

מרצה: פרופ' בני קימלפלד
מתרגלים: רואי קיסוס
חמודי סיף
גיא הורוביץ

סמסטר אביב תשפ"א

מסדי נתונים

236363

מועד ב'

8 באוקטובר 2021

פירוט החלקים והניקוד:

שאלה	נושא	ניקוד	הערות
1	ERD Design Theory	25	
2	RA, RC Datalog	20	
3	SQL	20	
4	Concurrency Control	11	
5	XML	12	יש לבחור 2 שאלות מתוך 5,6,7
6	Neo4j MongoDB	12	יש לבחור 2 שאלות מתוך 5,6,7
7	RDF	12	יש לבחור 2 שאלות מתוך 5,6,7

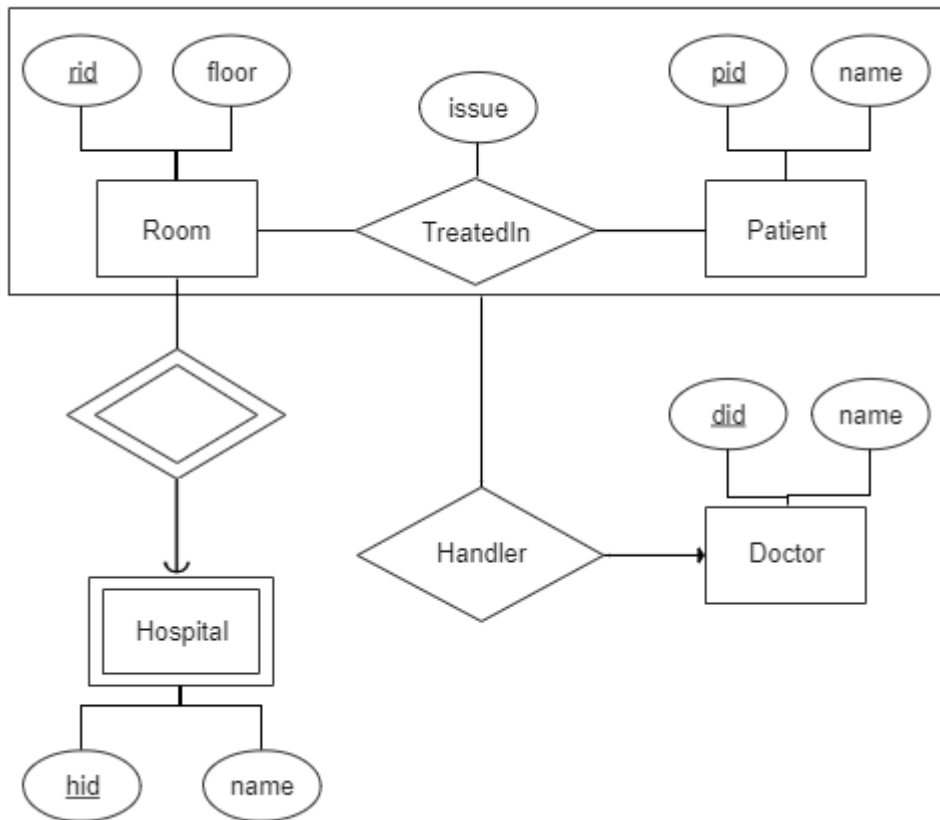
הנחיות לנבחנים:

1. כתבו את התשובות אך ורק בטופס הבחינה ובמקום המיועד להן, מחברת הטייטה לא תיבדק.
2. ניתן להביא למבחן חומר כתוב\מודפס על גבי 6 דפי A4 (דו צדדיים).
3. אין לקבל או להעביר חומר כלשהו בזמן הבחינה.
4. יש להשתמש רק בסימנים או פונקציות שנלמדו בתרגול או בהרצאה והמופיעים בשקפים של הקורס. כל שימוש בסימון שאינו כזה מחייב הסבר מלא של משמעות הסימון.
5. משך הבחינה הינו שלוש שעות, תכננו את הזמן בהתאם.
6. אין לכתוב בעפרון.

בהצלחה!

שאלה 1 – ERD, Design Theory

התבוננו בתרשים ה-ERD שלפניכם:



- א. תרגמו את תרשים ה-ERD לטבלאות המתאימות על פי הכללים שנלמדו בקורס. עבור כל טבלה, עליכם לרשום את סכמת הטבלה שתתקבל בתרגום, כולל **סימון מפתחות בקו תחתון וציון מפתחות זרים** (8 נק').
- המלבן הכפול של הישות החלשה הוא על Room ולא על Hospital.

