברסיטה שצפויים לפרוש באותה שנה לסכום המשוערך של המשכורות של אותם חברי סגל בעת הפרישה. מיינו את התוצאות לפי שם האוניברסיטה, בסדר עולה.

```
db.SeniorStaff.aggregate({
  {
    $group: {
      _id: { "university": "$university", "year": "$estimated_retirement_year" },
      totalSalary: { $sum: "$salary" },
      totalEstimatedRetirementSalary: { $sum: "$estimated_retirement_salary" }
    }
  },
  {
    $project: {
      _id: 0,
      "university": "$_id.university",
      "year": "$_id.year",
      "salaryDifference": { $subtract: [ "$totalSalary", "$totalEstimatedRetirementSalary" ] }
    }
  },
  { $sort: { "university": 1 } }
})
```

ד. כעת נבדוק באילו אוניברסיטאות המשכורות המשולמות גבוהות משמעותית מהמשכורות המשוערכות לאחר פרישה לגמלאות (כלומר צפוי "חסכון" ניכר לאחר פרישת חברי סגל). נאמר שחבר סגל הוא "יקר" אם המשכורת הנוכחית שלו גבוהה בלפחות 50% מהמשכורת הצפויה שלו בפנסיה. לדוגמה, טד מוסבי (שמתואר במסמך למעלה) הוא **לא** חבר סגל יקר היות ש-$\frac{33000}{29000} \approx 1.14$, כלומר המשכורת הנוכחית שלו גבוהה ב-14% בלבד מהמשכורת שלו לאחר הפרישה.

השתמשו ב-MapReduce על מנת להחזיר עבור כל אוניברסיטה את הערך 1 אם יש לה יותר מ-20 חברי סגל יקרים שצפויים לפרוש בין השנים 2025 ל-2030 (כולל) ו-0 אחרת. בסעיף זה ניתן להשתמש בתנאי if של javascript (הסינטקס זהה לתנאי if של C). בנוסף, במידת הצורך ניתן להשתמש באופרטורי ההשוואה המופיעים בתיעוד של MongoDB.

```
db.SeniorStaff.mapReduce(
  function() {
    if (this.salary >= 1.5 * this.estimated_retirement_salary &&
      this.estimated_retirement_year >= 2025 &&
      this.estimated_retirement_year <= 2030) {
      emit(this.university, 1);
    }
  },
  function(key, values) {
    var total = Array.sum(values);
    return (total > 20) ? 1 : 0;
  },
  {
    out: { inline: 1 }
  }
)
```