אוניברסיטת אריאל בשומרון
המחלקה למדעי המחשב

מבחן במסדי נתונים (ב')

ד"ר עמוס עזריה

7028010

סמסטר ב' מועד ב' **ט"ו באב** התשע"ז, 7.8.2017

הנחיות כלליות:

- משך הבחינה: 180 דקות.
- **יש לענות בגוף השאלון!** המחברת תשמש כטיוטא בלבד. על מענה במחברת יורדו נקודות!
- אין להכניס שום חומר עזר.
- השימוש במחשבון **אסור**.
- מומלץ לקרוא את ההוראות באנגלית ולפנות לעברית רק במקרה של חוסר הבנה.
- בסיום הבחינה - נא למסור את השאלון ואת המחברת.

	1	2	3	4	5	6	7	8	9	10	**Total**
Max points	15	13	12	8	10	7	10	9	5	14	**103**
Grade											

ב ה צ ל ח ה!

03.9066640
03.9066692
אריאל, 40700
אוניברסיטת אריאל בשומרון ע"ר

1. SQL (15 pt

נתונות הטבלאות הבאות לייצוג מרפאת שיניים:

Patients (patientId, firstName, lastName, YearOfBirth, gender)

TreatmentTypes (treatmentTypeId, treatmentName, price)

TreatmentForPatient (patientId, treatmentTypeId, dayAsInt)

ניתן להניח ש dayAsInt הוא מספר שלם המייצג את יום העסקים מאז המרפאה החלה לפעול.

כיתבו שאילתא שמחזירה את מספר הגברים (gender=0) שטופלו בכל יום ובתנאי שביום זה ההכנסה של המרפאה הייתה לפחות 1000 ש"ח.

```
SELECT dayAsInt, count(patientId) as total_men

FROM Patients p JOIN TreatmentForPatient tp

ON p.patientId =  tp.patientId JOIN TreatmentTypes tt

ON tp.treatmentTypeId = tt.treatmentTypeId

GROUP BY dayAsInt Having sum(price) >= 1000
```

2. Normalization (13 pt):

Given the following relation:

נתונה הרלציה הבאה והתלויות שלה:

R(A, B, C, D)

And the following dependencies:

{A,B}->C

{B} ->D

{C, D}->A

{C, D} -> B

a. What are the sets of candidate keys (5 pt):

מה הם מפתחות הקנדידייט?

{A, B}, {C, D}, {B, C} are the candidate-keys.

The “algorithm” to find them doesn’t work here since all the attributes occurs in the right side.

Then, we need to find them by trying possibilities.

Obviously {C, D} is easy to find.

If B→D then, {C, B} is also a candidate-key.

Finally, we can add {A, B}.

b. What are the **prime** attributes? (1 pt.):

אילו מהתכונות הן פריים?

The prime attributes are A, B, C, D. All the attribute occur at least once in the candidate-keys which explain the answer

c. What level of NF (normal form) does the relation adhere to? (1NF, 2NF, 3NF, BCNF(3.5NF), 4NF)? Show why the relation does not adhere to any higher NF (7 pt).