ב. הסתכלו על הסכמה (U, F) כאשר

$$U = \{A, B, C, D, E\}, \quad F = \{D \rightarrow C, C \rightarrow B, B \rightarrow A, AE \rightarrow D\}$$

כעת הסתכלו על הפירוק $\{AD, DE, ABCE\}$.

הראו הרצה על הסכמה והפירוק הנ"ל של האלגוריתם לבדיקת שימור מידע. האם הפירוק משמר מידע?

- אם כן, האם ניתן היה לוותר על אחת מתתי הסכמות ועדיין לקבל פירוק משמר מידע?
- אם לא, הראו דוגמא לטבלה ולרשומה שמתווספת לאחר פירוק וצירוף.

(8 נק')

- ג. נתונה הסכמה (U, F) כאשר שוב $U = \{A, B, C, D, E\}$ אבל עכשיו F אינה ידועה. ידוע כי:
- הפירוק $\{ABC, CDE\}$ הינו פירוק משמר מידע.
 - קיימת טבלה חוקית עבור (U, F) בה ישנן שתי רשומות שמסכימות על הערך של C (כלומר יש להן את אותו הערך עבור C) אבל לא על הערך של D .

לכל אחד מחמשת השדות ב- U ענו על השאלות הבאות:

1. האם השדה חייב להופיע בצד ימין של לפחות כלל אחד ב- F ?
2. האם השדה חייב להופיע בצד שמאל של לפחות כלל אחד ב- F ?

בכל פעם שהתשובה היא "כן", הסבירו למה. (9 נק')

שאלה 2 – RA, RC, Datalog

א. הזכרו כיצד הוגדרה פעולת החלוקה ע"י האלגברה הרלציונית:

$$R \div S := \pi_X R - \pi_X((\pi_X R \times S) - R)$$

האם ניתן היה להגדיר את פעולת החלוקה אך ורק בעזרת הפעולות בקבוצה $\{\times, \cup, \sigma, \rho, \pi\}$? אם כן, הראו כיצד. אם לא, הוכיחו כי הדבר בלתי אפשרי. (5 נק')

ב. להלן סכמות של שני יחסים: $R(A, B)$ ו- $S(A, B)$.

הראו כי ניתן לבטא את ההפרש $R - S$ ע"י הפעולות $\{\div, \times, \cup, \sigma, \pi, \rho\}$. (7 נק')

רמז: התחילו ב- $R \times S$ וחשבו כיצד ניתן להפעיל את יתר האופרטורים כדי להגיע ליחס $R - S$.

ג. הסכמה Teams(tid, member1, member2) מייצגת צוותים בגודל שתיים. כתבו בתחשיב היחסים (RC) שאילתה המוצאת את כל האנשים השייכים לצוות אחד בדיוק. (4 נק')

ד. אנו מייצגים גרף מכוון ע"י מסד נתונים התואם לסכמה עם שני יחסים:

- $V(\text{vertex})$ מייצג את הקודקודים.
- $E(\text{from, to})$ מייצג את הקשתות המכוונות.

כתבו תכנית Datalog שמוצאת את כל הקודקודים מהם ניתן להגיע במסלולים מכוונים לכל קודקוד אחר בגרף. אם אתם משתמשים בשלילה, אז הראו את הריבוד. (4 נק')

שאלה 3 – SQL

נתון מסד הנתונים הבא המייצג את מפעל השוקולדים של וילי וונקה:

Chocolates:

<u>CID</u>	Name	Type
------------	------	------

Inventory:

<u>IID</u>	Name	Price	AmountInStock
------------	------	-------	---------------

Ingredients:

<u>IID</u>	<u>CID</u>	AmountNeeded
------------	------------	--------------

הטבלה Chocolates שומרת את מזהה השוקולד ופרטים נוספים:

- CID – מזהה השוקולד.
- Name – שם השוקולד.
- Type - סוג השוקולד.

הטבלה Inventory שומרת מידע עבור המוצרים השונים הנצרכים במפעל:

- IID – מזהה המוצר.
- Name – שם המוצר.
- Price – מחיר המוצר ליחידה בודדת.
- AmountInStock – כמות המוצר במלאי.

הטבלה Ingredients שומרת את הרכב המתכון של כל שוקולד:

- IID – מזהה המוצר, מפתח זר ל-IID בטבלה Inventory.
- CID – מזהה השוקולד, מפתח זר ל-CID בטבלה Chocolates.
- AmountNeeded – הכמות הנדרשת למתכון, גדולה מ-0.

