אוניברסיטת אריאל בשומרון
המחלקה למדעי המחשב

מבחן במסדי נתונים

מירב שקרון

7029210-2

סמסטר קיץ מועד א' י"א אלול התשע"ט, 11.9.2019

הנחיות כלליות:

- משך הבחינה: 180 דקות.
- **יש לענות בגוף השאלון!** המחברת תשמש כטיוטא בלבד, מענה במחברת עלול לגרור ציון 0.
- אין להכניס שום חומר עזר.
- השימוש במחשבון **אסור**.
- בשאלות האמריקאיות רק תשובה אחת נכונה.
- בסיום הבחינה - נא למסור את השאלון ואת המחברת.

	1	2	3	4	5	6	7	8	9	Total
Max points	26	14	8	8	10	10	8	10	8	102
Grade										

ב ה צ ל ח ה !

03.9066640
03.9066692
אריאל, 40700
אוניברסיטת אריאל בשומרון ע"ר

שאלה 1- SQL (26 נק')

נתון בסיס הנתונים הרלציוני הבא המאחסן עבור חברת הביטוח "ביטוחי" את נתוני הלקוחות, הפוליסות והתביעות:

Customer (id, firstName, lastName, dateOfBirth)

טבלת לקוחות- ת.ז., שם פרטי, שם משפחה ותאריך לידה

Polisa (polisaNum, premia, trStart, trEnd, type)

טבלת פוליסות- מס' פוליסה, פרמיית הפוליסה, תאריך תחילת הפוליסה, תאריך תום הפוליסה וסוג הפוליסה

PolisaType (typeId, typeName)

טבלת סוגי פוליסות – קוד סוג, שם סוג (לדוגמא- 24 דירות, 16 רכב מקיף...)

PolisaCustomer (polisaNum, customerId, role)

טבלת לקוחות בפוליסה- מס' פוליסה, מס' ת.ז. וסוג התפקיד בפוליסה

rolesInPolisa(roleId, roleDesc)

טבלת תפקידים בפוליסה – קוד תפקיד ותיאור התפקיד (לדוגמא- משלם, בעל הנכס, נהג עיקרי, נהג נוסף...)

Tvia (tviaNum, polisaNum, trDate, amountAsked, amountPaid)

טבלת תביעות- מס' תביעה, מס' פוליסה, תאריך התביעה, סכום התביעה וסכום ששולם

א. כתבו שאילתא שתחזיר לכל סוג פוליסה את סכום הפרמיות בכל הפוליסות שבתוקף ושמס' הלקוחות בפוליסה הוא עד 2 (יש להציג את **שם** סוג הפוליסה) (10 נק')

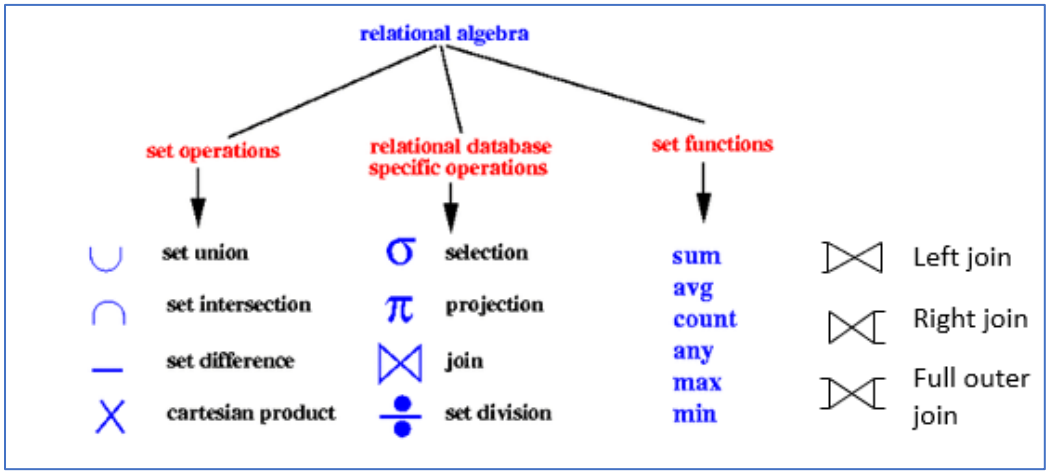
```
SELECT pt.typeName, SUM(p.premia) as totalSum
FROM polisaType as pt JOIN polisa as p

ON (pt.typename = p.type)
WHERE p.tEnd < data() AND
(SELECT count(*)

FROM polisaCustomer as pc
WHERE pc.polisaNum = p.polisaNumber ) < 3
```

