תרגיל 3 : SQL

תאריך הגשה : 19.06.24, 23:55.

הוראות הגשה:

בתרגיל זה אתם נדרשים להגיש קובץ zip בודד שיכלול את הקבצים הבאים :

- הקבצים: q1.sql, q2.sql, q3.sql, q4.sql, q5.sql, q6.sql, q7.sql, q8.sql, q9.sql, q10.sql, q11.sql
- example.sql – עבור חלק ב שאלה 2 סעיף ב.
- correct.sql – עבור חלק ב שאלה 2 סעיף ג.
- README שמכיל שורה בודדת ובו ה-login של הסטודנט שמגיש את התרגיל. אם התרגיל מוגש בזוגות, על שורה זאת להכיל את שני ה-login מופרדים בפסיק.

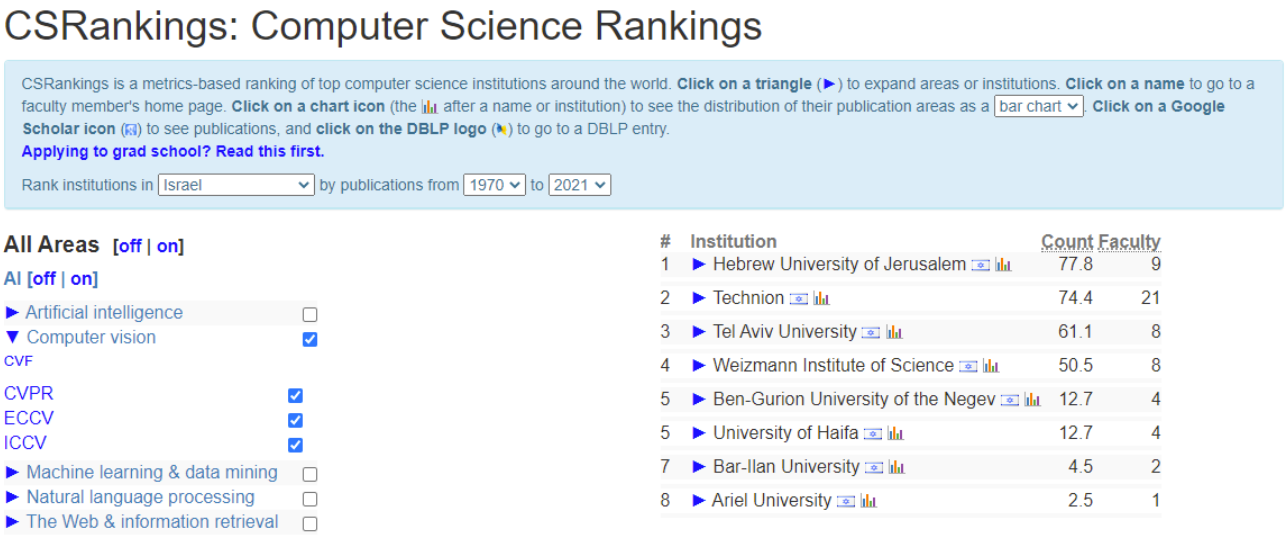
חלק א:

בתרגיל זה אנחנו נשתמש ביחסים ונתונים מתוך האתר http://csrankings.org/.

זהו אתר לדירוג מוסדות אקדמיים בתחום מדעי המחשב מסביב לעולם. האתר מדרג את המוסדות האקדמיים על סמך מספר הפרסומים של חברי סגל במוסד.

ניתן למצוא באתר מידע על כמות הפרסומים המדעיים של כל מוסד, לפי תחומים שונים של מדעי המחשב. ניתן לסנן את הנתונים לפי מדינות, לפי שנים ולפי תחומי המחקר. בנוסף ניתן לראות מידע על חברי סגל ספציפיים, וכן יש קישורים למידע מאתרים נוספים כמו dblp.

למשל, בצילום המסך הבא מוצג דירוג של המוסדות בישראל בתחום של Computer Vision בין השנים 1970-2021.



