

**מבחן במסדי נתונים**

**עמוס עזריה ומירב שקרון**

**7029210-1,3,4,5**

סמסטר ב' מועד א' ח' בתמוז התשע"ט, 11.7.2019

**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.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Total | 8 | 7 | 6 | 5 | 4 | 3 | 2 | 1 |  |
| 102 | 18 | 8 | 10 | 8 | 10 | 8 | 20 | 20 | Max points |
|  |  |  |  |  |  |  |  |  | Grade |

**ב ה צ ל ח ה!**

שאלה 1-SQL (20 נק')

נתונות הטבלאות הבאות לייצוג רשת בתי קולנוע:

cinema(cinema\_id, name, city)

hall(cinema\_id, hall\_num, seats)

film(film\_id, film\_name, director, type, length)

film\_hall(cinema\_id, hall\_num, film\_id, date, hour)

כיתבו את השאילתות הבאות:

1. לכל בית קולנוע מס' הסרטים מסוג קומדיות שהוקרנו בו (יש להציג את שם בית הקולנוע)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**Select c.name, count(film\_id)**

**From cinema as c join film\_hall as FH on c.cinema\_id = FH.cinema\_id**

**Join film as f on FH.film\_id = f.film\_id**

**where type = "comedy“**

**Group by c.cinema\_id**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

1. מס' הסרטים שהוקרנו בבתי קולנוע בהם מס' האולמות גדול מ-4

Select count(film\_id)

From film\_hall

where cinema\_id in

(Select cinema\_id

From hall

Group by cinema\_id

Having count(hall\_num) > 4)

Or

SELECT count(\*)

From cinema as c JOIN film\_hall as fh

ON ( c.cinema\_id = fh.cinema\_id)

WHERE (SELECT count(\*)

FROM hall as h

WHERE h.cinema\_id = fh.cinema\_id

GROUP BY h.hall\_num) > 4

שאלה 2 - Normalization + Relational DB + Cassandra (20 נק')

הניחו שעליכם ליצור בסיס נתונים לצורך ניהול מדפסות באוניברסיטה.

לכל מדפסת יש מספר מדפסת ייחודי, חברה, ודגם (יכולות להיות כמובן מספר מדפסות של אותה החברה ואותו הדגם). לכל מדפסת יש גם חדר בו היא נמצאת (ניתן להניח שמספר החדר הוא ייחודי, אבל יכולות להיות מספר מדפסות באותו החדר), החדר שייך לפקולטה מסויימת (פקולטה אחת בלבד).

לכל מדפסת יש מישהו/י שאחראי/ת עליה, לו/ה יש שם פרטי (בודד), שם משפחה (בודד), מספר(י) טלפון (אחד או יותר) ותעודת זהות (ייחודית).

**יש לכלול את כל המידע המופיע לעיל, אך אין צורך לכלול שום מידע נוסף שלא מפורט לעיל.**

1. אילו רלציות (טבלאות ותכונות) הייתם יוצרים לצורך דאטה-בייס במבנה הרלציוני? (יש לוודא רמת נרמול 4NF) (12 נק')

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

1. כעת הניחו שמדובר בבסיס נתונים ענק אותו אתם רוצים לבנות בCassandra. הניחו שמתבצעות אלפי קריאות (שאילתות) בדקה שנותנות את מספר החדר ומבקשות את כל מספרי הטלפון של האחראים על כל המדפסות באותו החדר. כמו כן, ישנן אלפי קריאות בדקה שנותנת שם ושם משפחה של אחראי/ת ומבקשות את כל מספרי הטלפון של האחראי/ת (ייתכן שישנם מספר אחראים עם אותו השם, במקרה זה יש להחזיר את מספרי הטלפון של כולם).

האם הייתם יוצרים רלציות שונות לצורך דאטה-בייס זה בCassandra? אם כן, פרטו, אם לא, נמקו. (8 נק')

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

שאלה 3 - Tf-idf (8 נק')

בכיתה ראינו את הנוסחא הבאה לחישוב tf-idf 

מהו הבסיס של ה-log בנוסחא זו:

1. כיוון שמדובר במימוש במדעי המחשב, סימון log ללא בסיס הכוונה היא לבסיס 2. כל מימוש אחר, עלול לגרור כמובן תוצאה שגוייה.
2. סימון log ללא בסיס הכוונה היא לבסיס 10, זה הבסיס המקובל (לבסיס הטבעי משתמשים בln). כל מימוש אחר, עלול לגרור כמובן תוצאה שגוייה.
3. כיוון שמדובר בנוסחא הסימון המקובל נלקח ממתמטיקה, לכן סימון log ללא בסיס הכוונה היא לבסיס הטבעי (e). כל מימוש אחר, עלול לגרור כמובן תוצאה שגוייה.
4. כיוון שמדובר בערך שמשמש רק לדירוג אפשר להשתמש בכל בסיס שרוצים (כל עוד משתמשים תמיד באותו הבסיס). תוצאות הדירוג יצאו זהות לחלוטין בכל בסיס.