מפתחות מסומנים בקו תחתון.

א. נגדיר כי ניתן להכין שוקולד אם כמות המצרכים הנדרשת להכנת השוקולד קיימת במלאי. כתבו שאילתה המחזירה את כל מזהי השוקולדים שניתן להכין אותם עפ"י המצב הנוכחי של מסד הנתונים. (6 נק')

--

ב. כתבו שאילתה המחשבת עבור כל סוג (Type) שוקולד את סכום עלויות הייצור של כל השוקולדים מסוג זה. (8 נק')

1. $\pi_{CID}(Chocolates) = \pi_{CID}(Ingredients)$
2. $\pi_{IID}(Inventory) = \pi_{IID}(Ingredients)$

ג. חמודי כתב את השאילתה הבאה על מנת לקבל את כמות המצרכים השונים (IID) הנדרשים עבור השוקולד עם CID=3.

```
SELECT COUNT(*)  
FROM Ingredients  
WHERE CID=3
```

למטרת סעיף זה הניחו כי:

- מסד הנתונים עושה שימוש בעץ B+ בעל $d=20$ כאשר גודל עלה הוא בלוק.
- יש 10,000 רשומות בטבלה Ingredients, ובכל בלוק יש 20 רשומות.

הציעו אינדקס שייעל את החישוב של שאילתה זו והסבירו. הניחו שתוצאתה של השאילתה של חמודי היא 100, מהו השיפור בזמן הריצה (Speedup) המתקבל עם הצעתכם ביחס למימוש הנאיבי של חמודי? הסבירו והראו מספר מדויק כפי שנלמד בהרצאה. (6 נק')

שאלה 4 – Concurrency Control

לכל אחד מהתזמונים הבאים, ציין והוכח אם הוא:

- בר סדרתיות מבטים.
- בר סדרתיות קונפליקטים.
- יכול להיווצר כתוצאה מהפעלת פרוטוקול 2PL.

שימו לב: בכל תזמון, ההבדל מהתזמון הקודם מודגש.

$R_1(x) \quad R_2(y) \quad W_3(y) \quad W_2(x) \quad W_1(x) \quad W_3(x)$		
ב"ס מבטים? כן / לא	ב"ס קונפליקטים? כן / לא	2PL? כן / לא
הוכחה:	הוכחה:	הוכחה:

$R_1(x) \quad R_2(y) \quad W_3(y) \quad \mathbf{W_1(x)} \quad \mathbf{W_2(x)} \quad W_3(x)$		
ב"ס מבטים? כן / לא	ב"ס קונפליקטים? כן / לא	2PL? כן / לא
הוכחה:	הוכחה:	הוכחה:

$R_1(x) \quad R_2(y) \quad W_3(y) \quad \mathbf{W_1(z)} \quad W_2(x) \quad W_3(x)$		
ב"ס מבטים? כן / לא	ב"ס קונפליקטים? כן / לא	2PL? כן / לא
הוכחה:	הוכחה:	הוכחה:

שאלה 5 - XML

להלן מסמך ה-DTD "exam.dtd" לייצוג בחינות:

