1



מבחן במסדי נתונים

**עמוס עזריה ומירב שקרון**

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

סמסטר ב' מועד ב' יט' באב התשפ"א, 28.7.2021

**הנחיות כלליות:**

 משך הבחינה: 150 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 | Total |
| Max points | 28 | 19 | 10 | 15 | 15 | 18 | 105 |
| Grade |  |  |  |  |  |  |  |



**ב ה צ ל ח ה !**

) 'נק 28( SQL -1 שאלה

נתונות הטבלאות הבאות המתארות נתונים מקופת חולים מסוימת הקשורים למחלת הקורונה.

(מפתחות הטבלה מסומנות עם קו תחתון)

patients(id, first\_name, last\_name, birth\_date, gender, first\_dose\_date, second\_dose\_date)

[טבלת מטופלים: תז, שם פרטי, שם משפחה, תאריך לידה, מגדר, תאריך מנת חיסון ראשונה, תאריך מנת חיסון

שניה]

isolated(patient\_id, start\_date, end\_date, isolation\_type\_id)

[טבלת מבודדים: תז, תאריך כניסה לבידוד, תאריך סיום הבידוד, מזהה סוג בידוד]

isolation\_types(id, desc)

[טבלת סוג הבידוד: מזהה סוג בידוד, תיאור סוג בידוד]

הערכים בטבלה זו הם:

|  |  |
| --- | --- |
| **id** | **desc** |
| 1 | חדר ושירותים נפרדים. בבית גרים אנשים נוספים |
| 2 | חדר נפרד בבית. שירותים עם שאר יושבי הבית |
| 3 | לבד בבית |
| 4 | מלונית |

confirmed\_positive\_cases(patient\_id, start\_date, is\_symptomatic, end\_date)

[טבלת חולים מאומתים: תז, תאריך תחילת מחלה, האם יש סימפטומים, תאריך סיום המחלה]

pcr\_test(patient\_id, test\_date, status, result, result\_date, test\_reason)

[טבלת בדיקות קורונה: תז, תאריך בדיקה, סטטוס בדיקה, תוצאות הבדיקה, תאריך תוצאות, סיבה להגעה

לבדיקה]

א. כיתבו שאילתא שתחזיר עבור כל בדיקה שבוצעה את תז של הנבדקת, תאריך הבדיקה ומספר

הבדיקות הכולל שבוצעו באותו יום (לכלל הנבדקים)

דוגמא למבנה התוצאה של השאילתא:

|  |  |  |
| --- | --- | --- |
| **מס' הבדיקות הכולל שבוצעו בתאריך**  **זה** | **תאריך בדיקה** | **תז** |
|  |  |  |
|  |  |  |
|  |  |  |

)'נק 11(

Select p.id, pcr.test\_date, agg.tests\_counter

From patients as p join pcr\_test as pcr

On(p.id=pcr.patient\_id) join

(Select test\_date, count(\*) as tests\_counter

from pcr\_test group by test\_date) as agg

on(pcr.test\_date = agg.test\_date);

ב. לצורך מחקר נדרש להוציא דוח המפרט את שמות שלושת מקומות הבידוד הפופולריים ביותר בקרב

מי שקיבלו תשובה חיובית במהלך בידודם , ומספר הבידודים הכולל שבוצעו במקום זה .

לפי הפופולריות (כך שמקום הבידוד בעל מספר הבידודים הגבוה יופיע

על הדוח להיות ממוין

ראשון.)

מבנה הדוח :

|  |  |
| --- | --- |
| **מספר הבידודים שנעשו במקום זה** | **מקום הבידוד** |
|  |  |
|  |  |
|  |  |

כתבו את השאילתא של הדוח הנ"ל 11( נק)'

Select isolation\_type\_id, count(\*) as isolations

from isolated group by (isolation\_type\_id)

order by isolations desc

ג. שימו לב שאם נבדקת כלשהי מקבלת תשובה חיובית לבדיקת ה,pcr\_test- היא צריכה להתווסף כרשומה בconfirmed\_positive\_case- . כיצד ניתן לעשות זאת באופן אוטומטי? יש לכתוב קוד

כדי לקבל ניקוד מלא. 6( נק)'

Delimiter $$

Create trigger positive\_pcr\_recieved

After insert on pcr\_test

For each row

begin

if new.result=1 then

Insert into confirmed\_positive\_case(patient\_id, start\_date)

values(new.patient\_id, new.result\_date)

End if;

End; $$

Delimiter;

) 'נק 19( Normalization – 2 שאלה

נגדיר את הform- Normal הבא, בשם :AANF

נאמר שרלציה R עונה על AANF אם היא מקיימת את שני התנאים הבאים:

.1 הרלציה ב1NF

.2 לכל תלות ,X->Y מתקיים ש X אינו תת קבוצה חלקית ממש לkey .candidate

שימו לב שאין כל דרישה על .Y

AANF אך לא על ?2NF אם כן, תנו דוגמא והסבירו, אם לא הוכיחו /

א. האם קיימות רלציות שעונות על

נמקו היטב.

