

מבחן במסדי נתונים

מירב שקרון

7029210-2

סמסטר קיץ מועד ב' כ"ה תשרי התש"פ, 24.10.2019

הנחיות כלליות:

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

	1	2	3	4	5	6	7	8	Total
Max points	28	14	8	12	10	10	6	12	100
Grade									

ב ה צ ל ח ה !

שאלה 1-SQL (28 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 וקוד סוג הפוליסה

PolisaType (typeId, typeName)

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

PolisaCustomer (polisaNum, customerId, role)

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

rolesInPolisa(roleId, roleDesc)

טבלת תפקידים בפוליסה – קוד תפקיד בפוליסה ותיאור התפקיד בפוליסה (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 בפועל

א. כתבו שאילתא שתחזיר את רשימת הלקוחות (שם פרטי, שם משפחה, ת.ז) שתבעו בפוליסות שהם בתפקיד "משלם" (10 נק')

SELECT c.fname, c.lastName , c.ID

FROM Customer as c JOIN PolisaCustomer as pc
ON (c.id = pc.customerID) JOIN rolesInPolisa as rp

ON (rp.roleID = pc.role)


WHERE rp.description='payer' AND c.polisaNum IN (SELECT t.polisaNum

FROM Tvia as t)

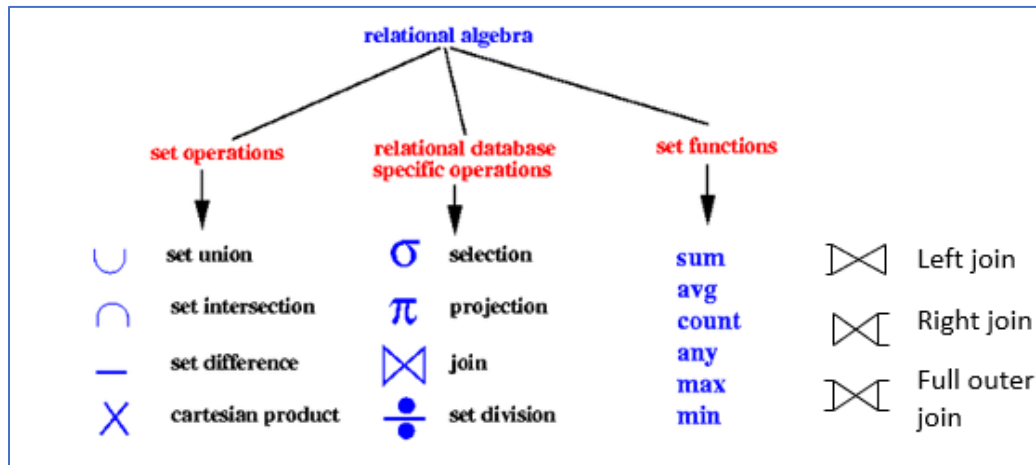
GROUP BY c.ID

ב. כתבו שאילתא שתחזיר את רשימת הפוליסות (מס' פוליסה, סוג הפוליסה, שם פרטי של הלקוח המשלם בפוליסה) שבהם נפתחה יותר מתביעה אחת. (10 נק')

```
SELECT  p.polisaNum, p.type, c.fName
FROM customer as c JOIN PolisaCustomer as pc
      ON (c.id = pc.customerID) JOIN Polisa as p ON (p.polisaNum = pc.polisaNum)
WHERE (SELECT count(*)
      FROM Tvia as t
      WHERE t.polisaNum = p.polisaNum) > 1
```

ג. כתבו שאילתא באלגברה רלציונית שתחזיר את רשימת הפוליסות שבהם יש את כל סוגי בעלי התפקידים האפשריים. השתמשו בסימן . (8 נק')

להזכירכם, אלו הסימונים שלמדנו:

[illegible]

שאלה 2 - Normalization (14 נק')

נתונה הרלציה $R(A,B,C,D,E,H)$

והתלויים הבאות:

 $AB \rightarrow CD$ $D \rightarrow HB$ $C \rightarrow H$ $CH \rightarrow AE$ $BC \rightarrow D$

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

 $(A,B), (B,C)$

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

1NF

H which is a non-prime is dependent on C which is a sub key of prime attribute

ג. האם ניתן להסיר את אחת התלויות כך ועדיין התשובות לסעיפים הקודמים (א' ו-ב') לא ישתנו? אם כן, מהי התלות? נמקו! אם לא, נמקו! (6 נק')

D -> HB

שאלה 3 - Tf-idf (8 נק')

נתונים 2 המשפטים הבאים:

- סוס טרף גמל

- גמל טרף סוס

אם נשתמש בנוסחת ה-tf-idf, מה נוכל לומר גבי התוצאות של המשפטים? נמקו!

להזכירכם, זוהי נוסחת ה-tf-idf שלמדנו בכיתה-

$$tfidf(d) = \sum_{k=0}^{|Q|} \frac{\#k \text{ in } d}{|d|} \log \left(\frac{|D|}{\#D \text{ with } k} \right)$$

we wouldnt be able to say anything for this equation does not give matter to location

but only to the number of words in a sentence, the count of words.

the result will be the same for both of them.

also there is not anything to compare to.