ב. כתבו שאילתא שתחזיר עבור כל תביעה של פוליסה מסוג דירה (24) את ההפרש בין סכום התביעה לסכום ששולם בפועל, רק במקרה שההפרש גדול מ-50. בנוסף, יש להציג את מס' התביעה, תאריך התביעה, מס' פוליסה. (10 נק')

```
SELECT  t.tviaNum, t.trDate, t.polisaNum, (t.amountAsked-t.amountpaid) as diff
FROM tvia
WHERE diff > 50 AND polisaNum in (SELECT polisaNum FROM Polisa as p JOIN
                                  PolisaType as pt ON pt.type = p.type
                                  WHERE pt.typeID=24)
```

ג. כתבו שאילתא באלגברה רלציונית שתחזיר עבור כל פוליסה בענף רכב מקיף (16) את מס' הפוליסה, מספרי התביעות שלה ות.זהות של הלקוחות הרשומים בה. (6 נק')

להזכירכם, אלו הסימונים שלמדנו:



<u>שאלה 2 - Normalization (14 נק')</u>

נתונה הרלציה R(A,B,C,D,E,F,G)

והתלויות הבאות:

AD → BF

D → EG

BD → F

E → D

א. מהם ה-candidate key/s של הרלציה R? (5 נק')

ב. מהי רמת הנירמול של הרלציה R? נמקו! (3 נק')

ג. תקנו את הרלציה R כך שרמת הנירמול שלה תהיה גבוהה בלפחות ברמה אחת **יותר** ממה שהתקבל בסעיף הקודם. סמנו את המפתחות. תארו ופרטו! (6 נק')

שאלה 3 - NoSQL (8 נק')

איזה מהמשפטים הבאים **לא** מהווה סיבה לזה ש-NoSQL הפך לפתרון פופולרי בארגונים מסוימים?

א. גישה מהירה יותר לנתונים מאשר בסיס נתונים רלציוני.

ב. האפשרות להחזיק נתונים על שרתים מרובים ביתר קלות.

ג. שיפור היכולת לשמור על עקביות הנתונים

ד. scalability יותר טובה

שאלה 4 - neo4j (8 נק')

נתון מסד נתונים גרפי מסוג neo4j המתאר יחסי משפחה וחברות בין אנשים.

הגרף מכיל צמתים מסוג **PERSON** בעל התכונות הבאות: firstName, lastName, age

צמתים אלו יכולים להיות מקושרים ע"י קשרים מהסוגים הבאים:

MARRIED, DIVORCED, PARENT, FRIEND

כתבו שאילתא ב-cypher המחזירה את שם המשפחה של כל האנשים שהם חברים של אנשים בני 40 ושיש להם יותר מ-3 ילדים ושהם נשואים.

```
MATCH [p1:person]-[:friend]-[p2:person{"age":40}]
WITH COLLECT (p2) as people
MATCH (p1:person)
WHERE ALL (p2 in people WHERE (p2)-[:married]-(p3) )
AND size ( (p2) -[:parent]->(p3) ) > 3
RETURN p2.name
```