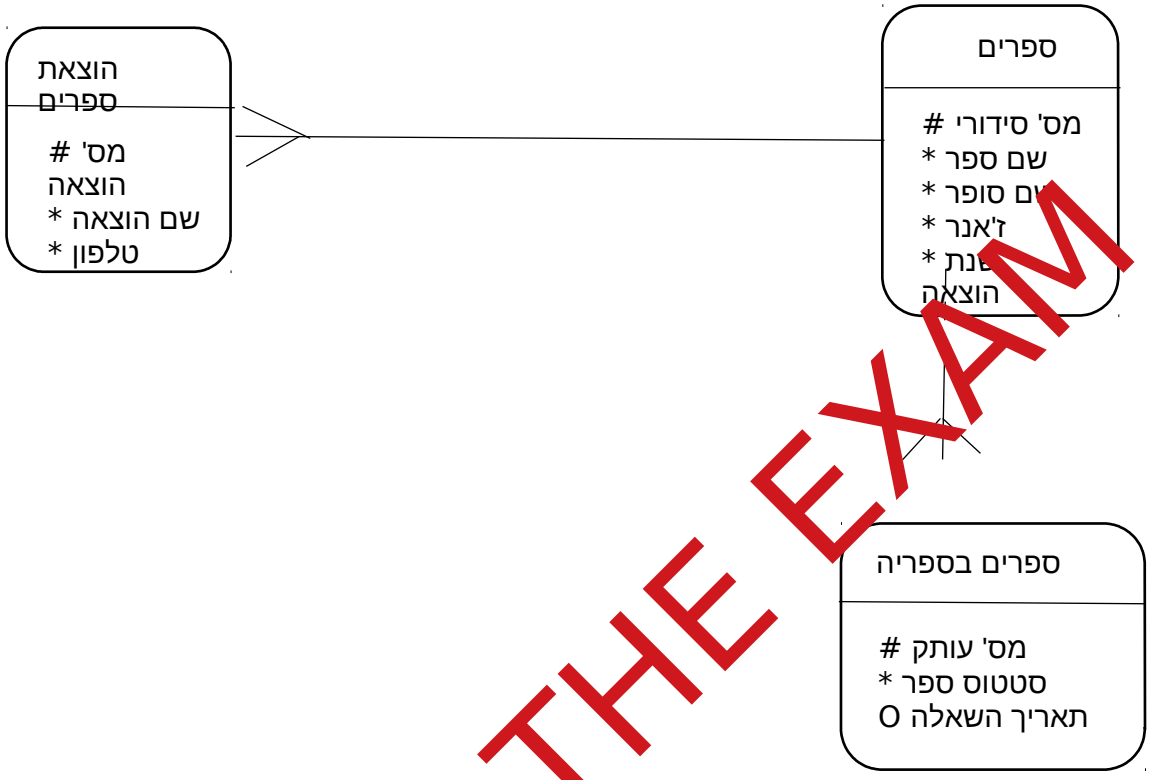
באיזה נורמל פורם הרלציה? נמק.

Assuming 1NF, and by the non-presence of any non-prime attribute, 2NF and 3NF are also true (by empty way), we want now to check 3.5NF.

As we can notice, D depends on B, which isn’t a key. We’re done, and the level of the NF is 3.

3. ERD

להלן מודל ERD עבור ספריית היישוב נווה הדרים:



שימו לב שהמודל שומר על כללי הנרמול לפחות עד 5NF. יש לשמור על נירמול זה בכל סעיף בו מתבצע שינוי למודל.

א. מהם מפתחות ה-foreign key בכל אחד מהטבלאות שנוצרו עבור המודל (4 נק')?

ב. הגיע ספר חדש בהוצאה משותפת של "ידיעות ספרים" ומשרד הביטחון. האם מודל ה-ERD הקיים תומך בו (4 נק')?
אם כן, נמקו
אם לא, תקנו את השירטוט הקיים.

ג. עקב התרחבות היישוב והקמת שכונות חדשות, הוחלט להקים ספריות נוספות. כיצד נתקן את המודל כך שיתמוך בכמה ספריות? תקנו את השרטוט לעיל **באופן מלא** (4 נק')

4. XML and XSD (8 pt):
State all the reasons why the following XML is not valid according to the given XSD:
כתבו את כל הסיבות מדוע האקס אם אל הבא לא תקף לפי האקס אס די הבא:

XML:
```
<computer>
 <keyboard>105 keys</keyboard>
 <ramSlot> 4GB </ramSlot>
 <mouse>3 button</mouse>
 <cpu> intel i5 </cpu>
 <monitor>vga</monitor>
</computer>
```

XSD:
```
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema">
 <xs:element name="computer">
   <xs:complexType>
     <xs:sequence>
       <xs:element name="keyboard"/>
       <xs:element name="ramSlot" minOccurs="2" maxOccurs="16"/>
       <xs:element name="mouse" />
       <xs:element name="monitor" />
       <xs:element name="cpu" />
       <xs:element name="joystick" minOccurs="0"/>
       <xs:element name="hardDrive" maxOccurs="unbounded"/>
     </xs:sequence>
   </xs:complexType>
 </xs:element>
</xs:schema>
```

1. ramSlot must occur at least twice but occurs only once.

2. swap between monitor and cpu need to be done.

3. hardDrive hasn’t a minOccurs number then by default it’s one but it doesn’t occurs at all.