נתונים היחסים הבאים מתוך מסד נתונים של האתר CSRankings:

authors (name, conference, year, institution, count, adjustedcount)

conferences (conference, area, subarea)

institutions (institution, region, country)

הערות :

- בטבלה של מחברים (authors) יש את המידע על פרסומים של מחברים בכנסים שונים :
  - name – שם המחבר.

- conference – שם הכנס שבו הוא פרסם.
- year – השנה שבה פורסם המאמר בכנס.
- institution – שם המוסד האקדמי של המחבר.
- totalcount – מספר המאמרים שהמחבר פרסם באותו הכנס.
- adjustedcount – מספר הפרסומים היחסי של המחבר בכנס. למשל אם פרסם מאמר אחד והיה אחד משני כותבים הספירה היחסית תהיה 0.5. אם היו שלשה כותבים למאמר ...0.33 וכו'.

- בטבלה של הכנסים (conferences) יש את המידע לגבי הכנסים:
  - conference – שם הכנס
  - area – תחום המחקר של הפרסומים בכנס
  - Subarea – תת-תחום המחקר.
- בטבלה של מוסדות (institutions) יש מידע על מוסדות אקדמים:
  - institution – שם המוסד
  - region – אזור גיאוגרפי בעולם.
  - country – המדינה בה נמצא המוסד מיוצגת בתקציר ע"י שני אותיות. למשל ישראל היא il.

הערה: הטבלה authors מכילה מידע שחוזר על עצמו, כמו המידע על המוסד האקדמי של המחבר, שחוזר שוב ושוב בכל רשומה של פרסום של המחבר.
בהמשך הקורס נלמד איך לתכנן טבלאות בצורה שבה נמנע מחזרתיות של מידע (דבר הנקרא "צורה נורמלית" של טבלה).
אבל בעולם האמיתי תתקלו לצערנו ברבה מידע שאינו מנורמל, וכך גם המידע הנ"ל.

באתר הקורס יש קובץ create.sql המכיל הגדרות עבור הטבלאות וקובץ drop.sql המכיל פקודות המוחקות את הטבלאות.
כמו כן, נתונים הקבצים:

- generated-author-info.csv
- conferences.csv
- country-info.csv

הקבצים מכילים מידע על מחברים, פרסומים, כנסים ומוסדות אקדמים. המידע משמש את האתר http://csrankings.org/ לדירוג מוסדות אקדמים בתחום מדעי המחשב.

את המידע המלא ניתן למצוא בלינקים הבאים:

https://raw.githubusercontent.com/cohensara/csrankings/main/conferences.csv
https://raw.githubusercontent.com/emeryberger/CSrankings/gh-pages/generated-author-info.csv
https://raw.githubusercontent.com/emeryberger/CSrankings/gh-pages/country-info.csv

ניתן למצוא את הקבצים גם במערכת המחשבים במעבדה בתיקיה:

~ db2/data/ex3/

ניתן להעתיק אותם לתיקיה שלכם.

על מנת לבדוק את התרגיל שלכם, יש ליצור את הטבלאות בעזרת create.sql, ולטעון לתוכן נתונים בעזרת הפקודות

```
cat generated-author-info.csv | psql -h dbcourse public -c "copy authors from STDIN DELIMITER ',' CSV HEADER"
cat conferences.csv | psql -h dbcourse public -c "copy conferences from STDIN DELIMITER ',' CSV HEADER"
cat country-info.csv | psql -h dbcourse public -c "copy institutions from STDIN DELIMITER ',' CSV HEADER"
```

שאילתות SQL:

כתבו את השאילתות הבאות בSQL. שם הקובץ שבו צריכה להופיע התשובה לכל שאלה נמצא בתחילת השאלה.

שימו לב:

- השתמשו ב SELECT DISTINCT כדי למנוע כפילויות בתשובות (אם כפילויות עלולות להוויצר בתשובה).

- בכל סעיף כתוב באיזה סדר למיין את התוצאות וכן את שמות העמודות בתוצאה.

1. **(q1.sql)** החזר את את שמות המוסדות האקדמים שנמצאים בישראל.
יש להחזיר טבלה עם העמודה institution, ממויין לפי institution.

