אוניברסיטת אריאל בשומרון
המחלקה למדעי המחשב

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 !

03.9066640
03.9066692
אריאל, 40700
אוניברסיטת אריאל בשומרון ע"ר

שאלה 1- SQL (28 נק')

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:

תז	תאריך בדיקה	מס' הבדיקות הכולל שבוצעו בתאריך זה

(11 נק')

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 .
על הדוח להיות ממוין לפי הפופולריות (כך שמקום הבידוד בעל מספר הבידודים הגבוה יופיע ראשון).
מבנה הדוח :

מקום הבידוד	מספר הבידודים שנעשו במקום זה

כתבו את השאילתא של הדוח הנ"ל (11 נק')

ג. שימו לב שאם נבדקת כלשהי מקבלת תשובה חיובית לבדיקת ה-pcr_test, היא צריכה להתווסף כרשומה ב-confirmed_positive_case . כיצד ניתן לעשות זאת באופן אוטומטי? יש לכתוב קוד כדי לקבל ניקוד מלא. (6 נק')

<u>שאלה 2 – Normalization (19 נק')</u>

נגדיר את ה-Normal form 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 לcandidate key.

שימו לב שאין כל דרישה על Y.

א. האם קיימות רלציות שעונות על AANF אך לא על 2NF? אם כן, תנו דוגמא והסבירו, אם לא הוכיחו / נמקו היטב.

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

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

<u>שאלה 3 - neo4j (10 נק')</u>

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- Id, Name, Age, HealthStatus (תז, שם מלא, מצב בריאותי- חולה/בריא בקורונה)
- צומת **Place**- מכילה את התכונות הבאות- Id, Name, Type (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 מ-3 מקומות.

<u>שאלה 4 – spark (15 נק')</u>

נתון קובץ המכיל תמלילים של שיחות של מאומתי קורונה עם אפידימיולוגית בהן המאומתים מספרים מה עשו והיכן שהו ב-3 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 כ-list 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 ב-pyspark 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).

<u>שאלה 5 - Naïve Bayes + TF-IDF (15 נק')</u>

לבעיית הSpam/Real, כלומר בהינתן דאטה מתוייגת, ובהינתן הודעה חדשה (Query) יש לדעת האם היא Spam או Real, הציעה גילה את הClassifier הבא:

לבחון את הQuery, ולדרג את כל ההודעות לפי ציון הTF-IDF שלהם, ואז לסווג את ההודעה לפי הClass של ההודעה עם ציון הTF-IDF הגבוה ביותר.

האם קיימים דאטה וQuery בהן ההצעה של גילה תחזיר תוצאה השונה מזו של Naïve Bayes? אם כן, תנו דוגמא מלאה (הכוללת דאטה, Query והסיווג שלה לפי ההצעה של גילה ולפי Naïve Bayes), אם לא, הוכיחו / נמקו היטב.

להזכירכם:

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

$$y^* = \underset{k \in \{1,..,K\}}{\operatorname{argmax}}\ p(y=k) \prod_{i=1}^{n} p(x_{ti} \mid y=k)$$

<u>שאלה 6 - Logistic regression (18 נק')</u>

תזכורת:

בlogistic regression : $h(x_i) = 1/(1+\exp(-(w_1x_{i1}+w_2x_{i2}+\ldots+b)))$

פונקציית הטעות בlogistic regression היא: $= -\frac{1}{m}\sum_{i=1}^{m}(y_i(\log(h(x_i)) + (1-y_i)\log(1-h(x_i))$

הנגזרת של פונקציית הטעות בlogistic regression לפי w_j היא $= \frac{1}{m}\sum_{i=1}^{m} x_{ij}(h(x_i) - y_i)$

נניח שאימנו מודל לצורך סיווג דוגמאות לשתי מחלקות, כאשר לכל דוגמא יש 3 פיצ'רים. לאחר האימון (כאשר פונקציית הטעות התכנסה למינימום) קיבלנו:

$w_1=3$,$w_2=-1$,$w_3=2$,$b=5$

(שימו לב שw_2 שלילי).

א. מה יהיה הסיווג של המודל ל $x_1=(0,1,2)$? 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 α
4. הסיווג עשוי להיות 0 או 1, הוא תלוי בגורמים נוספים שלא נתונים בשאלה, אך לא תלוי בα

_____

_____

_____

_____

_____

_____

ב. מה יהיה הסיווג של המודל ל $x_2=(0,1,-1)$? 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 α
4. הסיווג עשוי להיות 0 או 1, הוא תלוי בגורמים נוספים שלא מופיעים בשאלה, אך לא תלוי בα

_____

_____

_____

_____

_____

_____

ג. קיבלנו את הדוגמא החדשה הבאה: $x_3=(3,2,0)$, עם התיוג 0 (ה-label האמיתי של הדוגמא הוא 0).
החלטנו לעדכן את הפרמטרים שלנו w,b על סמך הדוגמא הזו בלבד (צעד אחד של stochastic gradient descent על סמך דוגמא אחת).
איזה מבין המשפטים הבאים נכון? נמקו את תשובתכם. (6 נק')

1. w_1 לא ישתנה כלל, אבל הפרמטרים האחרים ככל הנראה כן ישתנו
2. w_2 לא ישתנה כלל, אבל הפרמטרים האחרים ככל הנראה כן ישתנו
3. w_3 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

_____

_____

_____

_____

_____

_____