ב. האם קיימות רלציות שעונות על AANF אך לא על ?3NF אם כן, תנו דוגמא והסבירו, אם לא הוכיחו /

נמקו היטב.

R(A,B,C,D)

(A,B)->C

C->D

R violates 3NF because D is non prime and depends on C but C is non-prime then its not a super key.

R holds AANF, C depends on a full candidate key and D depends on C which is not a subset of the candi-key.

ג. האם קיימות רלציות שעונות על 3NF אך לא על ?AANF אם כן, תנו דוגמא והסבירו, אם לא הוכיחו /

נמקו היטב.

R(A,B,C,D,E)

(A,B,E)->)C,D)

(A,D)->B

Candi-keys: (A,B,E),(A,D,E) no non-prime attributes then 2,3NF holds trivially.

R doesn’t holds AANF since B depends on (A,D) which is a subset of (A,D,E)

which is a candi-key as mentioned

שאלה 3 - neo4j 10( נק' ) נתון בסיס נתונים גרפי של ,neo4j המתאר רשת של מידע על אנשים שביקרו בכל מיני מקומות, לצורך השתלטות

על מגיפת הקורונה.

סוגי הצמתים והמידע על כל הצומת:

 צומת-**Person** מכילה את התכונות הבאות- HealthStatus Age, Name, Id, (תז, שם מלא, מצב

בריאותי- חולה/בריא בקורונה)

 צומת-**Place** מכילה את התכונות הבאות- Type Name, Id, (מזהה מקום, שם מקום, סוג מקום-

מסעדה/ בר/ קניון/ בית חולים/ פארק/ בית ספר/ בית קולנוע)

בין צומת Person לצומת Place יכול להיות מוגדר הקשר – **visited\_at**

עליכם לכתוב שאילתא ב- cypher המחזירה את שמות כל האנשים החולים שביקרו ביותר מ3- מקומות.

match (m:Person)-[:visited\_at]->(p:Place) with m,count(\*) as c where c>3 return m.name

שאלה 4 – spark 15( נק' ) נתון קובץ המכיל תמלילים של שיחות של מאומתי קורונה עם אפידימיולוגית בהן המאומתים מספרים מה עשו

והיכן שהו ב3- ימים האחרונים לפני שאומתו.

המילה הראשונה בכל שורה מכילה את תז של המאומתת והמשך השורה מכילה את התיאור כפי שניתן על ידה.

שימו לב, כל שורה מכילה תיאור של מאומתת אחת.

כמו כן, נתונה רשימת מקומות ואירועים חשובים לחקירת התפשטות המגיפה. הרשימה מיוצגת כlist- בפייטון. לדוגמא- בר, מסעדה, קולנוע, חתונה וכו' . ניתן להניח שכל האיברים ברשימה הם מחרוזות בנות מילה אחת (ללא

רווחים.)

הפונקציה הבאה מקבלת מחרוזת ומחזירה list בפייטון שמכיל tuples של המילה הראשונה במחרוזת וכל אחת

מיתר המילים במחרוזת:

def firstWordWithRest(line): words = line.split)(

first = ((words[0]+ " ")\*len(words)).split)( return zip(first[1:], words[1:])

כתבו קטע קוד בpyspark- המקבל את כל תמלילי השיחות (שמור בקובץ בשם- )conversations ואת המשתנה

places\_events של רשימת המקומות והאירועים ומחזיר את מס' המקומות והאירועים ששהה בהם כל אדם. כלומר, יש להחזיר לכל מאומתת את תז שלה ומספר המציין את סך המקומות מהרשימה המופיעים בתמליל שלה

(אם מקום מופיע פעמיים ניתן לספור אותו כשניים.)

conversations\

.flatMap(firstWordWithRest)\

.filter(lambda t:t[1] in places\_events)\

.map(lambda t: (t[0],[t[1]]))\

.reduceByKey(lambda w1, w2: w1+w2)\

.map(lambda t: (t[0], len(t[1])))\

.collectAsMap()

שאלה 5 - TF-IDF + Bayes Naïve 15( נק' ) לבעיית ה,Spam/Real כלומר בהינתן דאטה מתוייגת, ובהינתן הודעה חדשה )Query( יש לדעת האם היא