5. NoSQL (10 pt.):

אתם מעוניינים לכתוב תוכנה שתעבוד עם דאטא רב. את הדאטא תרצו לאחסן במחשבים רבים ומעוניינים בשליפה ואחסון מהירים. לכל אחד מהמקרים הבאים ציינו באיזה מבסיסי הנתונים תשתמשו (תשובות אפשריות: Redis, MongoDB, Cassandra, ElasticSearch, Neo4J). שימו לב: למרות שהתשובה אינה בהכרח חד משמעית (בדר"כ ניתן להשתמש ביותר מבסיס נתונים אחד בדרכים שונות), אף על פי כן, ישנו בסיס נתונים אחד בלבד שהוא המתאים ביותר. נמקו!

א. אתם רוצים בסיס נתונים אשר יאחסן סיפורים, ויתן לכם אפשרות לקבל את הסיפורים הרלוונטיים ביותר לפי מילות מפתח.

Elactic Search and his tool TF-IDF offer the best match possible for a given question .

ב. מבחינת התוכנה שאתם כותבים, תפקידו של בסיס הנתונים אשר הוא לאחסן מסמך לפי id ולאחר מכן לשלוף את **כל** תוכן המסמך. בחלוף הזמן, תרצו להגדיר חלק מהמסכמים ככאלה שאתם לא צריכים עוד, אך לשמור אותם לעוד חודש ליתר בטחון.

Redis is a database based on the key-value search. Adding the function EXPIRE, you can delete an item after a given time

ג. אתם רוצים בסיס נתונים אשר יתן לכם לבטא קשרים בין ישויות שונות בסביבה בה יש קשרים מרובים. מעניין אתכם מי מקושר למי גם אם הקשר עקיף (למשל דרך ישות נוספת).

Neo4j is a graph database which give a particular importance to the link between all the item by determine a type for all the verticies

ד. אתם רוצים שהדאטא שלכם יהיה מאורגן בדומה למערכת קבצים. כדי לגשת למידע, אתם מעוניינים לתת את הערכים המדוייקים של כל ה attributes הקודמים. אתם מעוניינים בשפת שאילתא הדומה לזו של relational databases.

Cassandra is work as defined in the query. In fact, to ask any item in cassandra you need to search for the other attributes

ה. הדאטא שלכם בנוי ממסמכים בהם יש פרטים מספריים רבים. אתם מעוניינים במסד נתונים שיאפשר לכם שאילתות עם חישובים מורכבים על פרטים מתוך המסמכים.

MongoDB and the feature map reduce is strong and allows to handle with a lot of data

6. TF-IDF (7 pt)

Rank the following documents according to their TF-IDF score given the following query (no calculator is allowed or required); enter the most relevant document first:

Recall that the TF-IDF formula is: $tfidf(d) = \sum_{k=0}^{|Q|} \frac{\#k\ in\ d}{|d|} \log(\frac{|D|}{\#D\ with\ k})$

דרגו את המשפטים הבאים לפי ניקוד הטי אף איי די אף.

Q: elephants with horns
D1: Mary went on a date **with** Larry
D2: Google will lock **horns with** Microsoft
D3: I never knew **elephants** had **horns**
D4: I saw **elephants with** big trunks
D5: Bears **with** blue ears also have **horns**

First (most relevant): D3
Second: D4
Third: D2
Forth: D5
Fifth (least relevant): D1

See explanation in others exams.

7 Cassandra (10 pt)

Given the following two tables:

נתונות שתי הטבלאות הבאות:

T1 defined as: CREATE TABLE T1 (A INT, B INT, C INT, D INT, PRIMARY KEY((A, B), C, D)));

And:

T2 defined as: CREATE TABLE T2 (A INT, B INT, C INT, D INT, PRIMARY KEY(A, B, C));

For each of the following CQL queries, determine whether it is legal, illegal (e.g. requires ALLOW_FILTERING) or syntax error (e.g. “EAT C WITH T1”):

לכל אחת מהשאילתות הבאות כיתבו האם היא חוקית, לא חוקית, או בכלל לא בסינטאקס חוקי ל CQL.

a. SELECT * FROM T1 WHERE A=37 — illegal
b. SELECT * FROM T2 WHERE A=345 INNER JOIN T1 ON T1.C=T2.C — syntax
c. SELECT * FROM T2 WHERE A>54 — illegal
d. SELECT * FROM T1 WHERE A=145 AND B=356 AND D=43 — illegal
e. SELECT * FROM T1 WHERE C<679 AND A=465 AND B=195 — legal

See explanation in others exams.