א. המירו את **כל המידע** המופיע בסיפור לעיל לטבלה בRDF, אך אין לכלול מידע מיותר הניתן להסקה מהטבלה: (לדוגמא, יוסי הוא גבר, וגבר זה סוג של אדם, אז אין לציין בנוסף שיוסי הוא אדם). (6 נק')

This image shows a blank sheet of white paper with horizontal ruling lines. The lines are evenly spaced and run across the width of the page. There are no margins, text, or other markings on the paper.

ב. כתבו שאילתא SPARQL שבהינתן טבלה מחזירה את כל האנשים יחד עם כל הספרים שנכללים בקטגוריה שהם אוהבים. אם יש יותר מספר אחד בקטגוריה שמישהו/ אוהב/ת, הוא/היא צריכה/ להופיע מספר פעמים בתשובה, פעם אחת עם כל ספר. למשל בדוגמא לעיל, התשובה צריכה לכלול את 2 הספרים שדינה אוהבת. (6 נק')

This image shows a single sheet of white paper with horizontal blue or grey ruling lines. The lines are evenly spaced and run across the width of the page. There are no margins, text, or other markings on the paper.

שאלה 5 – spark (10 נק')

נתונה רשימת פוליסות שקיימות בחברת "ביטוחי בע"מ" המתועדות במבנה של json.

דוגמא לרשומה אחת ברשימה:

```
{
  polisaNum : "12345678900",
  trStart : new Date("Jan 01, 2019"),
  trEnd : new Date("Dec 31, 2019"),
  polisaType : "Apartments",
  status : "Deleted",
  premia : 1,258,
  customers : [ {id: "257812589", role: "payer"},
                 {id: "278119536", role: "owner"} ]
}
```

רוצים לשלוף את סך סכומי הפרמיות של הפוליסות שבתוקף מקובצות לפי סוג הפוליסה.

באיזה מבנה נתונים נוכל להשתמש (מבין אלה שלמדנו ב-spark) ואיזה מהם עדיף?

הסבירו ונמקו! (תשובה ללא נימוק לא תתקבל)

We will use DataFrames as they are organized and at time have a relational connections

that may look like a table . JSON file can be converted into this exactly.

We use that when we want to tell spark what to do, but not exactly how to perform that action.

SPARK will translate the dataframe into RDD by itself

DF is designed to make large data sets processing even easier.

שאלה 6 - java streams (10 נק')

הניחו שקיימת לנו ב-java רשימה המכילה בתי קולנוע ברשת מסוימת, כולל פירוט של האולמות שבכל בית קולנוע.

מבנה אובייקט בית קולנוע:

```
class Cinema {
    int cinemald;
    String name;
    String city;
    List<Hall> halls;
}
```

מבנה אובייקט אולם:

```
class Hall {
    int hall_id;
    int num_of_seats;
    int amount;
    bool is_accessible;
}
```

כתבו קטע קוד ב-java streams **בלבד**, המקבל את רשימת בתי הקולנוע ומחזיר את מס' בתי הקולנוע שבהם כל האולמות מוגשים

```
cinemas.stream().filter(c -> c.halls.stream().foreach(h->h.is_accessible)).map(c -> c.cinemaID)
.collect(Collectors.toList())
```

[illegible]

שאלה 7 – למידת מכונה (6 נק')

נתונות תוצאות שהתקבלו בבדיקת מודל Naïve Bayes על test set. המודל נבנה לצורך סיווג סטודנטים לאלה שיעברו במבחן מסדי נתונים לבין אלה שיכשלו במבחן.

	Classified as Pass	Classified as Fail
Really Pass	42	16
Really Fail	13	29

א. חשבו את ה-Accuracy (2 נק').

42+29 / 55+45

ב. חשבו את ה-recall לעבור את המבחן (2 נק').

42 / 55

ג. חשבו את ה-Precision לעבור את המבחן (2 נק').

42 / 58

שאלה 8 - Linear regression + Logistic regression (12 נק')

בהינתן הנתונים והתיאורים הבאים:

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

פונקציית ה-loss שלמדנו היא :

$$h(x) = xw + b$$

פונקציית הניבוי היא :

$$x = [3, 2, 0, -4]$$

$$y = [2, 1, 1, -2]$$

(הניחו שימוש ב-gradient decent)

א. מה הערכים של הגרדיאנט כאשר $w=0$, $b=0$? (5 נק')

ב. הניחו שאלפא שווה 0.1, מה יהיו הערכים של w ו- b באיטרציה הבאה? (4 נק')

ג. לפי הערכים שמצאתם בסעיף הקודם, מה יהיה הניבוי ל-3? (3 נק')
מתי הניבוי ל-3 מדויק יותר כש- $w=0, b=0$ או עם הערכים שמצאתם בסעיף הקודם? (יש להראות חישוב)
(5 נק')
