אוניברסיטת אריאל בשומרון
המחלקה למדעי המחשב

מבחן במסדי נתונים

מר נתנאל שקרון

7029210-2,6

סמסטר קיץ מועד א' ח תשרי התשפ"ב, 14.09.2021

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	18	15	10	10	100
Grade							

ב ה צ ל ח ה !

03.9066640
03.9066692
אריאל, 40700
אוניברסיטת אריאל בשומרון ע"ר

שאלה 1- SQL (28 נק')

נתון מסד הנתונים הבא.

מפתח ראשי לטבלה מסומן בכוכבית *.

DEPARTMENT
*department_id
depratment_name
manager_id
loaction_id

EMPLOYEES
*employee_id
first_name
last_name
salary
email
job_id
manager_id
department_id

JOBS
*job_id
job_title
min_salary
max_salary

א. כתבו שאילתה המחזירה את שם העובד, המשכורת שלו , כותרת התפקיד שלו והמשכורת הממוצעת של כלל העובדים בחברה **באותו תפקיד** (14 נק- ניקוד מלא יינתן למימוש ע"י פונקציית חלון)

first_name	last_name	salary	job_title	avg_salary
Jose Manuel	Urman	7800.00	Accountant	7920.000000
Luis	Popp	6900.00	Accountant	7920.000000
Nancy	Greenberg	12000.00	Finance Manager	12000.000000
Susan	Mavris	6500.00	Human Resources Represen...	6500.000000

ב. כתבו שאילתה המחזירה את רשימת המחלקות בהן מספר העובדים גדול מ-3 ומספר העובדים בכל אחת מהן (10 נק)

ג. הסבירו איך ניתן להבטיח שלא יימחקו מחלקות שיש בהן עובדים (4 נק)

שאלה 2 – Normalization (17 נק')

נתונה הטבלה הבאה :

ת"ז סטודנט	התמחות	מנחה	משרד
111	Physics	ד"ר אורי כהנה	11.1.3
111	Music	ד"ר ישי קפלן	9.4.7
320	Math	ד"ר נויימן - חזון	9.3.1
671	Physics	ד"ר ישראל בוחבוט	11.2.6
803	Physics	פרופסור עומר אלפרוביץ'	10.2.2

תחת ההנחות :

- כל סטודנט יכול להתמחות ביותר מהתמחות אחת
- בכל התמחות שנבחרה לסטודנט יכול להיות רק מנחה אחד
- לכל התמחות יש מספר מנחים
- כל מנחה מנחה תחום התמחות אחת בלבד
- כל מנחה עשוי להנחות יותר מסטודנט אחד

א. מה/ם המפתח/ות האפשרי/ים של הטבלה ? (5 נק)

_____

_____

_____

_____

ב. באיזו רמת נרמול נמצאת הרלציה הנ"ל ? (5 נק)

_____

_____

ג. במידה והרלציה דורשת נרמול - תאר את שלבי הנרמול הנדרשים בפירוט והצג את מבנה הטבלאות המתקבל בכל שלב . (7 נק)

_____

_____

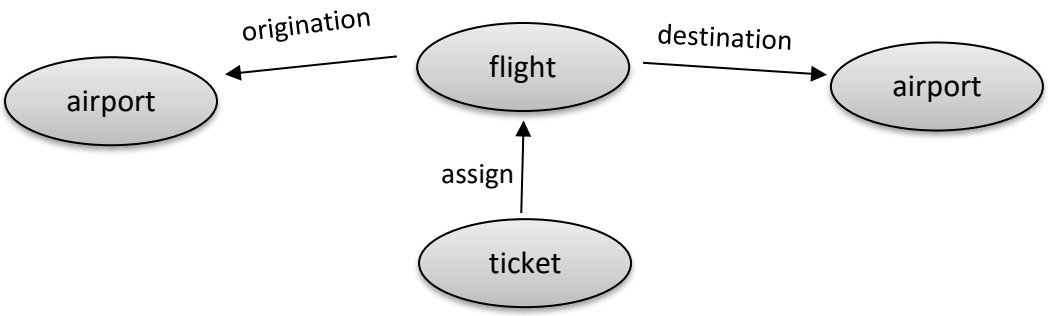
_____

_____

_____

שאלה 3 – Neo4J | NoSQl (18 נק')

א. נתון בסיס נתונים גרפי של neo4j, המתאר רשת של מידע על טיסות, שדות תעופה וכרטיסי טיסה.



<u>סוגי הצמתים והמידע על כל הצומת:</u>

צומת **Flight-** מכילה (בין השאר) את התכונות – תאריך ושעת טיסה, קוד חברת תעופה, משך הטיסה, מרחק הטיסה.