שאלה 4 - neo4j (10 נק')

נתון בסיס נתונים גרפי של neo4j, המתאר רשת חברתית המכילה את סוגי הצמתים – person, post, group ובינהם הקשרים האפשריים הבאים:

בין 2 צמתים מסוג person יכול להיות הקשר **friendOf**

בין person ל-Post יכול להיות הקשר **likes** והקשר **posted**

בין person ל-group יכול להיות הקשר **partOf**

כתבו שאילתא ב- cypher המחזירה את שמות כל האנשים שעשו like **לכל** הפוסטים שפורסמו ע"י מיכל.

MATCH (m:Person {name:'Michal'})-[:Posted]->(po:Post)

WITH COLLECT(po) as posts

MATCH (p:Person)

Where ALL (post in posts WHERE (p)-[:Likes]->(post))

RETURN p2.name

שאלה 5 – xQuery (8 נק')

נתון מסמך xml המתאר תפריט במסעדה. לדוגמא:

<breakfast\_menu>

<food>

<name>Belgian Waffles</name>

<price>$5.95</price>

<description>

Two of our famous Belgian Waffles with plenty of real maple syrup

</description>

<calories>650</calories>

</food>

<food>

<name>Strawberry Belgian Waffles</name>

<price>$7.95</price>

<description>

Light Belgian waffles covered with strawberries and whipped cream

</description>

<calories>900</calories>

</food>

...

<drinks>

...

</drinks>

</breakfast\_menu>

כתבו שאילתא ב-xQuery (ולא ב-xPath) שתחזיר את שמות המאכלים (ללא משקים) שכמות הקלוריות בהם גבוהה מהממוצע הקלורי של כל המאכלים במסעדה (ללא המשקים), מסודרות בסדר עולה לפי מחיר המאכל.

**f**or $x in /breakfast\_menu/food

**l**et $avg := avg(/breakfast\_menu/food /calories)

**w**here $x/calories > $avg

order by $x/price

**r**eturn $x/name

שאלה 6 - java streams (10 נק')

כתבו קטע קוד ב-java תוך שימוש ב-java stream בלבד, המקבל איבר ראשון בסדרה הנדסית (a1) ,מס' האיברים (n) ואת מנת הסדרה (q) ומחשב את מכפלת איברי הסדרה.

לדוגמא- עבור הנתונים הבאים a1=3 ; n=5 ; q=-2, מתקיימת הסדרה – 3,-6,12,-24,48 ומכפלתה הוא 248832.

IntStream.range(0,n)

.mapToDouble(x-> a1\*Math.pow(q,x))

.reduce(1,(x,y)->x\*y)

שאלה 7 - Naïve Bayes (8 נק')

לפי Bayes rule: p(y|x) = p(y)\*p(x|y) / p(x)

מדוע במודל של Naïve Bayes לא מחשבים הערכה ל p(x) (אלא רק לp(x|y) ולp(y))?

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

שאלה 8 - Linear regression + Logistic regression (18 נק')

הניחו שנתונה דאטה בגודל 1000 דוגמאות (700 בתור train ו300 בתור test). לכל דוגמא בדאטה יש 5 פיצ'רים: גיל, מספר ילדים, מספר נעליים, גובה, משקל, הניחו שאין שום תוספת של פיצ'רים (כלומר לא מוסיפים את הגיל בריבוע, ולא מוספים שום צירוף של מספר פיצ'רים וכד'). הניחו שלכל דוגמא נתון גם תוחלת החיים והאם האדם בריא או חולה, אך עלינו לבנות מודלים שמנסים לחזות מידע זה, (כלומר מידע זה לא משמש כפיצ'רים אלא כlabel - הY).

1. כמה פרמטרים יצטרך ללמוד מודל מסוג Linear regression שחוזה את תוחלת החיים (בשנים)? (נמקו) (5 נק')

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

1. כמה פרמטרים ייצטרך ללמוד מודל מסוג Logistic regression שמסווג לבריא/חולה? (נמקו) (5 נק')

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

1. נניח כי במודל הLinear regression שנלמד הw שכופלת את "מספר ילדים" נמצאה להיות 5- (מינוס חמש), איזו מהטענות הבאות נכונה? (הניחו שהמודל הנלמד חוזה את תוחלת החיים בצורה כמעט מדוייקת) (8 נק')
2. מהדאטה הנתונה נראה שילדים מפחיתים את תוחלת החיים, ייתרה מכך, נראה שכל ילד/ה נוספ/ת מפחית/ה את תוחלת החיים ב5 שנים.
3. מהדאטה הנתונה נראה שילדים מעלים את תוחלת החיים, ייתרה מכך, נראה שכל ילד/ה נוספ/ת מעלה את תוחלת החיים ב5 שנים.
4. לא ניתן להסיק כל קשר בין מספר הילדים לבין תוחלת החיים, יש לראות את תוצאות המודל כולו כדי להסיק קשר כלשהו (אם קיים).
5. מהדאטה הנתונה אפשר לומר רק באופן כללי ביותר שילד/ה בודד/ת מפחית/ה את תוחלת החיים, אך יש לראות את תוצאות המודל כולו כדי להסיק קשר כלשהו.
6. מהדאטה הנתונה אפשר לומר רק באופן כללי ביותר שילד/ה בודד/ת מעלה את תוחלת החיים, אך יש לראות את תוצאות המודל כולו כדי להסיק קשר כלשהו.