<u>שאלה 5 – xml (10 נק')</u>

נתון מסמך xml המתעד צמחים שונים:

```
<CATALOG>
      <PLANT ZONE="4">
            <COMMON>Bloodroot</COMMON>
            <BOTANICAL>Sanguinaria canadensis</BOTANICAL>
            <LIGHT>Mostly Shady</LIGHT>
            <PRICE>$2.44</PRICE>
            <AVAILABILITY>27</AVAILABILITY>
      </PLANT>
      <PLANT ZONE="3">
            <COMMON>Columbine</COMMON>
            <BOTANICAL>Aquilegia canadensis</BOTANICAL>
            <LIGHT>Mostly Shady</LIGHT>
            <PRICE>$9.37</PRICE>
            <AVAILABILITY>3</AVAILABILITY>
      </PLANT>
      <PLANT ZONE="4">
            <COMMON>Cowslip</COMMON>
            <BOTANICAL>Caltha palustris</BOTANICAL>
            <LIGHT>Mostly Shady</LIGHT>
            <PRICE>$9.90</PRICE>
            <AVAILABILITY>2</AVAILABILITY>
      </PLANT>
      <PLANT ZONE="2">
            <COMMON>Cardinal Flower</COMMON>
            <BOTANICAL>Lobelia cardinalis</BOTANICAL>
            <LIGHT>Shade</LIGHT>
            <PRICE>$3.02</PRICE>
            <AVAILABILITY>4</AVAILABILITY>
      </PLANT>
</CATALOG>
```

א. כתבו שליפה ב-**Xpath** שתחזיר את מס' האיזור (ZONE) של הצמחים שהזמינות (AVAILABILITY) שלהם גדולה מ-5 (5 נק')

//catalog/plant_zone[@AVAILABILITY>5]/TITLE

ב. כתבו שליפה ב- **Xquery** שתחזיר את הכמות הזמינה הממוצעת של הצמחים שמחירם נמוך $6 (5 נק')

FOR $plant in //CATALOG

LET avg :=AVG($plant/AVALIABILITY)

WHERE $plant/PRICE < 6$

RETURN AVG(

שאלה 6 - java streams (10 נק')

נתונה מחלקה בשם Polisa המגדירה נתוני פוליסות. השמות במחלקה הם:

int polisaNum

Date trStart

Date trEnd

String polisaType

int status

double premia

long customerId

האם ניתן לכתוב קטע קוד **שכולו** ב-java streams המקבל רשימה של פוליסות במבנה לעיל ומחזיר את סכום הפרמיות לכל סוג פוליסה? נמקו והוכיחו היטב!

polisot.Stream().map(p -> (p.polisaType,p.premia))

.colletct(Collectors.groupingBy(p -> p.first(), Collectors.summingDouble(p.second)))

שאלה 7 – Spark (8 נק')

איזה מהמשפטים הבאים **נכון** לגבי סוגי האובייקטים בSpark:

א. תמיד נעדיף להשתמש ב-RDD על פני dataframes

ב. כשאנחנו יודעים בדיוק מה נרצה לבצע, נעשה זאת בעזרת RDD. לעומת זאת, כשאנחנו לא יודעים מה בדיוק לבצע, נעשה זאת בעזרת dataframes.

ג. ניתן לממש את הפרדיגמה של map reduce רק בעזרת RDD

ד. ניתן להשתמש ב-dataframes כאשר הנתונים מאורגנים בצורה יחסית מובנית.

שאלה 8- Naïve Bayes (10 נק')

למדנו בכיתה על מודל Naïve Bayes המסייע לנו לסווג פריטים לקבוצות. אם אנחנו רוצים לייצר מודל חדש דומה למודל של Naïve Bayes אבל בשביל לשפר את ביצועיו נחליט שלא לעשות אותו נאיבי.

האם ואיזה מידע נוסף אנחנו צריכים שיהיה לנו בשביל שנוכל לבנות ולהריץ מודל כזה? (הסבירו ונמקו)

שאלה 9 - Linear regression + Logistic regression (8 נק')

איזה מהמשפטים הבאים נכון בהקשר של המשתנה α (אלפא) באלגוריתמים של רגרסיה לינארית ורגרסיה לוג'יסטית:

א. אלפא משמעותו קצב הלמידה ולכן לא משנה איך נאכלס אותו- אם נציב ערך יחסית גבוה קצב הלמידה יהיה גבוה ואם נציב ערך יחסית נמוך קצב הלמידה יהיה איטי יותר.

ב. אלפא משמעו קצב הלמידה, תמיד נבחר להציב בו את הערך 0.01 בשביל לקבל תוצאות מיטביות.

ג. לעיתים, בתהליך הלמידה, נצטרך לשנות את הערך של אלפא בשביל לשפר את המודל שלנו.

ד. אלפא ברגרסיה לינארית דומה לאלפא ברגרסיה לוג'יסטית אך לא זהה כי פונקציות הניבוי של 2 המודלים שונים.