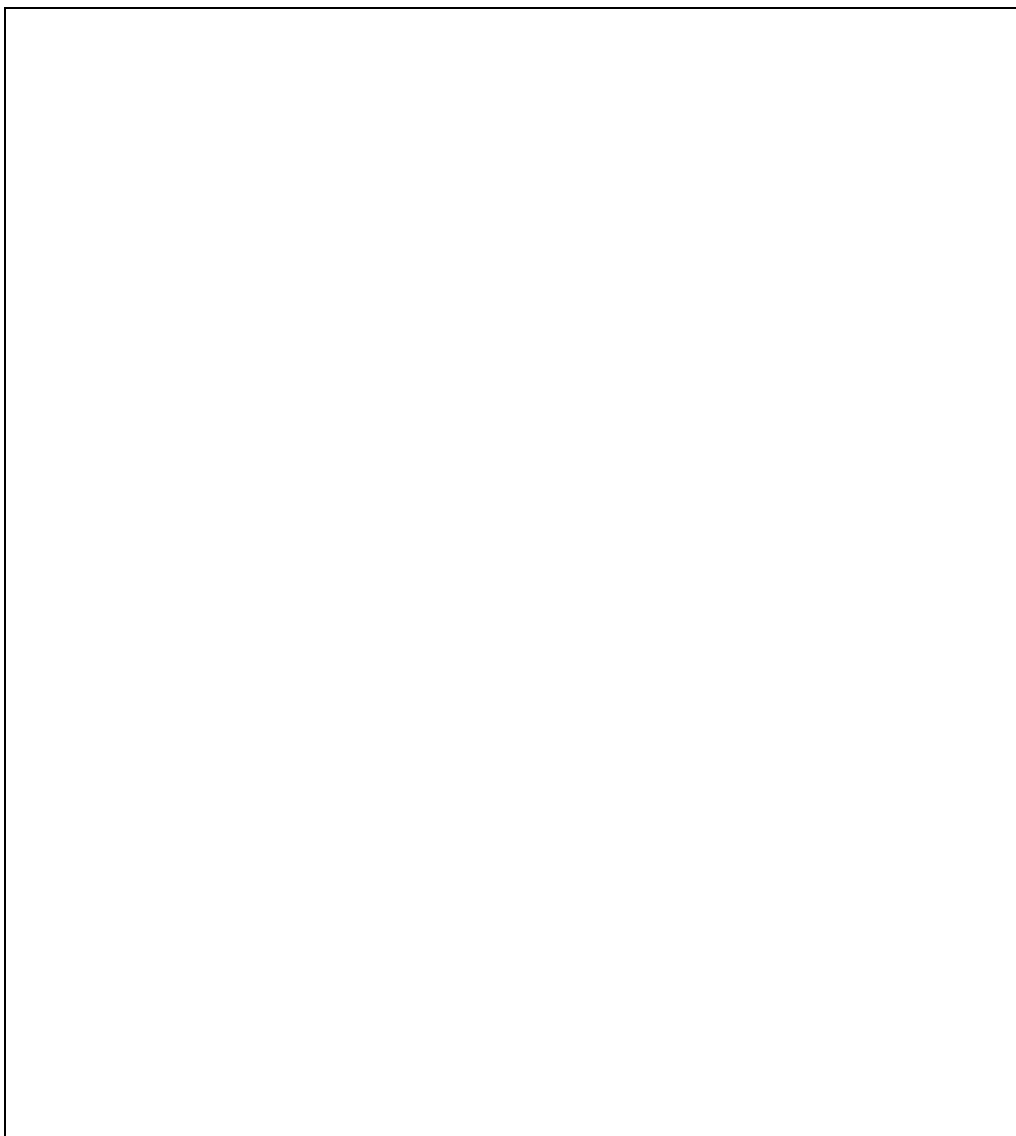
```
<!DOCTYPE exam [  
<!ELEMENT exam (lecturer, TA*, question+)>  
<!ATTLIST exam  
    course CDATA #REQUIRED>  
<!ELEMENT lecturer (#PCDATA | name)*>  
<!ATTLIST lecturer  
    id ID #REQUIRED>  
<!ELEMENT TA (#PCDATA | name)*>  
<!ATTLIST TA  
    id ID #REQUIRED>  
<!ELEMENT question (#PCDATA)>  
<!ATTLIST question  
    authors IDREFS #REQUIRED  
    difficulty (hard|easy) "hard"  
    subject CDATA #REQUIRED>  
<!ELEMENT name (#PCDATA)>  

```

בנוסף נתון המסמך לדוגמא הבא המציית לכללי ה-DTD הנ"ל:

```
<?xml version="1.0"?>  
<!DOCTYPE exam SYSTEM "exam.dtd">  
<exam course="MAMAN">  
    <lecturer id="id_1">  
        <name>Benny Kimelfeld</name>  
    </lecturer>  
    <TA id="id_2">  
        <name>Roei Kisous</name>  
    </TA>  
    <TA id="id_3">Guy Horowitz</TA>  
    <question authors="id_1 id_3" subject="design">  
        "look at the following ERD..."  
    </question>  
    <question authors="id_2" difficulty="easy" subject="XML">  
        "look at the following DTD..."  
    </question>  
</exam>
```

א) ציירו את העץ שמייצג המסמך לדוגמא. עבור כל אלמנט בעץ, ציינו את ערכי כל האטריביוטים שלו (2 נק').



ב) מהו המספר המינימלי של מזהים (IDs) שצריכים להופיע במסמך המציית ל-DTD הנ"ל? הסבירו (2 נק').



ג) עבור השאילתה

`//question/id(@authors)/text()`

הסבירו במילים מה מחזירה השאילתה על מסמך המציית לכללי ה-DTD וכתבו מה יהיה פלט השאילתה על מסמך הדוגמה (3 נק').