8. Mongo DB (9 pt):

Consider the car collection which contains the following documents;

```
db.car.insert([
{car_id:"c1", name:"Audi", color:"Black", current_speed:50},
{car_id:"c2", name:"Polo", color:"White", current_speed:65},
{car_id:"c3", name:"Alto", color:"White", current_speed:75},
{car_id:"c4", name:"Santro", color:"Black", current_speed:150},
{car_id:"c5", name:"Subaru", color:"Black", current_speed:100},
{car_id:"c6", name:"Zen", color:"Blue", current_speed:97} ] )
```

What will be the content of my_out after the following mapreduce call (i.e. what will we get when we type: db.my_out.find()):

```
db.car.mapReduce(
        function (){
                if ( this.current_speed < 120 ) {
                        emit(this.color, this.current_speed);
                }
        },
        function(key, val1) {
                var total =0;
                for (var i = 0; i < val1.length; i++) {
                        total += val1[i];
                }
                return total / val1.length;
        },
        {out: "my_out"});
```

מה תכיל my_out לאחר הרצת הקוד לעיל (כלומר, מה נקבל כשנקליד db.my_out.find()):

"Algorithm" for the answer:

First from all the item inserted, delete all the item in which the speed is > 120 (in red).

Then, since the key is color, reunite the item with the same color.

Since the value is the speed, and the function is calculate the average, do the same:

Black → 75

White → 70

Blue → 97

The last step is to properly write the answer:

"my_out" = {

{ _id:”Black “ , value:75},

{ _id:”White “ , value:70},

{ _id:”Blue“ , value:97},

}

9 JAVA Streams (5 pt):

What will be the output of the following Java streams program:

```
List<String> myList = Arrays.asList("camel", "zebra", "you", "apple", "banana", "me");
myList.stream().map(a->a+a).filter(a->a.length()>=7).sorted().forEach(System.out::println);
```

מה יהיה הפלט של התוכנית?

The output of the program will be:

appleapple

bananabanana

camelcamel

zebrazebra

Short explanation:

The map function make for all the elements in the list a + a, which mean elem + elem, then from camel, we arrive to camelcamel.

Then, by the filter utilisation, all the word (after the modification) which are shorter than 7 are deleted.

Then a sorted is effectuate, and we print the result for each words.

10 Linear regression (14 pt)

Recall that for linear regression, we have defined the following loss function

$J(w,b) = \frac{1}{2m}\sum_{i=1}^{m}(h(x_i) - y_i)^2$

Recall that the hypothesis in linear regression is: h(x) = xw + b

Given the following data: x = [3, 2, 0, -4] and y = [2, 1, 1, -2]

(Assume batch gradient descent).

a. What are the values of the **gradient** (both according to w and b) when w=0 and b=0 (provide numerical values) (7pt)?

מה **הגראדינט** ב w=0,b=0 (יש לתת ערך מספרי)?

Derived by W :

sum (from I = 1 to m) [-yi*xi] / m

Derived by B:

sum (from I = 1 to m) [-yi] / m

By setting the value we get:

for W : (- 6 -2 – 8) / 4 = -4

for B : (- 2 -1 – 1 + 2) / 4 = -0.5

Then, as a final answer: the gradient for W is -4 and for B is -0.5

b. Assume the learning rate (alpha) is 0.1, what will be the value of w and b in the next iteration (3pt)? Use the gradient calculated in the section a.

הניחו שאלפא=0.1, מה יהיו הערכים של w ו b באיטרציה הבאה? (השתמשו בגרדיאנט שחישבתם בסעיף א')

By the formulas

W = w - α*grad = 0 – 0.1 * - 4 = 2 / 5

B = b- α*grad = 0 – 0.1 – 1 / 2 = 1 / 20

where W and B are the new values and w and b are the old values (0 and 0).

c. According to the values you found for w and b, what will be the prediction for x=3? When is the prediction for x=3 more accurate, with w=0, b=0 or with the values you found in the previous section? (show your calculations) (4pt).

לפי הערכים שמצאתם בסעיף הקודם, מה יהיה הניבוי ל x=3? מתי הניבוי ל x=3 מדוייק יותר כש w=0,b=0 או עם הערכים שמצאתם בסעיף הקודם? (יש להראות חישוב)

By the new W and B's value, we have: $h(3) = 3 * (2 / 5) + 1 / 20 = 25 / 20$

With the old value we have : $h(3) = 0$.

When $x = 3$, we have $y = 2$ (by the index). Then, trivially, $25 / 20$ is a better prediction than 0

since $3 – 25 / 20 < 3 – 0$.