2. **(q2.sql)** החזר שמות המחברים ששייכים למוסד אקדמי שנמצא באפריקה, ואת שם המוסד בו הם עובדים.
יש להחזיר טבלה עם העמודות institution,name ממויין מיון ראשוני לפי institution ומיון שניוני לפי name.

3. **(q3.sql)** מצא את שמות המחברים ששייכים למוסד אקדמי שנמצא בישראל שפרסמו לפחות 2 מאמרים באותו הכנס ושם הכנס מתחיל במחרוזת "sig".
החזר את שמות המחברים ושמות המוסדות האקדמים שלהם הם שייכים.
יש להחזיר טבלה עם העמודות institution,name ממויין מיון ראשוני לפי institution ומיון שניוני לפי name.

4. **(q4.sql)** החזר את שמות כל החוקרים מהאוניברסיטה העברית (Hebrew University of Jerusalem) שפרסמו באותה שנה גם בכנס **בתת התחום** של 'ai' וגם בכנס בתת התחום של 'economics'.
יש להחזיר טבלה עם העמודות name,year, ממויין לפי name ומיון שניוני לפי year.

5. **(q5.sql)** החזר את את שמות המחברים שפרסמו אך ורק בתחום theory, וכן אך ורק לפני שנת 1980.
יש להחזיר טבלה עם העמודה name, ממויין לפי name.

6. **(q6.sql)** החזר את את שמות המחברים שפרסמו מאמר בכל אחד מהכנסים שבו- Omri Abend פרסם דווקא באותה שנה שבה עמרי פרסם. (גם עמרי עומד בתנאי וצריך לחזור בתוצאת השאילתה)
יש להחזיר טבלה עם העמודה name, ממויין לפי name.

7. **(q7.sql)** לכל מדינה החזר את מספר המוסדות האקדמיים הקיימים בה. לעמודת מספר המוסדות יש לקרוא בשם institutionCount
יש להחזיר טבלה עם העמודות country, institutionCount ממויינת לפי country .

8. **(q8.sql)** נאמר שמחבר הוא מומחה בתת התחום ml אם פרסם מאמרים בלפחות שלשה כנסים שונים כנסים בתת תחום זה. נאמר שמומחה ל-ml הוא עדכני אם הוא פרסם לפחות מאמר אחד בתחום ה- ml החל מ-2020. החזר את שמות כל המומחים העדכניים בתת התחום ה-ml.
יש להחזיר טבלה עם העמודה name ממויינת.

9. **(q9.sql)** לכל מדינה החזר את המוסד ממנו פורסמו הכי הרבה מאמרים באותה מדינה (על פי totalcount), ואת מספר המאמרים שפורסמו בו (אם ישנן כמה מוסדות מהם פורסם אותו מספר מקסימלי החזר את כולם).
יש להחזיר טבלה עם העמודות country, institution, countryCount ממויינת לפי country, ואז לפי institution.
<u>הדרכה -</u> בשאלה זו (וגם בשאלות הבאות) מומלץ להשתמש בפקודת with (הסבר נמצא בסרטון אחרון של שבוע 4 ואפשר למצוא גם פה https://www.postgresql.org/docs/current/queries-with.html בפסקה הראשונה).
הגדירו בעזרת with טבלת עזר ובה מספר הפרסומים שיש לכל מוסד, והאזור המדינה בה שוכן המוסד.

10. **(q10.sql)** נאמר שחבר סגל מהאוניברסיטה העברית הוא שיאן הפרסומים בעברית בתחום בינה מלאכותית (ai) בשנה x אם אין חבר סגל אחר מהאוניברסיטה העברית עם יותר פרסומים בתחום של ai (כלומר סכום totalcount גדול יותר) באותה שנה. שימו לב שיכולים להיות כמה שיאנים במקרה של שוויונות.
לכל האחד מהשנים 2000 עד 2020 החזירו שורה, או שורות, במקרה של מספר שיאנים, מהצורה (y,n) כאשר y הוא השנה ו-n הוא שם שיאן הפרסומים בעברית.
יש להחזיר טבלה עם העמודות year,name ממיונן במיון ראשוני לפי year ומיון שניוני לפי name

11. **(q11.sql)** נאמר שכנס הוא צעיר אם התקיים לכל היותר 15 פעמים (לאו דווקא בשנים רצופות). החזר את שמות המחברים שהשתתפו רק בכנסים צעירים.
יש להחזיר טבלה עם עמודה בודדת ממויינת של שמות המחברים הנקראת name.