צומת **Airport-** מכילה (בין השאר) את התכונות- קוד שדה התעופה, מיקום שדה התעופה.

צומת **Ticket-** מכילה (בין השאר) את התכונות- מס' מושב בטיסה, שם המחלקה בטיסה (/first business/ economy), מחיר הכרטיס

<u>קשרים בין הצמתים:</u>

בין צומת Ticket לצומת Flight יכול להיות מוגדר הקשר- **Assign** – בשביל לתאר שיוך בין כרטיס לטיסה.

בין צומת Flight לצומת Airport יכול להיות מוגדרים הקשרים – 1. **Origination** – בשביל לתאר את הקשר בין שדה תעופה ממנו יוצאת הטיסה לטיסה עצמה. 2. **Destination-** בשביל לתאר את הקשר בין שדה התעופה אליו מגיע הטיסה לטיסה עצמה.

עליכם לכתוב שאילתה ב- cypher המחזירה את המחירים של טיסות **ישירות** שיוצאות משדה התעופה באורלנדו (**ORD**), עבור כרטיסים **במחלקת עסקים**.

עבור כל תוצאה יש להציג את יעד הטיסה, מחיר הכרטיס וקוד חברת התעופה.(12 נק)

לדוגמא:

destiantion	price	airlineCode
TLV	1000	LY
CDG	850	AF
JFK	300	AA

ב. אלו מאפיינים של בסיסי מידע מסוג noSQL הופכים אותם למתאימים לעיבוד Big Data -פרט ונמק (6 נק)

שאלה 4 – spark (15 נק')

נתון מסמך במבנה json המתאר פרטי טיסות ברחבי העולם.

להלן תיאור טיסה אחת מתוך מסמך הג'ייסון:

```
{
  "id": 123,
  "airlineCode": "LY",
  "origin": "TLV",
  "destination": "ORD",
  "duration": 5.5,
  "departureTime": "01/01/2021, 22:00",
  "arrivalTime": "02/01/2021, 04:00"
}
```

כתבו קטע קוד ב-spark (pySpark) המקבל את המסמך ומחזיר את מס' הטיסות המגיעות לכל יעד , עבור טיסות בהן משך הטיסה קטן מ-7 שעות.

(ניתן להניח שכבר קיים rdd שנוצר ממסמך ה-json)

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

שאלה 5 - Naïve Bayes(10 נק')

להלן מידע על מזג האוויר שנאסף במגרש הגולף בקיסריה במשך 14 יום , כשבעמודה האחרונה מצוין האם התקיימו משחקי גולף באותו היום .

Outlook	Temp	Humidity	Windy	Play Golf
sunny	hot	hight	FALSE	No
sunny	hot	hight	TRUE	No
Overcast	hot	hight	FALSE	Yes
Rainy	Mild	hight	FALSE	Yes
Rainy	Cool	Normal	FALSE	Yes
Rainy	Cool	Normal	TRUE	No
Overcast	Cool	Normal	TRUE	Yes
Sunny	Mild	High	FALSE	No
Sunny	Cool	Normal	FALSE	Yes
Rainy	Mild	Normal	FALSE	Yes
Sunny	Mild	Normal	TRUE	Yes
Overcast	Mild	High	TRUE	Yes
Overcast	Hot	Normal	FALSE	Yes
Rainy	Mild	High	TRUE	No

בהינתן שידוע שיום חמישי האחרון היה שמשי (outlook= sunny), הטמפרטורות היו נמוכות (temp=cool)

הלחות הייתה גבוהה (Humidity = high) והיו רוחות (windy= true)

מצא ע"פ אלגוריתם *הנאייב ביי*ס האם התקיימו משחקי גולף ביום חמישי האחרון ?

יש לנמק ולפרט את דרך החישוב :

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

שאלה 6 - Linear regression (13 נק')

נתון הדאטה הבא בעל פיצ'ר אחד

x = [10, 17, 23, 40]

y = [15, 30, 44, 66]

סטודנט א' הציע את פונקציית הניבוי הבאה : $h(x) = x + 6$

סטודנט ב' הציע את פונקציית הניבוי הבאה: $h(x) = 2x + 1$

א. חשבו את ערך ה Loss בכל אחד מהאפשרויות

איזו פונקציית ניבוי עשויה לנבא ערך טוב יותר ל- y עבור דאטה חדש ? נמק (5 נק)

ב. שפר את פונקציית הניבוי של סטודנט א' באמצעות הפעלת איטרציה אחת של שיטת מורד הגרדיאנט , חשב עבורה את ה- loss החדש (5 נק)

ג. האם נכון להגיד כי ככל שמוספים יותר פיצ'רים ה- test error וה train error קטנים באותה מידה , נמקו (3 נק)

_____

_____

_____

_____