Spam או ,Real הציעה גילה את הClassifier הבא: לבחון את ה,Query ולדרג את כל ההודעות לפי ציון הTF-IDF שלהם, ואז לסווג את ההודעה לפי הClass של

/35ההודעה עם ציון הTF-IDF הגבוה ביותר. האם קיימים דאטה וQuery בהן ההצעה של גילה תחזיר תוצאה השונה מזו של Bayes ?Naïve אם כן, תנו

דוגמא מלאה (הכוללת דאטה, Query והסיווג שלה לפי ההצעה של גילה ולפי Bayes ,)Naïve אם לא, הוכיחו /

נמקו היטב.

להזכירכם:

|𝑄|

𝑡𝑓𝑖𝑑𝑓(𝑑) = ∑

#𝑘 𝑖𝑛 𝑑

log (

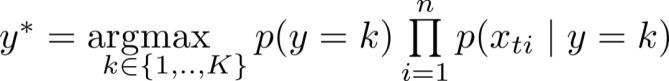
|𝐷|

)

𝑘=0

|𝑑|

#𝐷 𝑤𝑖𝑡ℎ 𝑘



‘get discount today’

’pay now come save money fast and don’t miss the chance’

’a meeting in the office today’

’come as you are

Query = come today

= − 1

) 'נק 18( Logistic regression - 6 שאלה

:תזכורת h(xi) = 1/(1+exp(-(w1xi1+w2xi2+…+b))) : logistic regressionב

𝑚 (𝑦 (log(ℎ(𝑥 )) + (1 − 𝑦 )log (1 − ℎ(𝑥 )) :היא logistic regressionב הטעות פונקציית

∑

𝑖

𝑖

𝑖

𝑖

𝑚 𝑖=1

1 𝑚

= ∑ 𝑥𝑖𝑗(ℎ(𝑥𝑖) − 𝑦𝑖) היא wj לפי logistic regressionב הטעות פונקציית של הנגזרת

𝑚

𝑖=1

נניח שאימנו מודל לצורך סיווג דוגמאות לשתי מחלקות, כאשר לכל דוגמא יש 3 פיצ'רים. לאחר האימון (כאשר

פונקציית הטעות התכנסה למינימום) קיבלנו:

w1=3 ,w2=-1 ,w3=2 ,b=5

(שימו לב שw2 שלילי.)

א. מה יהיה הסיווג של המודל ל ?x1=(0,1,2) נמקו את תשובתכם. 6( נק)'

.1 הסיווג יהיה 0 בוודאות

.2 הסיווג יהיה 1 בוודאות

.3 הסיווג עשוי להיות 0 או ,1 והוא תלוי בגודל הצעד

.4 הסיווג עשוי להיות 0 או ,1 הוא תלוי בגורמים נוספים שלא נתונים בשאלה, אך לא תלוי ב

Ans: 2 (the classification will be certainly 1) explanation:

ב. מה יהיה הסיווג של המודל ל ?x2=(0,1,-1) נמקו את תשובתכם. 6( נק)'

.1 הסיווג יהיה 0 בוודאות

.2 הסיווג יהיה 1 בוודאות

.3 הסיווג עשוי להיות 0 או ,1 והוא תלוי בגודל הצעד

.4 הסיווג עשוי להיות 0 או ,1 הוא תלוי בגורמים נוספים שלא מופיעים בשאלה, אך לא תלוי ב

Then classification will be certainly 1. Ans:2

ג. קיבלנו את הדוגמא החדשה הבאה: ,x3=(3,2,0) עם התיוג 0 (הlabel- האמיתי של הדוגמא הוא .(0 החלטנו לעדכן את הפרמטרים שלנו w,b על סמך הדוגמא הזו בלבד (צעד אחד של stochastic

descent gradient על סמך דוגמא אחת.)

איזה מבין המשפטים הבאים נכון? נמקו את תשובתכם. 6( נק)'

.1 w1 לא ישתנה כלל, אבל הפרמטרים האחרים ככל הנראה כן ישתנו .2 w2 לא ישתנה כלל, אבל הפרמטרים האחרים ככל הנראה כן ישתנו .3 w3 לא ישתנה כלל, אבל הפרמטרים האחרים ככל הנראה כן ישתנו .4 b לא ישתנה כלל, אבל הפרמטרים האחרים ככל הנראה כן ישתנו

.5 כיוון שהתיוג של הדוגמא הוא ,0 בכל מקרה, כל הפרמטרים צפויים להשאר ללא שינוי

calculating x3 in gradient gives:

Then ans: 3