חלק ב:

כפי שרובכם בוודאי כבר יודעים, ChatGPT הוא צ'אטבוט שפותח על-ידי חברת OpenAI על בסיס מודל השפה שלה. בין היכולות הרבות שלו, אפשר להשתמש ב-ChatGPT כדי לדמות כתיבת קוד בשפות שונות, ואף כדי לדמות כתיבת שאילתות ב-SQL.
אבל כמו שמחשבון הוא מאוד שימושי, אך לא פותר אותנו מלדעת מתמטיקה, כך גם ChatGPT יכול להיות כלי מאוד שימושי, אך אינו יכול להחליף את הצורך של מתכנתים לדעת לכתוב ולהבין קוד בעצמם (או את הצורך של סטודנטים לעבור את המבחן).
בשאלה זו נדגים את הבעייתיות בשימוש בChatGPT לכתיבת שאילתת SQL בלי לדעת לכתוב שאילתה נכונה בעצמנו.
נתונה שאלת המבחן הבאה שנשלחה ל-ChatGPT:

נתונות הגדרות הטבלאות הבאות:

```
CREATE TABLE Course (
        cid VARCHAR(255) PRIMARY KEY,
        department VARCHAR(255) );
CREATE TABLE Prerequisites (
        cid VARCHAR(255),
         preid VARCHAR(255),
        PRIMARY KEY (cid, preid)
        FOREIGN KEY (cid) REFERENCES Course(cid),
        FOREIGN KEY (preid) REFERENCES Course(cid));
```

כאשר:

- היחס Course מכיל מספרי קורסים של המחלקה לה הם שייכים.
- היחס Prerequisites מכיל זוגות של מספרי קורסים כך ש-preid מהווה דרישת קדם עבור cid.

כתבו שאילתת SQL אשר מחזירה את כל זוגות הקורסים אשר יש להם בדיוק אותם דרישות קדם. ניתן להניח שאין ערכי null בטבלאות.

החזר טבלה עם העמודות cid1,cid2 ממויינת לפי cid1 ואחר כך לפי cid2 בסדר עולה, כך שכל זוג יופיע פעם אחת ו-cid1 < cid2 .

התשובה שהתקבלה מ-:ChatGPT

```sql
SELECT
   c1.cid AS cid1,
   c2.cid AS cid2
FROM
   (SELECT cid, STRING_AGG(preid, ',') AS PrereqList
    FROM Prerequisites
    GROUP BY cid) AS c1
JOIN
   (SELECT cid, STRING_AGG(preid, ',') AS PrereqList
    FROM Prerequisites
    GROUP BY cid) AS c2
ON c1.PrereqList = c2.PrereqList AND c1.cid < c2.cid
ORDER BY c1.cid, c2.cid;
```

התשובה שגויה.
(הערה: הפקודה STRING_AGG משרשרת מחרוזות, אפשר למצוא תיעוד בלינק הבא:
https://www.postgresqltutorial.com/postgresql-aggregate-functions/postgresql-string_agg-function/)

א. תן דוגמה מינימאלית של טבלה שמדגימה שהשאילתה מחזירה תשובה שגויה.
תוכלו להשתמש בקישור הבא ל fiddle כדי לנסות דוגמאות שונות ולהריץ עליהן את השאילתה השגויה:
SQL_FIDDLE
את התשובה יש להגיש בצורה של פקודות insert לטבלאות בקובץ בשם example.sql.
(**אין** לכלול בקובץ את פקודות ה-create של הטבלאות).

ב. כתוב שאילתה נכונה עבור השאלה.
את התשובה יש להגיש בקובץ correct.sql.
(**אין** לכלול בקובץ את פקודות ה-create של הטבלאות או פקודות insert).

בהצלחה!