ד) עבור השאילתה

`//question[count(/id(@authors)[. = //TA]) = 1][@difficulty = "hard"]/@subject`

הסבירו במילים מה מחזירה השאילתה על מסמך כללי המציית לכללי ה-DTD וכתבו מה יהיה פלט השאילתה על מסמך הדוגמה (5 נק').

שאלה 6 – mongoDB, Neo4j

בקורס "פרסור ברשת האינטרנט" ניתנה מטלת בית ובה יש צורך בפרסור (Parsing) של אובייקטי JSON הנמצאים ברשת האינטרנט. רואי, סטודנט לא חרוץ במיוחד, לא הקשיב בשיעור ועל כן הוא מבקש את עזרתכם בכתיבת השאילתות המתאימות ב-mongoDB.

רואי הצליח באורח פלא להוריד את מאגר הנתונים מהרשת לתוך מסד נתונים הבא ב-mongoDB ובו אוסף (Collection) בודד הנקרא Parser אשר מכיל מידע על כל אובייקטי ה-JSON שהורדו. כל מסמך באוסף הוא מסוג קורס או סטודנט ומהצורה הבאה:

```
{
  _id: <ObjectId>,
  type: <string>,
  id: <int>,
  name: <string>,
  institute: <string>,
  list: [
    {
      id: <int>,
      grade: <int>,
      year: <int>,
    },
    ...,
    {
      id: <int>,
      grade: <int>,
      year: <int>,
    }
  ]
}
```

דוגמא למסמך אפשרי של קורס:

```
{
  "_id": ObjectId("056ab83901a09b"),
  "type": "Course",
  "id": 236363,
  "name": "maman",
  "institute": "Technion",
}
```

```

    "list": [
      {
        "id": 123456789,
        "grade": 100,
        "year": 2021
      },
      {
        "id": 123456729,
        "grade": 50,
        "year": 2020
      }
    ]
  }

```

כאשר עבור קורס, כל איבר ב-list הוא סטודנט שעשה את הקורס.

דוגמא למסמך אפשרי של סטודנט:

```

{
  "_id": ObjectId(056ab83901b09b),
  "type": "Student",
  "id": 123456789,
  "name": "Roei",
  "institute": "Technion",
  "list": [
    {
      "id": 236363,
      "grade": 100,
      "year": 2021
    }
  ]
}

```

כאשר עבור סטודנט כל איבר ב-list הוא קורס שעשה הסטודנט.

א. רואי כתב את השאילתה הבאה על מנת להחזיר את הממוצע ההיסטורי (לאורך כל השנים) של כל קורס:

```
db.Parser.aggregate([
    $group: { _id: "$id" , value: { $avg: { "$list.grade" } } }
])
```

כאשר \$list.grade מהווה גישה לשדה grade ב-list.

רואי קיבל תוצאות לא נכונות ולא הבין למה, הסבירו במילים מה הבעיה בשאילתה והציעו במילים דרך לתקן אותה? (3 נק')

ב. רואי, שהתייאש, פנה לגיא שניסח את השאילתה הבאה עבור אותה מטרה:

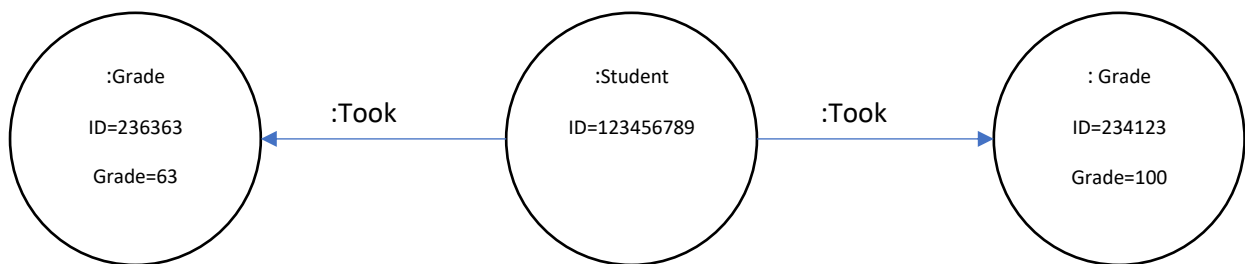
```
db.Parser.mapReduce(
    function(){
        for(var idx=0; idx<this.list.length; idx++){
            var val = this.list[idx].id;
            var val1 = this.list[idx].grade;
            emit(val, val1);
        }
    },
    function(key, values){ return (Array.sum(values)/values.length); },
    {
        out: {"result"},
        query: { type: "Student" }
    }
)
```

רואי טען שהשאילתה שגויה, בעוד גיא התעקש כי נקבל את התשובה הנכונה. הסבירו מי לדעתכם צודק ולמה? בנוסף, הסבירו את דרך הביצוע של השאילתה המצופה ממנוע של MongoDB ללא תלות בתשובתכם. (4 נק')

לרואי נמאס סופית והוא החליט לייצג את מסד הנתונים ב-Neo4j בצורה הבאה:

1. כל צומת בגרף מחזיק בתווית (label) אחת בדיוק מבין האפשרויות הבאות: **Grade**, **Student**.
2. צמתים בעלי תווית **Grade** מחזיקים בתכונות בשם **ID**, **Grade**.
3. צמתים בעלי תווית **Student** מחזיקים בתכונה בשם **ID**.
4. כל צומת בעל תווית **Student** יכול להיות מחובר לכל היותר לצומת אחד בעל תווית **Grade** עם ID המייצג את ציונו בקורס זה, ע"י קשר בעל תווית (label) **Took**.

להלן **דוגמא** לגרף המקיים את ארבעת הכללים:



ג. לכל אחת מהשאלות, הסבירו בקצרה האם היא מחשבת את הממוצע או לא: (5 נק')

- i. `MATCH () -[:Took] ->(C:Grade)`
`WITH sum(*) as s, count(*) as c`
`RETURN s/c`
- ii. `MATCH () -[:Took] ->(C:Grade)`
`WITH C.ID as id, sum(C.grade) as s, count(*) as c`
`RETURN id, s/c`
- iii. `MATCH (S:Student) -[:Took] ->(C:Grade)`
`WITH sum(C.grade) as s, count(*) as c`
`RETURN C.ID, s/c`
- iv. `MATCH () -[:] ->(C:Grade)`
`WITH C.ID, sum(C.grade) as s, count(C) as c`
`RETURN C.ID, s/c`

שאלה 7 – RDF

בשאלה זו הניחו את קיום ה-namespaces הבאים:

- rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
- rdfs: <http://www.w3.org/2000/01/rdf-schema#>
- ex: <http://example.maman.cs.technion/>

התבוננו בגרף הבא המיוצג ב-RDF:

ex:ArtemDolgopyat	ex:sportBranch	ex:Gymnastics
ex:LinoyAshram	ex:sportBranch	ex:Gymnastics
ex:AvishagSemberg	ex:sportBranch	ex:Taekwondo
ex:IsraelJudoTeam	ex:sportBranch	ex:Judo
ex:ItayShanny	ex:sportBranch	ex:Archery
ex:IsraelJudoTeam	rdf:type	ex:nationalTeam
ex:IsraelGymnasticsTeam	rdf:type	ex:nationalTeam
ex:ArtemDolgopyat	ex:wonMedal	ex:gold
ex:LinoyAshram	ex:wonMedal	ex:gold
ex:AvishagSemberg	ex:wonMedal	ex:bronze
ex:IsraelJudoTeam	ex:wonMedal	ex:bronze

(א) רשמו את תוצאת ההפעלה של כל אחת מן השאילתות הבאות על הגרף הנתון:

1. (3 נק')

```
SELECT ?s ?b ?t {  
    ?s ex:sportBranch ?b.  
    ?s ex:wonMedal ?m  
    OPTIONAL {  
        ?s rdf:type ?t}  
    MINUS {  
        ?s rdf:type ex:nationalTeam}  
}
```

2. (3 נק')

```
SELECT ?s1 ?s2 {  
  ?s1 ex:sportBranch ?b  
  OPTIONAL {  
    ?s1 ex:wonMedal ?m}.  
  ?s2 ex:wonMedal ?m  
  FILTER (?s1 != ?s2)  
}
```

3. (3 נק')

```
SELECT ?s ?b {  
  {?s ex:sportBranch ?b.  
   ?s ex:wonMedal ?m  
   FILTER (?m = ex:gold)}  
  UNION  
  {?s rdf:type ex:nationalTeam}  
}
```

ב) הוסיפו לגרף הנתון שלישייה חדשה, כך שכאשר השאילתה

```
SELECT ?s { ?s rdf:type ex:team }
```

תרוץ על מנוע שתומך ב-RDFS, היא תניב את הפלט הבא (3 נק'):

```
{?s ← ex:IsraelJudoTeam }, {?s ← ex:IsraelGymnasticsTeam }
```

--	--	--

דפים נוספים לתשובות:

